Preface

We were very pleased to once again extend to the delegates and, we are pleased to say, our friends the warmest of welcomes to the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems at Wellington Institute of Technology in Wellington, New Zealand.

The KES conferences attract a wide range of interest. The broad focus of the conference series is the theory and applications of computational intelligence and emergent technologies. Once purely a research field, intelligent systems have advanced to the point where their abilities have been incorporated into many conventional application areas. The quest to encapsulate human knowledge and capabilities in domains such as reasoning, problem solving, sensory analysis, and other complex areas has been avidly pursued. This is because it has been demonstrated that these abilities have definite practical applications. The techniques long ago reached the point where they are being exploited to provide commercial advantages for companies and real beneficial effects on profits. KES 2004 provided a valuable mechanism for delegates to obtain a profound view of the latest intelligent systems research into a range of algorithms, tools and techniques. KES 2004 also gave delegates the chance to come into contact with those applying intelligent systems in diverse commercial areas. The combination of theory and practice represents a uniquely valuable opportunity for appreciating the full spectrum of intelligent-systems activity and the “state of the art”.

For the first time in the short history of KES, the conference came to New Zealand. KES 2004 aimed at providing not only a high-tech forum for presenting results on theory and applications of intelligent systems and techniques, but focused on some significant emerging intelligent technologies including evolvable hardware (EHW), evolutionary computation in computational intelligence, DNA computing, artificial immune systems (AIS), bioinformatics using intelligent and machine learning techniques, and intelligent Web mining.

The impressive audience of the KES conferences series was confirmed, and we broke some KES records, such as: about 500 attendants from 55 countries, and for the first time in the conference history, more than one third of the participant presenting high-quality papers were Ph.D. students from all over the world. This last detail is relevant for the major role played by the KES organization and conferences with respect to support and education for practitioners who are acting in the area of intelligent systems and emergent technologies.

Thanking all the individuals who contributed to a conference like this is always fraught with difficulty, as someone is always unintentionally omitted. The WelTec team, including Gary Hartley, the conference administrator, Michael Hyndman, the conference Web page designer, and the Local Organizing Committee, chaired by Dr. Linda Sissons, WelTec CEO, all worked hard to bring the conference to a high level of organization. We would like to arrange a special appreciation on behalf of the KES 2004 General Chair for the hard work done by David Pritchard from the WelTec Centre for Computational Intelligence. We would like to extend our praise and thanks to them.
An important distinction of the KES conferences over others is the Invited Session Program. Invited sessions give new and dedicated researchers an opportunity to present a “mini-conference” of their own. By this means they can bring to public view a topic at the leading edge of intelligent science and technology. This mechanism for feeding new blood into the research is immensely valuable, and strengthens KES conferences enormously. For this reason we must extend thanks to the Invited Session Chairs who contributed in this way.

We would like to thank the KES 2004 International Program Committee and the KES 2004 Reviewers Team who were essential in providing their reviews of the papers. We are immensely grateful for this service, without which the conference would not have been possible. We thank the high-profile keynote speakers and invited tutorial lecturers for providing interesting and informed talks to catalyze subsequent discussions.

In some ways, the most important contributors to KES 2004 were the authors, presenters and delegates without whom the conference could not have taken place. So we thank them for their contributions. Finally we thank the “unsung heroes” the army of administrators, caterers, hoteliers, and the people of Wellington, for welcoming us and providing for the conference.

We hope the attendees all found KES 2004 a worthwhile, informative and enjoyable experience. We hope to see them in Melbourne for KES 2005, which will be hosted by La Trobe University, Melbourne, Australia.

June 2004                                          Prof. Mircea Gh. Negoita
                                             Dr. R.J. Howlett
                                             Prof. Lakhmi C. Jain
KES 2004 Conference Organization

General Chair

Mircea Negoita
Centre for Computational Intelligence
School of Information Technology
Wellington Institute of Technology (WelTec), Wellington, New Zealand
Co-director of NZ-German School on Computational Intelligence at KES 2004

Conference Founder and Honorary Programme Committee Chair

Lakhmi C. Jain
Knowledge-Based Intelligent Information and Engineering Systems Centre
University of South Australia, Australia

KES Executive Chair

Bob Howlett
Intelligent Systems and Signal Processing Laboratories/KTP Centre
University of Brighton, UK

KES 2004 Invited Co-chair

Bernd Reusch
Department of Computer Science
University of Dortmund, Germany
Co-director of NZ-German School on Computational Intelligence at KES 2004

KES Journal General Editor

Bogdan Gabrys
University of Bournemouth, UK
Local Organizing Committee

Linda Sissons – Chair, WelTec CEO
Gary Hartley, Mircea Gh. Negoita, Murray Wills
Wellington Institute of Technology (WelTec), New Zealand

KES 2004 Web Page Designer

Michael Hyndman
Wellington Institute of Technology (WelTec), New Zealand

Technical Emergence Desktop Team

Doug StJust
Ali Rashid Mardani
Wellington Institute of Technology (WelTec), New Zealand

KES 2004 Liaison Officer

Lesley Lucie-Smith
Wellington Institute of Technology (WelTec), New Zealand

Proceedings Assembling Team

David Pritchard
Paulene Mary Crook
Ian Hunter
Terry Jeon
Des Kenny
Sara Rule
Nick Tullock
Wellington Institute of Technology (WelTec), New Zealand
International Program Committee

Hussein Abbass, University of New South Wales, Australia
Peter Andreae, Victoria University, Wellington, New Zealand
Viorel Ariton, “Danubius” University of Galatzi, Romania
Akira Asano, Hiroshima University, Higashi-Hiroshima, Japan
K. Vijayan Asari, Old Dominion University, Norfolk, Virginia, USA
Norio Baba, Osaka Kyoiku University, Japan
Robert Babuska, Delft University of Technology, Delft, The Netherlands
Andrzej Bargiela, Nottingham Trent University, UK
Marius Bazu, Institute of Microtechnology, Bucharest, Romania
Yevgeniy Bodyanskiy, Kharkiv National University of Radioelectronics, Ukraine
Patrick Bosc, IRISA/ENSSAT, Lanion, France
Pascal Bouvry, Luxembourg University of Applied Sciences, Luxembourg
Phillip Burrell, South Bank University, London, UK
Yen-Wei Chen, University of the Ryukyus, Okinawa, Japan
Vladimir Cherkassky, University of Minnesota, USA
Krzysztof Cios, University of Colorado at Denver, USA
Carlos A. Coello, LANIA, Mexico
George Coghill, Auckland University, Auckland, New Zealand
David W. Corne, University of Exeter, UK
David Cornforth, Charles Sturt University, Albury, Australia
Ernesto Damiani, University of Milan, Italy
Da Deng, University of Otago, Dunedin, New Zealand
Da Ruan, Belgian Nuclear Research Centre (SCK · CEN), Belgium
Vladan Devedzic, University of Belgrade, Belgrade, Serbia
Didier Dubois, IRIT, Université Paul Sabatier, Toulouse, France
Duncan Earl, Oak Ridge National Laboratory, USA
Madjid Fathi, National Magnet Lab., Florida, USA
Marcus Frean, Victoria University, Wellington, New Zealand
Peter Funk, Mälardalen University, Västerås, Sweden
Bogdan Gabrys, University of Bournemoth, UK
Boris Galitsky, Birkbeck College, University of London, UK
Hugo de Garis, Utah State University, USA
Max H. Garzon, University of Memphis, USA
Tamas Gedeon, Murdoch University, Murdoch, Australia
Mitsuo Gen, Waseda University, Kytakyushu, Japan
Vladimir Gorodetski, St. Petersburg Institute of Informatics, Russian Academy of Sciences, Russia
Manuel Grana, Facultad de Informatic, UPV/EHU, Spain
David Gwaltney, NASA George C. Marshall Space Flight Center, Huntsville, USA
Lars Kai Hansen, Technical University of Denmark, Lyngby, Denmark
Chris Harris, University of Southampton, UK
Lars Hildebrand, Dortmund University, Dortmund, Germany
Tetsuya Highchi, National Institute of Advanced Industrial Science and Technology, Japan
Yuzo Hirai, University of Tsukuba, Japan
Dawn Holmes, University of California, Santa Barbara, USA
Daniel Howard, University of Limerick, Ireland
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Keiichi Horio, Kyushu Institute of Technology, Japan
Hitoshi Iba, University of Tokyo, Tokyo, Japan
Florin Ionescu, University of Applied Sciences, Konstanz, Germany
Hisao Ishibuchi, Osaka Prefecture University, Osaka, Japan
Naohiro Ishii, Aichi Institute of Technology, Toyota City, Japan
Mo M. Jamshidi, University of New Mexico, Albuquerque, USA
Norbert Jesse, Dortmund University, Dortmund, Germany
Seong-Joon Yoo, Sejong University, Seoul, Korea
Janusz Kacprzyk, Polish Academy of Sciences, Poland
Nikos Karacapilidis, University of Patras, Greece
Vojislav Kecman, Auckland University, Auckland, New Zealand
Rajiv Khosla, La Trobe, University, Melbourne, Australia
Laszlo T. Koczy, Budapest University of Technology and Economics, Budapest and Szechenyi Istvan University, Gyor, Hungary
Hiroyasu Koshimizu, Chukyo University, Toyota, Japan
Susumu Kunifujii, Japan Advanced Institute of Science & Technology, Japan
Andrew Kusiak, University of Iowa, Iowa City, USA
W.K. Lai, MIMOS Bhd., Kuala Lumpur, Malaysia
Pier Luca Lanzi, Polytechnic Institute, Milan, Italy
Raymond Lee, Hong Kong Polytechnic University, Kowloon, Hong Kong
Chee-Peng Lim, University of Science Malaysia, Penang, Malaysia
Jason Lohn, NASA Ames Research Center, Mountain View, CA, USA
Ignac Lovrek, University of Zagreb, Croatia
Bruce MacDonald, Auckland University, Auckland, New Zealand
Bob McKay, University of NSW, Australian Defence Force Academy, Australia
Luis Magdalena-Layos, EUSFLAT & Universidad Politecnica de Madrid, Spain
Dan C. Marinescu, University of Central Florida, Orlando, USA
Jorma K.Mattila, Lappeenranta University of Technology, Finland
Radko Mesiar, Slovak Technical University, Bratislava, Slovakia
Claudio Moraga, University of Dortmund, Germany
Hirofumi Nagashino, University of Tokushima, Tokushima, Japan
Noriko Nagata, Kwansei Gakuin University, Japan
Ryohei Nakatsu, Kwansei Gakuin University, Japan
Koji Nakajima, Tohoku University, Sendai, Japan
Akira Namatame, National Defense Academy, Yokosuka, Japan
Victor Emil Neagoe, Technical University Bucharest, Romania
Ciprian Daniel Neagu, University of Bradford, UK
Charles Nguyen, Catholic University of America, Washington, DC, USA
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland
Toyoaki Nishida, University of Tokyo, Japan
Nikhil R. Pal, Indian Statistical Institute, Calcutta, India
Vasile Palade, Oxford University, UK
Costas Papis, University of Piraeus, Greece
Ian C. Parmee, University of the West of England, Bristol, UK
Carlos-Andrés Pena-Reyes, Swiss Federal Institute of Technology–EPFL, Lausanne, Switzerland
Theodor Popescu, National Institute for Research and Development Informatics, Bucharest, Romania
John A. Rose, University of Tokyo, Tokyo, Japan
Eugene Roventa, York University, Toronto, Canada
Rajkumar Roy, Cranfield University, UK
Takeshi Samatsu, Kyushu Tokai University, Japan
Elie Sanchez, Université de la Méditerranée, Marseille, France
Marc Schoenauer, INRIA Rocquencourt, Le Chesnay, France
Udo Seiffert, Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany
Barry Smyth, University College Dublin, Ireland
Flavio Soares Correa da Silva, Instituto de Matematica e Estatistica, University of São Paulo, Brazil
Von-Wun Soo, National Tsing Hua University, Taiwan
Adrian Stoica, NASA Jet Propulsion Laboratory, Pasadena, USA
Noriaki Suetake, Yamaguchi University, Japan
Sarawut Sujitjorn, Suranaree University of Technology, Thailand
Mieko Tanaka-Yamawaki, Tottori University, Japan
Takushi Tanaka, Fukuoka Institute of Technology, Japan
Eiichiro Tazaki, Toin University of Yokohama, Japan
Jon Timmis, University of Kent at Canterbury, UK
Jim Torresen, University of Oslo, Norway
Kazuhiko Tsuda, University of Tsukuba, Japan
Andy M. Tyrrell, University of York, UK
Eiji Uchino, University of Yamaguchi, Japan
Angel Navia Vazquez, Universidad Carlos III de Madrid, Spain
Jose Luis Verdegay, University of Granada, Granada, Spain
Dianhui Wang, La Trobe University, Melbourne, Australia
Pei Wang, Temple University, Philadelphia, USA
Junzo Watada, Waseda University, Kitakyushu, Fukuoka, Japan
Keigo Watanabe, Saga University, Japan
Takeshi Yamakawa, Kyushu Institute of Technology, Graduate School of Life Science and Systems Engineering, Japan
Xin Yao, University of Birmingham, UK
Kaori Yoshida, Kyushu Institute of Technology, Japan
Lotfi A. Zadeh, University of California at Berkeley, USA
Ricardo Zebulum, NASA Jet Propulsion Laboratory, Pasadena, USA
Invited Session Chairs Committee

Akinori Abe, ATR Intelligent Robotics & Communication Labs, Kyoto, Japan
Yoshinori Adachi, Chubu University, Japan
Alicia d’Anjou, Universidad del Pais Vasco, Spain
Norio Baba, Osaka Kyoiku University, Japan
Pascal Bouvry, Luxembourg University of Applied Sciences, Luxembourg
Malu Castellanous, Hewlett-Packard Laboratories, Palo Alto, CA, USA
Yen-Wei Chen, Ritsumeikan University, Japan
George G. Coghill, Auckland University, New Zealand
Ernesto Damiani, University of Milan, Italy
Vladan Devedzic, University of Belgrade, Serbia and Montenegro
Marijan Druzovec, University of Maribor, Slovenia
Richard Duro, Universidad de A Coruña, Spain
Minoru Fukumi, University of Tokushima, Japan
Boris Galitsky, Birkbeck College, University of London, UK
Max H. Garzon, University of Memphis, USA
Wanwu Guo, Edith Cowan University, Australia
Manuel Graña, Universidad Pais Vasco, Spain
Jerzy M. Grzymala-Busse, University of Kansas, USA
Robert F. Harrison, University of Sheffield, UK
Philip Hingston, Edith Cowan University, Australia
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Nikhil Ichalkaranje, University of South Australia, Adelaide, Australia
Takumi Ichimura, Hiroshima University, Japan
Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan
Yoshiteru Ishida, Toyohashi University of Technology, Japan
Naohiro Ishii, Aichi Institute of Technology, Japan
Yuji Iwahori, Chubu University, Japan
Lakhmi C. Jain, University of South Australia, Adelaide, Australia
Taki Kanda, Bunri University of Hospitality, Japan
Radoslaw P. Katarzyniak, Wroclaw University of Technology, Poland
Le Kim, University of South Australia, Adelaide, Australia
Tai-hoon Kim, Korea Information Security Agency (KISA), Korea
Rajiv Khosla, La Trobe University, Melbourne, Australia
Peter Kokal, University of Maribor, Slovenia
Naoyuki Kubota, Tokyo Metropolitan University, Tokyo, Japan
Mineichi Kudo, Hokkaido University, Japan
Chiaki Kuroda, Tokyo Institute of Technology, Tokyo, Japan
Susumu Kunifujji, Japan Advanced Institute of Science and Technology, Japan
Weng Kim Lai, MIMOS Berhad, Technology Park, Malaysia
Dong Chun Lee, Hongwon University, Korea
Huey-Ming Lee, Chinese Culture University, Taiwan
Raymond Lee, Hong Kong Polytechnic University, Kowloon, Hong Kong
KES 2004 Reviewers

R. Abdulah, University of Science Malaysia, Malaysia
A. Abe, ATR Intelligent Robotics & Communication Labs., Kyoto, Japan
Y. Adachi, Chubu University, Aichi, Japan
P. Andreae, Victoria University, Wellington, New Zealand
A. Asano, Hiroshima University, Higashi-Hiroshima, Japan
K.V. Asari, Old Dominion University, Norfolk, Virginia, USA
N. Ashidi, KES 2004 Reviewers Team
D. Arita, Kyushu University, Fukuoka, Japan
N.A. Aziz, MIMOS, Malaysia
N. Baba, Osaka Kyoiku University, Japan
R. Babuska, Delft University of Technology, Delft, The Netherlands
O. Boissier, École des Mines de Saint-Étienne, France
P. Bosc, IRISA/ENSSAT, France
P. Bouvry, Luxembourg University of Applied Sciences, Luxembourg
G. Bright, Massey University, Auckland, New Zealand
D.A. Carnegie, Waikato University, Hamilton, New Zealand
M. Castellaneous, Hewlett-Packard Laboratories, Palo Alto, CA, USA
C.-T. Chang, National Cheng Kung University, Taiwan
Y.-W. Chen, Ritsumeikan University, Japan
S.-C. Chi, Huafan University, Taiwan
B.-C. Chien, I-Shou University, Taiwan
G.G. Coghill, Auckland University, Auckland, New Zealand
D.W. Corne, University of Exeter, UK
D. Cornforth, Charles Sturt University, Albury, Australia
A. Czyzewski, Gdansk University of Technology, Gdansk, Poland
E. Damiani, University of Milan, Italy
R.J. Deaton, University of Arkansas, USA
Da Deng, University of Otago, Dunedin, New Zealand
V. Devedzic, University of Belgrade, Serbia and Montenegro
P.M. Drezet, University of Sheffield, UK
R. Dunlog, University of Canterbury, Christchurch, New Zealand
C. Elamvazuthi, MIMOS, Malaysia
T. Ejima, Aichi University of Education, Aichi, Japan
M. Fathi, National Magnet Lab., Florida, USA
M. Frean, Victoria University, Wellington, New Zealand
W. Friedrich, Industrial Research Limited, Auckland, New Zealand
T. Fujinami, JAIST, Japan
P. Funk, Mälardalen University, Västerås, Sweden
B. Gabrys, Bournemouth University, UK
M.H. Garzon, University of Memphis, USA
B. Galitsky, Birkbeck College, University of London, UK
T. Gedeon, Murdoch University, Murdoch, Australia
V. Gorodetski, St. Petersburg Institute of Informatics, Russia
M. Grana, Universidad Pais Vasco, Spain
J.W. Grzymala-Busse, University of Kansas, USA
N. Guelfi, Luxembourg University of Applied Sciences, Luxembourg
F. Guinand, Le Havre University, France
W. Guo, Edith Cowan University, Australia
M. Hagiya, University of Tokyo, Japan
L.K. Hansen, Technical University of Denmark, Lyngby, Denmark
A. Hara, Hiroshima City University, Japan
R.F. Harrison, University of Sheffield, UK
Y. Hayakawa, Tohoku University, Japan
L. Hildebrand, University of Dortmund, Germany
P. Hingston, Edith Cowan University, Australia
K. Hirayama, University of Kitakyushu, Kitakyushu, Japan
O.S. Hock, University of Malaya, Malaysia
T.-P. Hong, National University of Kaohsiung, Taiwan
K. Horio, Kyushu Institute of Technology, Fukuoka, Japan
D. Howard, University of Limerick, Ireland
T. Ichikawa, Shizuoka University, Japan
T. Ichimura, Hiroshima City University, Japan
N. Ichalkaranje, University of South Australia, Australia
F. Ishida, University of Electro-communications, Japan
Y. Ishida, Toyohashi University of Technology, Japan
N. Ishii, Aichi Institute of Technology, Japan
S. Ito, ATR, Japan
Y. Iwahori, Chubu University, Aichi, Japan
S. Iwamoto, Kyushu University, Fukuoka, Japan
M.E. Jefferies, Waikato University, Hamilton, New Zealand
N. Jesse, University of Dortmund, Germany
K. Juszczyszyn, Wroclaw University of Technology, Poland
D. Khadraoui, CRP Tudor, Luxembourg
K. Kakusho, Kyoto University, Kyoto, Japan
T. Kanda, Bunri University of Hospitality, Japan
T. Kanai, Meijin-gakuin University, Japan
N. Karakapilidis, University of Patras, Greece
R.P. Katarzyniak, Wroclaw University of Technology, Poland
N. Katayama, Tohoku University, Japan
P. Kazienko, Wroclaw University of Technology, Poland
V. Kecman, Auckland University, New Zealand
S.J. Kia, New Zealand
C.W. Kian, Ohio Northern University, USA
L. Kim, University of Canberra, Australia
C.P. Lian, DSTO, Australia
C.-P. Lim, University of Science Malaysia, Malaysia
D.N.C. Ling, Multimedia University, Malaysia
M. Kinjo, Tohoku University, Japan
Y. Kinouchi, University of Tokushima, Japan
A.T. Khader, University of Science Malaysia, Malaysia
R. Khosla, La Trobe University, Melbourne, Australia
T. Koda, Kyoto University, Japan
T. Komatsu, Future University Hakodate, Hakodate, Japan
T. Kondo, KES 2004 Reviewers Team
B. Kostec, Gdansk University of Technology, Gdansk, Poland
N. Kubota, Tokyo Metropolitan University, Tokyo, Japan
M. Kudo, University of Hokkaido, Japan
N. Kulathuramaiyer, University Malaysia Sarawak, Malaysia
S. Kumamoto, University of Kytkyushu, Japan
S. Kunifuji, Japan Advance Institute of Science and Technology (JAIST), Japan
H.-C. Kuo, National Chiayi University, Taiwan
M. Kurano, Chiba University, Japan
C. Kuroda, Tokyo Institute of Technology, Japan
T. Kuroda, KES 2004 Reviewers Team
S. Kurohashi, University of Tokyo, Japan
Y. Kurosawa, Hiroshima City University, Japan
A. Kusiai, University of Iowa, Iowa City, USA
S. Kurohashi, University of Tokyo, Japan
Y. Kurosawa, Hiroshima City University, Japan
W.K. Lai, MIMOS Berhad, Technology Park, Malaysia
D.C. Lee, Howon University, Korea
H.-M. Lee, Chinese Culture University, Taiwan
R. Lee, Hong Kong Polytechnic University, Hong Kong
C.P. Lian, KES 2004 Reviewers Team
J.-H. Lin, I-Shou University, Taiwan
W.-Y. Lin, I-Shou University, Taiwan
D.N.C. Ling, KES 2004 Reviewers Team
C.-P. Lim, University of Science Malaysia, Penang, Malaysia
H. Li, Edith Cowan University, Australia
C. Liu, Shenyang Institute of Technology, Shenyang, China
I. Lovrek, University of Zagreb, Croatia
B. MacDonald, Auckland University, New Zealand
B. McKay, University of New South Wales, Australian Defence Force Academy, Australia
David McG. Squire, Monash University, Australia
Z. Ma, Northeast Normal University, China
L. Magdalena-Layos, EUSFLAT and Universidad Politecnica de Madrid, Spain
N.A. Matisa, University of Science, Malaysia, Malaysia
C. Messom, Massey University, Auckland, New Zealand
C. Moraga, University of Dortmund, Germany
N. Mort, University of Sheffield, UK
K. Mera, Hiroshima City University, Japan
M. Minoh, ACCMS, Kyoto University, Japan
M. Miura, JAIST, Japan
Y. Mizugaki, University of Electro-communications, Japan
T. Mizuno, Shizuoka University, Japan
S. Sato, Tohoku University, Japan
R. Sakamoto, JAIST, Japan
E. Sanchez, Université de la Méditerranée, Marseille, France
C. Schommer, Luxembourg University of Applied Sciences, Luxembourg
S. Scott, Asia Pacific Institute of Technology, Malaysia
N. Seeman, New York University, USA
U. Seifert, Leibniz Institute of Plant Genetics and Crop Plant Research, Germany
F. Seredynski, PJWSTK/IPIPAN, Poland
T. Shimooka, Hokkaido University, Sapporo, Japan
F.S. Correa da Silva, Instituto de Matematica e Estatistica, University of São Paulo, Brazil
V.-W. Soo, National Tsing Hua University, Taiwan
U. Sørger, Luxembourg University of Applied Sciences, Luxembourg
P. Sturm, University of Trier, Germany
N. Suetsuke, Yamaguchi University, Japan
K. Sugiyama, JAIST, Japan
M. Suka, St. Marianna University, Japan
S. Suwitjorn, Suranaree University of Technology, Thailand
Y. Sumi, Kyoto University, Kyoto, Japan
N. Surayana, Multimedia University, Malaysia
A. Suyama, University of Tokyo, Japan
M. Takano, University of Tokyo, Japan
H. Taki, Wakayama University, Japan
M. Takano, University of Tokyo, Japan
H. Taki, Wakayama University, Japan
Y.-H. Tao, National Pingtung University of Technology and Science, Taiwan
T. Tanaka, Fukuoka Institute of Technology, Fukuoka, Japan
R. Taniguchi, Kyushu University, Fukuoka, Japan
E.H. Tat, Multimedia University, Malaysia
J. Timmis, University of Kent at Canterbury, UK
J. Torresen, University of Oslo, Norway
K. Tsuda, University of Tsukuba, Tokyo, Japan
C. Turchetti, Università Politecnica delle Marche, Ancona, Italy
E. Uchino, University of Yamaguchi, Japan
H. Ueda, Hiroshima City University, Japan
K. Ueda, University of Tokyo, Japan
K. Umemoto, JAIST, Japan
K. Unsworth, Auckland University, New Zealand
K. Uosaki, Osaka University, Japan
J. Xiao, Edith Cowan University, Australia
N. Xiong, KES 2004 Reviewers Team
H. Yamaba, Miyazaki University, Japan
T. Yamakami, ACCESS, Japan
Y. Yamashita, Tohoku University, Japan
H. Yan, Duke University, USA
X. Yao, University of Birmingham, UK
M. Yasuda, Chiba University, Japan
S.-J. Yoo, Sejong University, Seoul, Korea
J. Yoon, Institute of Science and Technology, Korea
K. Yoshida, St. Marianna University, Japan
Y. Yoshida, University of Kitakyushu, Japan
T. Yoshino, Wakayama University, Japan
K.-M. Yu, Chung-Hua University, Taiwan
D.C.K. Yuen, Auckland University, New Zealand
T. Yuizono, Shimane University, Japan
D. Wang, La Trobe University, Melbourne, Australia
P. Wang, Temple University, Philadelphia, USA
S.-L. Wang, New York Institute of Technology, USA
X. Wang, Hebei University, China
J. Watada, Waseda University, Japan
K. Watanabe, Saga University, Japan
Y. Watanabe, Toyohashi University of Technology, Japan
E. Weidert, Luxembourg University of Applied Sciences, Luxembourg
T. Welzer, University of Maribor, Slovenia
S. Wilk, Poznan University of Technology, Poland
C.-H. Wu, Shu-Te University, Taiwan
V. Zharkova, University of Bradford, UK
A. Zomaya, University of Sydney, Australia
C. Zhao, Edith Cowan University, Australia
Z. Zheng, Chinese Academy of Sciences, Beijing, China

Sponsors

"IPE NZ is proud to be associated with this knowledge event for engineers."

New Zealand's Economic and Trade Development Agency
Table of Contents, Part III

Engineering of Ontology and Multi-agent System Design

Implementing EGAP-Based Many-Valued Argument Model for Uncertain Knowledge
Taro Fukumoto, Takehisa Takahashi, Hajime Sawamura..............................1

Ontology Revision Using the Concept of Belief Revision
Seung Hwan Kang, Sim Kim Lau.................................................................8

A Robust Rule-Based Event Management Architecture for Call-Data Records
C. W. Ong, J. C. Tay...............................................................................16

Adaptive Agent Integration in Designing Object-Based Multiagent System
Jaya Sil.........................................................................................................24

Ontological Representations of Software Patterns
Jean-Marc Rosengard, Marian F. Ursu.......................................................31

Intelligent Multimedia Solution and the Security for the Next Generation Mobile Networks

Dynamic Traffic Grooming and Load Balancing for GMPLS-Centric All Optical Networks
Hyuncheol Kim, Seongjin Ahn, Jinwook Chung........................................38

Probabilistic Model of Traffic Breakdown with Random Propagation of Disturbance for ITS Application
Bongsoo Son, Taewan Kim, Hyung Jin Kim, Soobeom Lee..........................45

Novel Symbol Timing Recovery Algorithm for Multi-level Signal
Kwang Ho Chun, Myoung Seob Lim..........................................................52

Development Site Security Process of ISO/IEC TR 15504
Eun-ser Lee, Tai-hoon Kim........................................................................60

Improving CAM-DH Protocol for Mobile Nodes with Constraint Computational Power
Yong-Hwan Lee, Il-Sun You, Sang-Surm Rhee............................................67

Space Time Code Representation in Transform Domain
Gi Yean Hwang, Jia Hou, Moon Ho Lee.....................................................74
A Multimedia Database System Using Mobile Indexing Agent in Wireless Network
Jong-Hee Lee, Kwang-Hyoung Lee, Moon-Seog Jun, Keun-Wang Lee......................81

Bus Arrival Time Prediction Method for ITS Application
Bongsoo Son, Hyung Jin Kim, Chi-Hyun Shin, Sang-Keon Lee.........................88

RRAM Spare Allocation in Semiconductor Manufacturing for Yield Improvement
Youngshin Han, Chilgee Lee.................................................................95

A Toolkit for Constructing Virtual Instruments for Augmenting User Interactions and Activities in a Virtual Environment
Kyoung S. Park, Yongjoo Cho.................................................................103

Mobility Grouping Scheme to Reduce HLR Traffic in IMT-2000 Networks
Dong Chun Lee, Gwang-Hyun Kim, Seung-Jae Yoo.................................110

Security Requirements for Software Development
Tai-hoon Kim, Myong-chul Shin, Sang-ho Kim, Jae Sang Cha.........................116

Operations Research Based on Soft Computing

Intelligent Control Model of Information Appliances
Huey-Ming Lee, Ching-Hao Mao, Shu-Yen Lee........................................123

Effective Solution of a Portofolio Selection Based on a Block of Shares by a Meta-controlled Boltzmann Machine
Teruyuki Watanabe, Junzo Watada......................................................129

Soft Computing Approach to Books Allocation Strategy for Library
Junzo Watada, Keisuke Aoki, Takayuki Kawaura.................................136

Analysis of Human Feelings to Colors
Taki Kanda..........................................................143

Possibilistic Forecasting Model and Its Application to Analyze the Economy in Japan
Yoshiyuki Yabuuchi, Junzo Watada......................................................151

A Proposal of Chaotic Forecasting Method Based on Wavelet Transform
Yoshiyuki Matsumoto, Junzo Watada......................................................159

Fuzzy Multivariant Analysis
Junzo Watada, Masato Takagi, Jaeseok Choi........................................166
## Web Mining and Personalization

- Using Coherent Semantic Subpaths to Derive Emergent Semantics  
  D.V. Sreenath, W.I. Grosky, F. Fotouhi ............................................. 173

- Retrieval of Product Reputations from the WWW  
  Takahiro Hayashi, Yosuke Kinosita, Rikio Onai ................................... 180

- A Logic-Based Approach for Matching User Profiles  
  Andrea Calì, Diego Calvanese, Simona Colucci, Tommaso Di Noia,  
  Francesco M. Donini .......................................................... 187

## Learning and Soft Computing with Support Vector Machines (SVM) and RBF NNs

- Pose Classification of Car Occupant Using Stereovision and Support Vector Machines  
  Min-Soo Jang, Yong-Guk Kim, Hyun-Gu Lee, Byung-Joo Lee, Soek-Joo Lee,  
  Gwi-Tae Park .......................................................... 196

- A Fully Automatic System Recognizing Human Facial Expressions  
  Yong-Guk Kim, Sung-Oh Lee, Sang-Jun Kim, Gwi-Tae Park ......................... 203

- A Study of the Radial Basis Function Neural Network Classifiers Using Known Data of Varying Accuracy and Complexity  
  Patricia Crowther, Robert Cox, Dharmendra Sharma ................................ 210

## Novel Methods in Evolutionary Computation

- Top Down Modelling with Genetic Programming  
  Daniel Howard .......................................................... 217

- A Two Phase Genetic Programming Approach to Object Detection  
  Mengjie Zhang, Peter Andreae, Urvesh Bhowan ..................................... 224

- Mapping XML Schema to Relations Using Genetic Algorithm  
  Vincent Ng, Chan Chi Kong, Stephen Chan ......................................... 232

- Diagnosing the Population State in a Genetic Algorithm Using Hamming Distance  
  Radu Belea, Sergiu Caraman, Vasile Palade ......................................... 246

- Optimizing a Neural Tree Using Subtree Retraining  
  Wanida Pensuwon, Rod Adams, Neil Davey ......................................... 256
XXIV Table of Contents, Part III

Bioinformatics Using Intelligent and Machine Learning Techniques

Cluster Analysis of Gene Expression Profiles Using Automatically Extracted Seeds
   Miyoung Shin, Seon-Hee Park ................................................................. 263

Prediction of Plasma Membrane Spanning Region and Topology Using Hidden Markov Model and Neural Network
   Min Kyung Kim, Hyun Seok Park, Seon Hee Park ............................... 270

Speed Control and Torque Ripple Minimization in Switch Reluctance Motors Using Context Based Brain Emotional Learning
   Mehran Rashidi, Farzan Rashidi, Mohammad Hossein Aghdaei,
   Hamid Monavar ................................................................................. 278

Practical Common Sense Reasoning

Reasoning in Practical Situations
   Pei Wang ............................................................................................ 285

Commonsense Reasoning in and Over Natural Language
   Hugo Liu, Push Sing ............................................................................ 293

A Library of Behaviors: Implementing Commonsense Reasoning About Mental World
   Boris Galitsky ...................................................................................... 307

Handling Default Rules by Autistic Reasoning
   Don Peterson, Boris Galitsky ............................................................ 314

Systems for Large-scale Metadata Extraction and Maintenance

An Ontology-Driven Approach to Metadata Design in the Mining of Software Process Events
   Gabriele Gianini, Ernesto Damiani ..................................................... 321

Knowledge Extraction from Semi-structured Data Based on Fuzzy Techniques
   Paolo Ceravolo, Maria Cristina Nocerino, Marco Viviani ......................... 328

Managing Ontology Evolution Via Relational Constraints
   Paolo Ceravolo, Angelo Corallo, Gianluca Elia, Antonio Zilli ............... 335
### Table of Contents, Part III

#### Soft Computing in Fault Detection and Diagnosis

- **Using Design Information to Support Model-Based Fault Diagnosis Tasks**  
  Katsuaki Tanaka, Yoshikiyo Kato, Shin’ichi Nakasuka, Koichi Hori
  - Page 350

- **Fault Detection and Diagnosis Using the Fuzzy Min-Max Neural Network with Rule Extraction**  
  Kok Yeng Chen, Chee Peng Lim, Weng Kin Lai
  - Page 357

- **Refinement of the Diagnosis Process Performed with a Fuzzy Classifier**  
  C. D. Bocaniala, J. Sa da Costa, V. Palade
  - Page 365

- **ANN-Based Structural Damage Diagnosis Using Measured Vibration Data**  
  Eric W.M. Lee, H.F. Lam
  - Page 373

- **Induction Machine Diagnostic Using Adaptive Neuro Fuzzy Inferencing System**  
  Mohamad Shukri, Marzuki Khalid, Rubiyah Yusuf, Mohd Shafawi
  - Page 380

#### Intelligent Feature Recognition and Classification in Astrophysical and Medical Images

- **Real Time Stokes Inversion Using Multiple Support Vector Regression**  
  David Rees, Ying Guo, Arturo López Ariste, Jonathan Graham
  - Page 388

- **Extracting Stellar Population Parameters of Galaxies from Photometric Data Using Evolution Strategies and Locally Weighted Linear Regression**  
  Luis Alvarez, Olac Fuentes, Roberto Terlevich
  - Page 395

- **Using Evolution Strategies to Find a Dynamical Model of the M81 Triplet**  
  Juan Carlos Gomez, Olac Fuentes, Lia Athanassoula, Albert Bosma
  - Page 404

- **Automated Classification of Galaxy Images**  
  Jorge de la Calleja, Olac Fuentes
  - Page 411

- **Automatic Solar Flare Tracking**  
  Ming Qu, Frank Shih, Ju Jing, Haimin Wang, David Rees
  - Page 419

- **Source Separation Techniques Applied to Astrophysical Maps**  
  E. Salerno, A. Tonazzini, E. E. Kuruoğlu, L. Bedini, D. Herranz, C. Baccigalupi
  - Page 426

- **Counting Magnetic Bipoles on the Sun by Polarity Inversion**  
  Harrison P. Jones
  - Page 433
Correlation of the He I 1083 nm Line Width and Intensity as a Coronal Hole Identifier

Olена Malanushenko, Harrison P. Jones…………………………………………………………..439

Automated Recognition of Sunspots on the SOHO/MDI White Light Solar Images

S. Zharkov, V. Zharkova, S. Ipson, A. Benkhalil……………………………………………………446

A Procedure for the Automated Detection of Magnetic Field Inversion in SOHO MDI Magnetograms

S.S. Ipson, V.V. Zharkova, S.I. Zharkov, A. Benkhalil…………………………………………………453

Automatic Detection of Active Regions on Solar Images

A. Benkhalil, V. Zharkova, S. Ipson, S. Zharkov………………………………………………………….460

Automatic Detection of Solar Filaments Versus Manual Digitization

N. Fuller, J. Aboudarham…………………………………………………………………………………………..467

Adaptation of Shape Dendritic Spines by Genetic Algorithm

A. Herzog, V. Spravedlyvyy, K. Kube, E. Korkotian, K. Braun, E. Michaelis……..476

Detection of Dynamical Transitions in Biomedical Signals Using Nonlinear Methods

Patrick E. McSharry……………………………………………………………………………………………………..483

Applications of Machine Learning Concepts

On Retrieval of Lost Functions for Feedforward Neural Networks Using Re-Learning

Naotake Kamiura, Teijiro Isokawa, Kazuharu Yamato, Nobuyuki Matsui……..491

Analyzing the Temporal Sequences for Text Categorization

Xiao Luo, A. Nur Zincir-Heywood…………………………………………………………………………………………..498

Prediction of Women’s Apparel Sales Using Soft Computing Methods

Les M. Sztandera, Celia Frank, Balaji Vemulapali………………………………………………………………………506

A Try for Handling Uncertainties in Spatial Data Mining

Shuliang Wang, Guoqing Chen, Deyi Li, Deren Li, Hanning Yuan………………..513

Combining Evidence from Classifiers in Text Categorization

Yaxin Bi, David Bell, Jiwen Guan……………………………………………………………………………………………521

Predicting the Relationship Between the Size of Training Sample and the Predictive Power of Classifiers

Natthaphan Boonyanunta, Panlop Zeephongseku………………………………………………………………………529
Topographic Map Formation Employing kMER with Units Deletion Rule
Eiji Uchino, Noriaki Suetake, Chuhei Ishigaki .................................................. 536

Neuro-Fuzzy Hybrid Intelligent Industrial Control and Monitoring

Study on Weld Quality Control of Resistance Spot Welding Using a Neuro-Fuzzy Algorithm
Yansong Zhang, Guanlong Chen, Zhongqin Lin .................................................. 544

Exploring Benefits of Neuro Fuzzy Controller with Vehicle Health Monitoring
Preeti Bajaj, Avinash Keskar ................................................................. 551

Improvement of Low Frequency Oscillation Damping in Power Systems Via an Adaptive Critic Based NeuroFuzzy Controller
Farzan Rashidi, Behzad Moshidi ................................................................. 559

Use of Artificial Neural Networks in the Prediction of the Kidney Transplant Outcomes
Fariba Shadabi, Robert Cox, Dharmendra Sharma, Nikolai Petrovsky .................. 566

Intelligent Hybrid Systems for Robotics

An SoC-Based Context-Aware System Architecture
Keon Myung Lee, Bong Ki Sohn, Jong Tae Kim, Seung Wook Lee,
Ji Hyong Lee, Jae Wook Jeon, Jundong Cho .................................................. 573

An Intelligent Control of Chaos in Lorenz System with a Dynamic Wavelet Network
Yusuf Oysal ..................................................................................................... 581

Intelligent Robot Control with Personal Digital Assistants Using Fuzzy Logic and Neural Network
Seong-Joo Kim, Woo-Kyoung Choi, Hong-Tae Jeon ......................................... 589

Mobile Robot for Door Opening in a House
Dongwon Kim, Ju-Hyun Kang, Chang-Soon Hwang, Gwi-Tae Park ................... 596

Hybrid Fuzzy-Neural Architecture and Its Application to Time Series Modeling
Dongwon Kim, Sam-Jun Seo, Gwi-Tae Park .................................................... 603

Techniques of Computational Intelligence for Affective Computing

Accelerometer Signal Processing for User Activity Detection
Jonghun Baek, Geehyuk Lee, Wonbae Park, Byoung-Ju Yun ......................... 610
Neural Network Models for Product Image Design
Yang-Cheng Lin, Hsin-Hsi Lai, Chung-Hsing Yeh.................................618

Evaluation of Users’ Adaptation by Applying LZW Compression Algorithm to Operation Logs
Hiroshi Hayama, Kazuhiro Ueda.........................................................625

Study on Segmentation Algorithm for Unconstrained Handwritten Numeral Strings
Zhang Chuang, Wu Ming, Guo Jun.....................................................632

Information Agents on the Internet and Intelligent Web Mining

Wavelet-Based Image Watermaking Using the Genetic Algorithm
Prayoth Kumsawat, Kitti Attkitmongcol, Arthit Srikaew, Sarawut Sujitjorn......643

Extraction of Road Information from Guidance Map Images
Hirokazu Watabe, Tsukasa Kawaoka..................................................650

Dynamic Customer Profiling Architecture Using High Performance Computing
Qiubang Li, Rajiv Khosla, Chris Lai....................................................657

Intelligent Information Systems Using Case-Based Reasoning or Search Engineering

Predicting Business Failure with a Case-Based Reasoning Approach
Angela Y.N. Yip.................................................................................665

Capturing and Applying Lessons Learned During Engineering Equipment Installation
Ian Watson..........................................................................................672

Case-Based Adaptation for UML Diagram Reuse
Paulo Gomes, Francisco C. Pereira, Paulo Carreiro, Paulo Paiva, Nuno Seco, José L. Ferreira, Carlos Bento..................................................678

Harmonic Identification for Active Power Filters Via Adaptive Tabu Search Method
Thanatchai Kulworawanichpong, Kongpol Areerak, Kongpan Areerak, Sarawut Sujitjorn.................................................................687

Active Power Filter Design by a Simple Heuristic Search
Thanatchai Kulworawanichpong, Kongpol Areerak, Sarawut Sujitjorn.................................................................695

Stochastic Local Search for Incremental SAT and Incremental MAX-SAT
Malek Mouhoub, Changhai Wang.......................................................702
Finite Convergence and Performance Evaluation of Adaptive Tabu Search
Deacha Puangdownreong, Thanatchai Kulworawanichpong, Sarawut Sujitjorn .......................................................... 710

Applications of Computational Intelligence to Signal and Image Processing

Knowledge-Based Method to Recognize Objects in Geo-Images
Serguei Levachkine, Miguel Torres, Marco Moreno, Rolando Quintero .............. 718

Fast Design of 2-D Narrow Bandstop FIR Filters for Image Enhancement
Pavel Zahradnik, Miroslav Vlček .......................................................... 726

Fast Design of Optimal Comb FIR Filters
Pavel Zahradnik, Miroslav Vlček .......................................................... 733

Artificial Intelligence Methods in Diagnostics of the Pathological Speech Signals
Andrzej Izworski, Ryszard Tadeusiewicz, Wieslaw Wszolek .......................... 740

Intelligent Sub-patch Texture Synthesis Algorithm for Smart Camera
Jhing-Fa Wang, Han-Jen Hsu, Hong-Ming Wang ........................................ 749

Exploration of Image Features for Describing Visual Impressions of Black Fabrics
Chie Muraki Asano, Satoshi Hirakawa, Akira Asano .................................. 756

Emergent Global Behaviors of Distributed Intelligent Engineering and Information Systems

Distributed Resource Allocation via Local Choices: General Model and a Basic Solution
Marian F. Ursu, Botond Virginas, Chris Voudouris .................................... 764

Behavior Profiling Based on Psychological Data and Emotional States
Rajiv Khosla, Chris Lai, Tharanga Goonesekera ......................................... 772

Extension of Multiagent Data Mining for Distributed Databases
Ayahiko Niimi, Osamu Konishi ............................................................... 780

Agent-Based Approach to Conference Information Management
Hee-Seop Han, Jae-Bong Kim, Sun-Gwan Han, Hyeoncheol Kim ...................... 788

Mining Frequency Pattern from Mobile Users
John Goh, David Taniar ............................................................................. 795

Semi-supervised Learning from Unbalanced Labeled Data – An Improvement
Te Ming Huang, Vojislav Kecman ............................................................ 802
Posters

Handling Emergent Resource Use Oscillations
Mark Klein, Richard Metzler, Yaneer Bar-Yam

A Practical Timetabling Algorithm for College Lecture-Timetable Scheduling
Kyoung-Soon Hwang, Keon Myung Lee, Joongnam Jeon

Java Bytecode-to-.NET MSIL Translator for Construction of Platform Independent Information Systems
YangSun Lee, Seungwon Na

A Scale and Viewing Point Invariant Pose Estimation
M. Y. Nam, P. K. Rhee

A Novel Image Preprocessing by Evolvable Neural Network
M.Y. Nam, W.Y. Han, P.K. Rhee

Transition Properties of Higher Order Associative Memory of Sequential Patterns
Hiromi Miyajima, Noritaka Shigei, Yasuo Hamakawa

Morphological Blob-Mura Defect Detection Method for TFT-LCD Panel Inspection
Young-Chul Song, Doo-Hyun Choi, Kil-Houm Park

A Recommendation System for Intelligent User Interface: Collaborative Filtering Approach
Ju-Hyoung Yoo, Kye-Soon Ahn, Jeong Jun, Phill-Kyu Rhee

Fast Half Pixel Motion Estimation Based on the Spatial Correlation
Hyo Sun Yoon, Guee Sang Lee

A New Vertex Selection Scheme Using Curvature Information
Byoung-Ju Yun, Si-Woong Lee, Jae-Soo Cho, Jae Gark Choi, Hyun-Soo Kang

Author Index
# Table of Contents, Part I

## Keynote Lecturers

Web Intelligence, World Knowledge and Fuzzy Logic – The Concept of Web IQ (WIQ)  
*Loﬁ A. Zadeh*..................................................................................................................1

Industrial Applications of Evolvable Hardware  
*Tetsuya Higchi*................................................................................................................6

Equilibrium Modelling of Oligonucleotide Hybridization, Error, and Efficiency for DNA-Based Computational Systems  
*John A. Rose*..............................................................................................................8

Chance Discovery with Emergence of Future Scenarios  
*Yukio Ohsawa*................................................................................................................11

Brain-Inspired SOR Network and Its Application to Trailer Track Back-up Control  
*Takanori Koga, Takeshi Yamakawa*.............................................................................13

Dual Stream Artificial Neural Networks  
*Colin Fyfe*..................................................................................................................16

## Session Papers

### DNA-Based Semantic Information Processing

Improving the Quality of Semantic Retrieval in DNA-Based Memories with Learning  
*Andrew Neel, Max Garzon, Phani Penumatsa*...............................................................18

Conceptual and Contextual DNA-Based Memory  
*Russell Deaton, Junghuei Chen*..................................................................................25

Semantic Model for Artificial Intelligence Based on Molecular Computing  
*Yusei Tsuboi, Zuwairie Ibrahim, Osamu Ono*.............................................................32

The Fidelity of the Tag-Antitag System III. Robustness in the Excess Limit: The Stringent Temperature  
*John A. Rose*..............................................................................................................40
## Emergent Computational Intelligence Approaches – Artificial Immune Systems and DNA Computing

- Robust PID Controller Tuning Using Multiobjective Optimization Based on Clonal Selection of Immune Algorithm  
  *Dong Hwa Kim, Jae Hoon Cho* ................................................................. 50

- Intelligent Tuning of PID Controller With Robust Disturbance Rejection Function Using Immune Algorithm  
  *Dong Hwa Kim* .......................................................................................... 57

- The Block Hidden Markov Model for Biological Sequence Analysis  
  *Kyoung-Jae Won, Adam Prügel-Bennett, Anders Krogh* .......................... 64

## Innovations in Intelligent Agents and Their Applications

- Innovations in Intelligent Agents and Applications  
  *Gloria E. Phillips-Wren, Nikhil Ichalkaranje* ........................................... 71

- An Intelligent Aircraft Landing Support System  
  *Steve Thatcher, Lakhmi Jain, Colin Fyfe* .................................................. 74

- Teaming Humans and Agents in a Simulated World  
  *Christos Sioutis, Jeffrey Tweedale, Pierre Urlings, Nikhil Ichalkaranje, Lakhmi Jain* ................................................................. 80

- Contextual-Knowledge Management in Peer to Peer Computing  
  *E.V. Krishnamurthy, V.K. Murthy* ............................................................ 87

- Collaborating Agents in Distributed Networks and Emergence of Collective Knowledge  
  *V.K. Murthy, E.V. Krishnamurthy* ............................................................. 95

- Intelligent Decision Making in Information Retrieval  
  *Gloria E. Phillips-Wren, Guiseppi A. Forgionne* ........................................ 103

- Innovations in Intelligent Agents, Web and Their Applications  
  *Gloria E. Phillips-Wren, Nikhil Ichalkaranje* ............................................ 110

- Novel Intelligent Agent-Based System for Study of Trade  
  *Tomohiro Ikai, Mika Yoneyama, Yasuhiko Dote* ....................................... 113

- Testing of Multi-agent-based System in Ubiquitous Computing Environment  
  *Ken’ichi Takahashi, Satoshi Amamiya, Tadashige Iwao, Guoqiang Zhong, Makoto Amamiya* .......................................................... 124

- Helping Users Customize Their Pedagogical Agents: Issues, Approaches and Examples  
  *Anders I. Mørch, Jan Eirik B. Nævdal* ....................................................... 131
Intelligent Web Site: Understanding the Visitor Behavior
Juan D. Velásquez, Pablo A. Estévez, Hiroshi Yasuda, Terumasa Aoki, Eduardo Vera..................................................140

Data Mining and Knowledge Discovery

Mining Transformed Data Sets
Alex Burns, Andrew Kusiak, Terry Letsche...................................................148

Personalized Multilingual Web Content Mining
Rowena Chau, Chung-Hsing Yeh, Kate A. Smith........................................155

Intelligent Multimedia Information Retrieval for Identifying and Rating Adult Images
Seong-Joon Yoo............................................................................................164

Using Domain Knowledge to Learn from Heterogeneous Distributed Databases
Sally McClean, Bryan Scotney, Mary Shapcott........................................171

A Peer-to-Peer Approach to Parallel Association Rule Mining
Hiroshi Ishikawa, Yasuo Shioya, Takeshi Omi, Manabu Ohta, Karoru Katayama.................................................................178

FIT: A Fast Algorithm for Discovering Frequent Itemsets in Large Databases
Jun Luo, Sanguthevar Rajasekaran................................................................189

Frequency-Incorporated Interdependency Rules Mining in Spatiotemporal Databases
Ickjai Lee........................................................................................................196

Robotics: Intelligent Control and Sensing

Theoretical Considerations of Multiple Particle Filters for Simultaneous Localisation and Map-Building
David C.K. Yuen, Bruce A. MacDonald........................................................203

Continuous Walking Over Various Terrains – A Walking Control Algorithm for a 12-DOF Locomotion Interface
Jungwon Yoon, Jeha Ryu...............................................................................210

Vision Controlled Humanoid Robot Tool-Kit
Chris Messom..............................................................................................218

Modular Mechatronic Robotic Plug-and-Play Controller
Jonathan R. Zybalo, Glen Bright, Olaf Diegel, Johan Potgieter ............225

The Correspondence Problem in Topological Metric Mapping - Using Absolute Metric Maps to Close Cycles
Margaret E. Jefferies, Michael C. Cosgrove, Jesse T. Baker, Wai-Kiang Yeap...........................................................................232
Intelligent Tutoring Systems

Developing a “Virtual Student” Model to Test the Tutor and Optimizer Agents in an ITS
  Mircea Gh. Negoita, David Pritchard .................................................................240

Considering Different Learning Styles when Transferring Problem Solving Strategies from Expert to End Users
  Narin Mayiwar, Anne Håkansson .................................................................253

ULMM: A Uniform Logic Modeling Method in Intelligent Tutoring Systems
  Jinxin Si, Cungen Cao, Yuefei Sui, Xiaoli Yue, Nengfu Xie ............................263

Mining Positive and Negative Fuzzy Association Rules
  Peng Yan, Guoqing Chen, Chris Cornelis, Martine De Cock, Etienne Kerre .................................................................270

Intelligence and Technology in Educational Applications

An Adaptation Framework for Web Based Learning System
  T.T. Goh, Kinshuk .................................................................277

Ontologies for Creating Learning Object Content
  Dragan Gašević, Jelena Jovanović, Vladan Devedžić ..................................284

PASS: An Expert System with Certainty Factors for Predicting Student Success
  Ioannis Hatzilygeroudis, Anthi Karatrantou, C. Pierrakeas ...........................292

Student Modeling in Design Pattern ITS
  Zoran Jeremić, Vladan Devedžić .................................................................299

Supporting Self-Explanation in an Open-Ended Domain
  Amali Weerasinghe, Antonija Mitrovic ..........................................................306

Creativity Support Systems

Evaluation of the IRORI: A Cyber-Space that Catalyzes Face-to-Face Informal Communication
  Masao Usuki, Kozo Sugiyama, Kazushi Nishimoto, Takashi Matsubara ..........314

Information Sharing System Based on Location in Consideration of Privacy for Knowledge Creation
  Toshiyuki Hirata, Susumu Kunifuji .................................................................322

A Method of Extracting Topic Threads Towards Facilitating Knowledge Creation in Chat Conversations
  Kanayo Ogura, Masato Ishizaki, Kazushi Nishimoto .....................................330

Support Systems for a Person with Intellectual Handicap from the Viewpoint of Universal Design of Knowledge
  Toshiaki Ikeda, Susumu Kunifuji .................................................................337
# Table of Contents, Part I

## Intelligent Media Technology for Communicative Intelligence – Knowledge Management and Communication Model

Intelligent Conversational Channel for Learning Social Knowledge Among Communities

*S.M.F.D. Syed Mustapha* ................................................................. 343

An Algorithm for Avoiding Paradoxical Arguments Among the Multi-agent in the Discourse Communicator

*S.M.F.D. Syed Mustapha* ................................................................. 350

**Gallery: In Support of Human Memory**

*Hung-Hsuan Huang, Yasuyuki Sumi, Toyoaki Nishida* ........................................ 357

Evaluation of the Communication Atmosphere

*Tomasz M. Rutkowski, Koh Kakusho, Victor Kryssanov, Michihiko Minoh* ........... 364

A Method for Estimating Whether a User is in Smooth Communication with an Interactive Agent in Human-Agent Interaction

*Takanori Komatsu, Shoichiro Ohtsuka, Kazuhiro Ueda, Takashi Komeda, Natsuki Oka* ................................................................. 371

A Meaning Acquisition Model Which Induces and Utilizes Human’s Adaptation

*Atsushi Utsumoniya, Takanori Komatsu, Kazuhiro Ueda, Natsuki Oka* ................. 378

## Intelligent Media Technology for Communicative Intelligence – Interaction and Visual Content

Video Content Manipulation by Means of Content Annotation and Nonsymbolic Gestural Interfaces

*Burin Anuchitkittikul, Masashi Okamoto, Sadao Kurohashi, Toyoaki Nishida, Yoichi Sato* ................................................................. 385

Structural Analysis of Instruction Utterances Using Linguistic and Visual Information

*Tomohide Shibata, Masato Tachiki, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, Toyoaki Nishida* ........................................ 393

Video Contents Acquisition and Editing for Conversation Scene

*Takashi Nishizaki, Ryo Ogata, Yuichi Nakamura, Yuichi Ohta* ......................... 401

Video-Based Interactive Media for Gently Giving Instructions

*Takuya Kosaka, Yuichi Nakamura, Yoshinari Kameda, Yuichi Ohta* .................. 411

**Real-Time Human Proxy: An Avatar-Based Interaction System**

*Daisaku Arita, Rin-ichiro Taniguchi* ................................................ 419

## Soft Computing Techniques in the Capital Markets

Reliability and Convergence on Kohonen Maps: An Empirical Study

*Marcello Cattaneo Adorno, Marina Resta* ........................................... 426
A New Trial for Improving the Traditional Technical Analysis in the Stock Markets  
*Norio Baba, Tomoko Kawachi* .................................................................434

Prediction of Business Failure by Total Margin Support Vector Machines  
*Yeboon Yun, Min Yoon, Hirotaka Nakayama, Wataru Shiraki* .........................441

Tick-Wise Predictions of Foreign Exchange Rates  
*Mieko Tanaka-Yamawaki* ............................................................................449

**Knowledge-Based Systems for e-Business**

A Rule-Based System for eCommerce Applications  
*Jens Dietrich* ........................................................................................................455

Analyzing Dynamics of a Supply Chain Using Logic-Based Genetic Programming  
*Ken Taniguchi, Takao Terano* .........................................................................464

From Gaming Simulation to Case Method – Empirical Study on Business Game Development and Evaluation  
*Kenji Nakano, Takao Terano* ...........................................................................472

A Study of a Constructing Automatic Updating System for Government Web Pages  
*Keeiichiro Mitani, Yoshikatsu Fujita, Kazuhiko Tsuda* .................................480

Efficient Program Verification Using Binary Trees and Program Slicing  
*Masakazu Takahashi, Noriyoshi Mizukoshi, Kazuhiko Tsuda* ..........................487

An Efficient Learning System for Knowledge of Asset Management  
*Satoru Takahashi, Hiroshi Takahashi, Kazuhiko Tsuda* ....................................494

Extracting Purchase Patterns in Convenience Store E-Commerce Market Using Customer Cube Analysis  
*Yoshinori Fukue, Kessoku Masayuki, Kazuhiko Tsuda* ..................................501

A Study of Knowledge Extraction from Free Text Data in Customer Satisfaction Survey  
*Yukari Iseyama, Satoru Takahashi, Kazuhiko Tsuda* .......................................509

Network Information Mining for Content Delivery Route Control in P2P Network  
*Yoshikatsu Fujita, Jun Yoshida, Kenichi Yoshida, Kazuhiko Tsuda* ..................516

A Method of Customer Intention Management for a My-Page System  
*Masayuki Kessoku, Masakazu Takahashi, Kazuhiko Tsuda* ............................523

New Hierarchy Technique Using Co-occurrence Word Information  
*El-Sayed Atlam, Elmarhomy Ghada, Masao Fuketa, Kazuhiro Morita, Jun-ichi Aoe* .................................................................530
A New Method of Detecting Time Expressions for E-mail Messages
Toru Sumitomo, Yuki Kadoya, El-Sayed Atlam, Kazuhiro Morita, Shinkaku Kashiji, Jun-ichi Aoe ........................................................................541

A New Classification Method of Determining the Speaker’s Intention for Sentences in Conversation
Yuki Kadoya, El-Sayed Atlam, Kazuhiro Morita, Masao Fuketa, Toru Sumitomo, Jun-ichi Aoe ..............................................................549

A Fast Dynamic Method Using Memory Management
Shinkaku Kashiji, Toru Sumitomo, Kazuhiro Morita, Masaki Ono, Masao Fuketa, Jun-ichi Aoe ........................................................................558

A Method of Extracting and Evaluating Popularity and Unpopularity for Natural Language Expressions
Kazuhiro Morita, Yuki Kadoya, El-Sayed Atlam, Masao Fuketa, Shinkaku Kashiji, Jun-ichi Aoe ........................................................................567

Intelligent Hybrid Systems for Medical Diagnosis

Evaluating a Case-Based Reasoner for Clinical Decision Support
Anna Wills, Ian Watson .............................................................................575

Early Detection of Breast Cancer Using Mathematical Morphology
Özgür Özsen .............................................................................................583

Diagnosis of Cervical Cancer Using Hybrid Multilayered Perceptron (HMLP) Network
Dzati Athiar Ramli, Ahmad Fauzan Kadmin, Mohd. Yousoff Mashor, Nor Ashidi, Mat Isa .................................................................591

Mammographic Image and Breast Ultrasound Based Expert System for Breast Diseases
Umi Kalthum Ngah, Chan Choyi Ping, Shalihatun Azlin Aziz ...................599

A Study on Nonparametric Classifiers for a CAD System of Diffuse Lung Opacities in Thin-Section Computed Tomography Images
Yoshihiro Mitani, Yusuke Fujita, Naofumi Matsunaga, Yoshihiko Hamamoto ..................................................................................608

Techniques of Computational Intelligence for Web Applications

Recognition of Grouping Areas in Trademarks Considering Proximity and Shape Similarity
Koji Abe, Debabrata Roy, John P. Eakins .................................................614

Multidimensional Visualization and Navigation in Search Results
Will Archer Arentz, Aleksander Øhrn ......................................................620
A Hybrid Learning Approach for TV Program Personalization  
Zhiwen Yu, Xingshe Zhou, Zhiyi Yang ................................................................. 630

An Adaptive-Learning Distributed File System  
Joseph D. Gradecki, Ilkeyun Ra ............................................................... 637

**Intelligent Information Processing for Remote Sensing**

Review of Coding Techniques Applied to Remote Sensing  
Joan Serra-Sagrista, Francesc Auli, Fernando Garcia, Jorge Gonzales,  
Pere Guitart ............................................................................................................. 647

Efficient and Effective Tropical Cyclone Eye Fix Using Genetic Algorithms  
Chi Lap Yip, Ka Yan Wong ................................................................................... 654

Spectral Unmixing Through Gaussian Synapse ANNs in Hyperspectral Images  
J.L. Crespo, R.J. Duro, F. López-Peña ................................................................. 661

A Hyperspectral Based Multisensor System for Marine Oil Spill Detection, Analysis and Tracking  
F. López-Peña, R.J. Duro ....................................................................................... 669

Some Experiments on Ensembles of Neural Networks for Hyperspectral Image Classification  
Carlos Hernández-Espinosa, Mercedes Fernández-Redondo,  
Joaquín Torres Sospedra ..................................................................................... 677

A Modular Approach to Real-Time Sensorial Fusion Systems  
F. Gil-Castiñeira, P.S. Rodríguez-Hernández, F.J. Gonzáles-Castaño,  
E. Costa-Montenegro, R. Asorey-Cacheda, J.M. Pousada Carballo .................. 685

Feature Extraction by Linear Spectral Unmixing  
M. Graña, A. D’Anjou ............................................................................................ 692

**Intelligent and Knowledge-Based Solutions for Mobile and Ad-Hoc Networks**

Decision Support System on the Grid  
M. Ong, X. Ren, J. Allan, V. Kadirkamanathan, HA Thompson, PJ Fleming ...... 699

Representing Knowledge in Controlled Natural Language: A Case Study  
Rolf Schwitter ....................................................................................................... 711

Supporting Smart Applications in Multihop Ad-Hoc Networks - The GecGo Middleware -  
Peter Sturm, Hannes Frey, Daniel Görgen, Johannes Lehner ......................... 718

A Heuristic for Efficient Broadcasting in the Metropolitan Ad hoc Networks  
Luc Hogie, Frederic Guinand, Pascal Bouvry .................................................... 727

ADS as Information Management Service in an M-Learning Environment  
Matthias R. Brust, Daniel Görgen, Christian Hutter, Steffen Rothkugel ...... 734
# Rough Sets - Theory and Applications

| Title                                                                 | Authors                                      | Page |
|----------------------------------------------------------------------|----------------------------------------------|------|
| Noise Reduction in Audio Employing Spectral Unpredictability Measure and Neural Net | Andrzej Czyzewski, Marek Dziubinski           | 743  |
| Forming and Ranking Musical Rhythm Hypotheses                        | Bozena Kostek, Jaroslaw Wojcik               | 750  |
| A Comparison of Two Approaches to Data Mining from Imbalanced Data    | Jerzy W. Grzymala-Busse, Jerzy Stefanowski, Szymon Wilk | 757  |
| Measuring Acceptance of Intelligent System Models                     | James F. Peters, Sheela Ramanna             | 764  |
| Rough Set Based Image Texture Recognition Algorithm                   | Zheng Zheng, Hong Hu, Zhongzhi Shi           | 772  |
| Sets of Communicating Sequential Processes. A Topological Rough Set Framework | L. Polkowski, M. Serneniuk-Polkowska         | 779  |

# Soft Computing Techniques and Their Applications

| Title                                                                 | Authors                                      | Page |
|----------------------------------------------------------------------|----------------------------------------------|------|
| Robust System Identification Using Neural Networks                    | Shigenobu Yamawaki, Lakhmi Jain              | 786  |
| A Consideration on the Learning Behaviors of the HSLA Under the Nonstationary Multiteacher Environment and Their Application to Simulation and Gaming | Norio Baba, Yoshio Mogami                | 792  |
| Genetic Lips Extraction Method with Flexible Search Domain Control   | Takuya Akashi, Minoru Fukumi, Norio Akamatsu | 799  |
| Medical Diagnosis System Using the Intelligent Fuzzy Systems          | Yasue Mitsukura, Kensuke Mitsukura, Minoru Fukumi, Norio Akamatsu, Witold Pedrycz | 807  |
| Music Compression System Using the GA                                 | Hiroshi Kawasaki, Yasue Mitsukura, Kensuke Mitsukura, Minoru Fukumi, Norio Akamatsu | 827  |
| Effects of Chaotic Exploration on Reinforcement Maze Learning         | Koichiro Morihiro, Nobuyuki Matsui, Haruhiko Nishimura | 833  |
| Face Search by Neural Network Based Skin Color Threshold Method       | Takashi Imura, Minoru Fukumi, Norio Akamatsu, Kazuhiro Nakaura | 840  |
| Face Edge Detection System by Using the GAs                           | Hideaki Sato, Katsuhiro Sakamoto, Yasue Mitsukura, Norio Akamatsu | 847  |
| A Feature Extraction of EEG with Individual Characteristics          | Shin-ichi Ito, Yasue Mitsukura, Norio Akamatsu | 853  |
Proposal of Neural Recognition with Gaussian Function and Discussion for Rejection Capabilities to Unknown Currencies
   Baiqing Sun, Fumiaki Takeda ................................................................. 859

Development of DSP Unit for Online Tuning and Application to Neural Pattern Recognition System
   Hironobu Satoh, Fumiaki Takeda ................................................................. 866

Face Identification Based on Ellipse Parameter Independent of Varying Facial Pose and Lighting Condition
   Hironori Takimoto, Yasue Mitsukura, Norio Akamatsu.............................. 874

Object Extraction System by Using the Evolutionaly Computations
   Seiki Yoshimori, Yasue Mitsukura, Minoru Fukumi, Norio Akamatsu.............. 881

Wrist EMG Pattern Recognition System by Neural Networks and Multiple Principal Component Analysis
   Yuji Matsumura, Minoru Fukumi, Norio Akamatsu, Fumiaki Takeda ............... 891

Age Classification from Face Images Focusing on Edge Information
   Miyoko Nakano, Fumiko Yasukata, Minoru Fukumi.................................. 898

Evolutionary Computation in the Soft Computing Framework

Why Do Machine Learning Based Techniques Fail to Accelerate the Evolution of Neural Networks?
   Hugo de Garis, Thayne Batty................................................................. 905

An Optimiser Agent that Empowers an ITS System to “on-the-fly” Modify Its Teaching Strategies
   Mircea Gh. Negoita, David Pritchard..................................................... 914

A Constraint-Based Optimization Mechanism for Patient Satisfaction
   Chi-I Hsu, Chaochang Chiu, Pei-Lun Hsu............................................... 922

Optimizing Beam Pattern of Adaptively Linear Array Antenna by Phase Perturbations Using Genetic Algorithms
   Chao-Hsing Hsu, Chun-Hua Chen........................................................... 929

The Optimal Airline Overbooking Strategy Under Uncertainties
   Chaochang Chiu, Chanhsi Tsao............................................................. 937

Determination of Packet Priority by Genetic Algorithm in the Packet Switching Networks
   Taner Tuncer, Ali Karci ................................................................. 946

A New Encoding for the Degree Constrained Minimum Spanning Tree Problem
   Sang-Moon Soak, David Corne, Byung-Ha Ahn...................................... 952
| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|-------------------------------------------------------------------------|------|
| Neurodynamics and Its Hardware Implementation                         | Towards Cortex Sized Artificial Nervous Systems                          | 959  |
|                                                                      | Christopher Johansson, Anders Lansner                                     |      |
|                                                                      | A Memory Model Based on Dynamical Behaviour of the Hippocampus           | 967  |
|                                                                      | Hatsuo Hayashi, Motoharu Yoshida                                         |      |
|                                                                      | Analysis of Limit-Cycles on Neural Networks with Asymmetrical Cyclic     | 974  |
|                                                                      | Connections Using Approximately Activation Functions                     |      |
|                                                                      | Shinya Suenaga, Yoshihiro Hayakawa, Koji Nakajima                        |      |
|                                                                      | Inverse Function Delayed Model for Optimization Problems                 | 981  |
|                                                                      | Yoshihiro Hayakawa, Tatsuaki Denda, Koji Nakajima                        |      |
|                                                                      | Switched-Capacitor Large-Scale Chaotic Neuro-Computer Prototype and Chaotic| 988  |
|                                                                      | Search Dynamics                                                          |      |
|                                                                      | Yoshihiko Horio, Takahide Okuno, Koji Mori                               |      |
|                                                                      | A Convolutional Neural Network VLSI Architecture Using Thresholding and   | 995  |
|                                                                      | Weight Decomposition                                                     |      |
|                                                                      | Osamu Nomura, Takashi Morie, Keisuke Korekado, Masakazu Matsugu, Atsushi |      |
|                                                                      | Iwata                                                                    |      |
|                                                                      | Pulse Codings of a Spiking Neuron Having Quantized State                 | 1002 |
|                                                                      | Hiroyuki Torikai, Hiroshi Hamanaka, Toshimichi Saito                      |      |
|                                                                      | Design of Single Electron Circuitry for a Stochastic Logic Neural Network | 1010 |
|                                                                      | Hisanao Akima, Shigeo Sato, Koji Nakajima                               |      |
| Advances, in Design, Analysis and Applications of Neural/Neuro-Fuzzy   | An Improved Time Series Prediction Scheme Using Fuzzy Logic Inference     | 1017 |
|                                                                      | Bin Qiu, Xiaoxiang Guan                                                  |      |
|                                                                      | Fuzzy Classification of Secretory Signals in Proteins Encoded by the     | 1023 |
|                                                                      | Plasmodium falciparum Genome                                             |      |
|                                                                      | Erica Logan, Richard Hall, Nectarios Klonis, Susanna Herd, Leann Tilley  |      |
|                                                                      | Web Users’ Classification Using Fuzzy Neural Network                     | 1030 |
|                                                                      | Fang Yuan, Huanrui Wu, Ge Yu                                             |      |
|                                                                      | Enhancing Generalization Capability of SVM Classifiers with Feature Weight| 1037 |
|                                                                      | Xizhao Wang, Qiang He                                                    |      |
|                                                                      | GREN-Networks in WDI-Based Analysis of State Economies                   | 1044 |
|                                                                      | Iveta Mrázová                                                            |      |
|                                                                      | Learning Pseudo Metric for Multimedia Data Classification and Retrieval  | 1051 |
|                                                                      | Dianhui Wang, Xiaohang Ma                                                |      |
| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|-------------------------------------------------------------------------|------|
| Several Aspects in Ubiquitous Pattern Recognition Techniques          | Akira Tanaka, Ichigaku Takigawa, Hideyuki Imai, Mineichi Kudo, Masaaki Miyakoshi | 1058 |
| Projection Learning Based Kernel Machine Design Using Series of Monotone Increasing Reproducing Kernel Hilbert Spaces | Akira Tanaka, Ichigaku Takigawa, Hideyuki Imai, Mineichi Kudo, Masaaki Miyakoshi | 1065 |
| Combination of Weak Evidences by D-S Theory for Person Recognition    | Masafumi Yamada, Mineichi Kudo                                          | 1072 |
| Time-Frequency Decomposition in Gesture Recognition System Using Accelerometer | Hidetoshi Nonaka, Masahito Kurihara                                    | 1079 |
| A Method of Belief Base Revision for Extended Logic Programs Based on State Transition Diagrams | Yasuo Kudo, Tetsuya Murai                                              | 1085 |
| Monotonic and Nonmonotonic Reasoning in Zoom Reasoning Systems        | Tetsuya Murai, M. Sanada, Yasuo Kudo, Y. Sato                          | 1092 |
| Interaction and Intelligence                                          |                                                                        |      |
| An Exoskeleton for Human Shoulder Rotation Motion Assist              | Kazuo Kiguchi                                                          |      |
| Networked Intelligent Robots by Ontological Neural Networks           | Eri Sato, Jun Kawakatsu, Toru Yamaguchi                                |      |
| Some Emergences of Mobilgence in the Pursuit Game                    | Seiichi Kawata, Kazuya Morohashi, Takeshi Tateyama                     |      |
| Use of Successful Policies to Relearn for Induced States of Failure in Reinforcement Learning | Tadahiko Murata, Hiroshi Matsumoto                                     |      |
| A Perceptual System for a Vision-Based Mobile Robot Under Office Automation Floors | Naoyuki Kubota, Kazuhiko Taniguchi, Atsushi Ueda                       |      |
| Performance Evaluation of a Distributed Genetic Algorithm with Cellular Structures on Function Optimization Problems | Tadahiko Murata, Kenji Takada                                         |      |
| New Development, Trends and Applications of Intelligent Multi-Agent Systems |                                                        |      |
| On-Line Update of Situation Assessment Based on Asynchronous Data Streams | Vladimir Gorodetsky, Oleg Kasaev, Vladimir Samoilov                     |      |
| Mobility Management for Personal Agents in the All-mobile Network    | Ignac Lovrek, Vjekoslav Sinkovic                                        |      |
| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|------------------------------------------------------------------------|------|
| A Multi-agent Perspective on Data Integration Architectural Design   | Stéphane Faulkner, Manuel Kolp, Tai Nguyen, Adrien Coyette             | 1150 |
| Identification of Structural Characteristics in Product Spectra      | Maik Maurer, Udo Lindemann                                             | 1157 |
| Policies, Rules and Their Engines: What do They Mean for SLAs?       | Mark Perry, Michael Bauer                                              | 1164 |
| Forecasting on Complex Datasets with Association Rules               | Marcello Bertoli, Andrew Stranieri                                     | 1171 |
| Using a Multi-agent Architecture to Manage Knowledge in the Software | Oscar M. Rodríguez, Aurora Vizcaíno, Ana I. Martínez, Mario Piattini, Jesús Favela | 1181 |

### Engineering Techniques and Developments of Intelligent Systems

| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|------------------------------------------------------------------------|------|
| Evolution Strategies Based Particle Filters for Nonlinear State Estimation | Katsuji Uosaki, Yuuya Kimura, Toshiharu Hatanaka                       | 1189 |
| Coordination in Multiagent Reinforcement Learning Systems            | M.A.S. Kamal, Junichi Murata                                           | 1197 |
| Measurement of Shaft Vibration Using Ultrasonic Sensor in Sump Pump Systems | Shogo Tanaka, Hajime Morishige                                         | 1205 |
| Behavior Learning of Autonomous Agents in Continuous State Using Function Approximation | Min-Kyu Shon, Junichi Murata                                           | 1213 |
| Some Experiences with Change Detection in Dynamical Systems          | Theodor D. Popescu                                                     | 1220 |

### Computational Intelligence for Fault Diagnosis

| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|------------------------------------------------------------------------|------|
| The KAMET II Approach for Knowledge-Based System Construction        | Osvaldo Cairó, Julio César Alvarez                                     | 1227 |
| A Recursive Component Boundary Algorithm to Reduce Recovery Time for Microreboots | Chanwit Kaewkasi, Pitchaya Kaewkasi                                    | 1235 |
| Electric Power System Anomaly Detection Using Neural Networks        | Marco Martinelli, Enrico Tronci, Giovanni Dipoppa, Claudio Balducelli | 1242 |
| Capturing and Applying Lessons Learned During Engineering Equipment Installation | Ian Watson                                                            | 1249 |
Moving Towards a New Era of Intelligent Protection Through Digital Relaying in Power Systems
   Kongpan Areerak, Thanatchai Kulworawanichpong, Sarawut Sujitjorn.........1255

Capacitor Switching Control Using a Decision Table for a 115-kV Power Transmission System in Thailand
   Phinit Srithorn, Kasem Khojulklang, Thanatchai Kulworawanichpong.........1262

Author Index..........................................................1269
# Table of Contents, Part II

## Methods of Computational Intelligence with Applications for Product Development and Human Resource Recruitment

Integration of Psychology, Artificial Intelligence and Soft Computing for Recruitment and Benchmarking of Salespersons  
*Rajiv Khosla, Tharanga Goonesekera*  

1

FHP: Functional Heuristic Planning  
*Joseph Zalaket, Guy Camilleri*  

9

Planning with Recursive Subgoals  
*Han Yu, Dan C. Marinescu, Annie S. Wu, Howard Jay Siegel*  

17

Development of a Generic Computer Aided Deductive Algorithm for Process Parameter Design  
*K.P. Cheng, Daniel C.Y. Yip, K.H. Lau, Stuart Barnes*  

28

Epistemic Logic and Planning  
*Shahin Maghsoudi, Ian Watson*  

36

Tátari: An Open Source Software Tool for the Development and Evaluation of Recommender System Algorithms  
*Halal Hassan, Ian Watson*  

46

DCPP: Knowledge Representation for Planning Processes  
*Takushi Tanaka, Koki Tanaka*  

53

An IS Framework to Support the Collaborative Design of Supply Chains  
*Nikos Karacapilidis, Emmanuel Adamides, Costas P. Pappis*  

62

## Knowledge-Based Interface Systems

A New Similarity Evaluation Function for Writer Recognition of Chinese Character  
*Yoshinori Adachi, Min Liu, Masahiro Ozaki*  

71

Development of Teaching Materials Which Dynamically Change in Learning Process  
*Masahiro Ozaki, Koji Koyama, Saori Takeoka, Yoshinori Adachi*  

77

Analog VLSI Layout Design of Motion Detection for Artificial Vision Model  
*Masashi Kawaguchi, Takashi Jimbo, Masayoshi Umeno, Naohiro Ishii*  

83

Development of High-Precise and No-Contacting Capacitance Measuring System Using Dipmeter  
*Shoji Suzuki, Yoshinori Adachi*  

89

Similarity of Documents Using Reconfiguration of Thesaurus  
*Tomoya Ogawa, Nobuhiko Inuzuka*  

95
| Title                                                                 | Page |
|----------------------------------------------------------------------|------|
| On Refractory Parameter of Chaotic Neurons in Incremental Learning   | 103  |
| Toshinori Deguchi, Naohiro Ishii                                      |      |
| Automatic Virtualization of Real Object Based on Shape Knowledge in   | 110  |
| Mixed Reality                                                          |      |
| Kenji Funahashi, Kazunari Komura, Yuji Iwahori, Yukie Koyama          |      |
| Generation of Virtual Image from Multiple View Point Image Database    | 118  |
| Haruki Kawanaka, Nobuaki Sado, Yuji Iwahori                           |      |
| Correlation Computations for Movement Detection in Neural Networks     | 124  |
| Naohiro Ishii, Masahiro Ozaki, Hiroshi Sasaki                         |      |
| **Intelligent Human Computer Interaction Systems**                     |      |
| Information Acquisition Using Chat Environment for Question Answering | 131  |
| Calkin A.S. Montero, Kenji Araki                                     |      |
| Design and Implementation of Natural Language Interface for Impression-| 139  |
| Based Music-Retrieval Systems                                          |      |
| Tadahiko Kumamoto                                                     |      |
| InTREND: An Interactive Tool for Reflective Data Exploration Through  | 148  |
| Natural Discourse                                                     |      |
| Mitsunori Matsushita, Kumiyo Nakaoji, Yasuhiro Yamamoto, Tsuneaki    |      |
| Kato                                                                   |      |
| Using Mitate-shi Related to the CONTAINER Schema for Detecting the    | 156  |
| Container-for-Contents Metonymy                                        |      |
| Yoshiaki Kurosawa, Takumi Ichimura, Teruaki Aizawa                    |      |
| Character Learning System Using Inter-stroke Information              | 165  |
| Jungpil Shin, Aisushi Takeda                                           |      |
| Construction of Conscous Model Using Reinforcement Learning           | 175  |
| Masafumi Kozuma, Hirokazu Taki, Noriyuki Matsuda, Hirokazu Miura,     |      |
| Satoshi Hori, Norihiro Abe                                           |      |
| Advice Recording Method for a Lesson with Computers                   | 181  |
| Katsuyuki Harada, Noriyuki Matsuda, Hirokazu Miura, Hirokazu Taki,    |      |
| Satoshi Hori, Norihiro Abe                                           |      |
| Acquiring After-Sales Knowledge from Human Motions                    | 188  |
| Satoshi Hori, Kota Hirose, Hirokazu Taki                              |      |
| Emotion Analyzing Method Using Physiological State                    | 195  |
| Kazuya Mera, Takumi Ichimura                                         |      |
| **Posters**                                                           |      |
| A Lyapunov Function Based Direct Model Reference Adaptive Fuzzy Control| 202  |
| Youngwan Cho, Yangsun Lee, Kwangyup Lee, Euntai Kim                   |      |
Table of Contents, Part II

Semi-automatic Video Object Segmentation Method Based on User Assistance and Object Tracking
  J. G. Choi, S. W. Lee, B. J. Yun, H. S. Kang, S. H. Hong, J. Y. Nam.........................211

Design and Evaluation of a Scale Patching Technique for VOD Servers
  Hyo-Young Lee, Sook-Jeong Ha, Sun-Jin Oh, Ihn-Han Bae........................................219

Optimal Gabor Encoding Scheme for Face Recognition Using Genetic Algorithm
  Inja Jeon, Kisang Kwon, Phill-Kyu Rhee.................................................................227

T-shape Diamond Search Pattern for New Fast Block Matching Motion Estimation
  Mi Gyoung Jung, Mi Young Kim..................................................................................237

Motion Estimation Using Cross Center-Biased Distribution and Spatio-Temporal Correlation of Motion Vector
  Mi Young Kim, Mi Gyoung Jung...............................................................................244

A Fast Motion Estimation Using Prediction of Motion Estimation Error
  Hyun-Soo Kang, Seong-Mo Park, Si-Woong Lee, Jae-Gark Choi, Byoungh-Ju Yun........253

Ontology Revision Using the Concept of Belief Revision
  Seung Hwan Kang, Sim Kim Lau..................................................................................261

Novelty in the Generation of Initial Population for Genetic Algorithms
  Ali Karci......................................................................................................................268

Framework for Personalized e-Mediator
  Dong-Hwee Kim, Soon-Ja Kim......................................................................................276

Advances in Intelligent Data Processing Techniques and Applications

Weightless Neural Networks for Typing Biometrics Authentication
  Shereen Yong, Weng Kin Lai, George Goghill..........................................................284

Intelligent Pressure-Based Typing Biometrics System
  Azweeda Dahalan, M.J.E. Salami, W.K. Lai, Ahmad Faris Ismail.................................294

Classifiers for Sonar Target Differentiation
  C.K. Loo, W.S. Lim, M.V.C. Rao................................................................................305

Design and Development of Intelligent Fingerprint-Based Security System
  Suriza Ahmad Zabidi, Momoh-Jimoh E. Salami.........................................................312

Weightless Neural Networks: A Comparison Between the Discriminator and the Deterministic Adaptive RAM Network
  Paul Yee, George Coghill............................................................................................319

Extracting Biochemical Reaction Kinetics from Time Series Data
  Edmund J. Crampin, Patrick E. McSharry, Santiago Schnell.......................................329
PCA and ICA Based Signal and Image Processing

Image Feature Representation by the Subspace of Nonlinear PCA  
Yen-Wei Chen, Xiang-Yan Zeng ..........................................................337

Improving ICA Performance for Modeling Image Appearance with the Kernel Trick  
Qingshan Liu, Jian Cheng, Hangqin Lu, Songde Ma ................................344

Random Independent Subspace for Face Recognition  
Jian Cheng, Qingshan Liu, Hangqing Lu, Yen-Wei Chen .......................352

An RDWT Based Logo Watermark Embedding Scheme with Independent Component Analysis Detection  
Thai Duy Hien, Zensho Nakao, Yen-Wei Chen ....................................359

Real-Time Independent Component Analysis Based on Gradient Learning with Simultaneous Perturbation Stochastic Approximation  
Shuxue Ding, Jie Huang, Daming Wei, Sadao Omata ............................366

Intelligent Data Processing in Process Systems and Plants

Extraction Operation Know-How from Historical Operation Data – Using Characterization Method of Time Series Data and Data Mining Method –  
Kazuhiro Takeda, Yoshifumi Tsuge, Hisayoshi Matsuyama ........................375

Handling Qualitative Aspects of Human Knowledge in Diagnosis  
Viorel Ariton ......................................................................................382

Qualitative Analysis for Detection of Stiction in Control Valves  
Yoshiyuki Yamashita ........................................................................391

Agent-Based Batch Process Control Systems  
Masaru Sakamoto, Hajime Eguchi, Takashi Hamaguchi, Yutaka Ota, Yoshinhiro Hashimoto, Toshiaki Itoh ................................................398

Acquisition of AGV Control Rules Using Profit Sharing Method and Evaluation of the Rules  
Hisaaki Yamaba, Hitoshi Yoshioka, Shigeyuki Tomita ............................405

Dynamic Acquisition of Models for Multiagent-Oriented Simulation of Micro Chemical Processes  
Naoki Kimura, Hideyuki Matsumoto, Chiaki Kuroda ..............................412

Acquisition of Engineering Knowledge on Design of Industrial Cleaning System through IDEF0 Activity Model  
Tetsuo Fuchino, Takao Wada, Masahiko Hirao ......................................418

Intelligent Systems for Spatial Information Processing and Imaging

Exchanging Generalized Maps Across the Internet  
Min Zhou, Michela Bertolotto ..............................................................425
Adaptive Spatial Data Processing System (ASDPS)

Wanwu Guo ............................................................................................................432

Modified ASDPS for Geochemical Data Processing

Chi Liu, Hui Yu .......................................................................................................440

Gravity Data Processing Using ASDPS

Kai Ding, Baishan Xu .............................................................................................447

Remote Sensing Image Processing Using MCDF

Zhiqiang Ma, Wanwu Guo .....................................................................................454

Coarse-Grained Parallel Algorithms for Spatial Data Partition and Join Processing

Jitian Xiao ..............................................................................................................461

Image Processing and Intelligent Information Applications

Multi-agents for Decision Support

Manoj Achuthan, Bala Balachandran, Dharmendra Sharma ...............................469

Dynamic Scheduling Using Multiagent Architecture

Dharmendra Sharma, Dat Tran ..................................................................................476

Using Consensus Ensembles to Identify Suspect Data

David Clark ..............................................................................................................483

Fuzzy Analysis of X-Ray Images for Automated Disease Examination

Craig Watman, Kim Le ...............................................................................................491

New Background Speaker Models and Experiments on the ANDOSL Speech Corpus

Dat Tran, Dharmendra Sharma ..................................................................................498

Immunity-Based Systems and Approaches

An Approach for Self-repair in Distributed System Using Immunity-Based Diagnostic Mobile Agents

Yuji Watanabe, Shigeyuki Sato, Yoshiteru Ishida .........................................................504

Artificial Immune System for Personal Identiﬁcation with Finger Vein Pattern

Toshiyuki Shimooka, Koichi Shimizu ........................................................................511

A Switching Memory Strategy in an Immune Network Model

Kouji Harada .............................................................................................................519

A Process Algebra Model of the Immune System

Raúl Monroy ..............................................................................................................526

Mechanism for Generating Immunity-Based Agents that Detect Masqueraders

Takeshi Okamoto, Takayuki Watanabe, Yoshiteru Ishida ...........................................534
Machine and Computer Vision, Neural Networks, Intelligent Web Mining and Applications

False Alarm Filter in Neural Networks for Multiclass Object Detection
   Mengjie Zhang, Bunna Ny  .................................................................541

iJADE Scene Segmentator – A Real-Time Scene Segmentation System Using Watershed-Based Neuro-Oscillatory Network
   Gary C.L. Li, Raymond S.T. Lee ..........................................................549

Visual Tracking by Using Kalman Gradient Vector Flow (KGVF) Snakes
   Toby H.W. Lam, Raymond S.T. Lee ......................................................557

Chart Patterns Recognition and Forecast Using Wavelet and Radial Basis Function Network
   Jamec N.K. Liu, Raymond W.M. Kwong, Feng Bo .................................564

Appearance-Based Face Recognition Using Aggregated 2D Gabor Features
   King Hong Cheung, Jane You, James Liu, Tony W.H. Ao Ieong .............572

Ontology-Based Web Agents Using Concept Description Flow
   Nengfu Xie, Cungen Cao, Bingxian Ma, Chunxia Zhang, Jinxin Si ...........580

Web Page Recommendation Model for Web Personalization
   Abdul Manan Ahmad, Mohd. Hanafi Ahmad Hijazi .................................587

iJADE Face Recognizer - A Multi-agent Based Pose and Scale Invariant Human Face Recognition System
   Tony W.H. Ao Ieong, Raymond S.T. Lee .............................................594

Neural Networks for Data Mining

Piecewise Multivariate Polynomials Using a Four-Layer Perceptron
   Yusuke Tanahashi, Kazumi Saito, Ryohei Nakano .................................602

Learning an Evaluation Function for Shogi from Data of Games
   Satoshi Tanimoto, Ryohei Nakano .....................................................609

Extended Parametric Mixture Model for Robust Multi-labeled Text Categorization
   Yuji Kaneda, Naonori Ueda, Kazumi Saito ............................................616

Visualisation of Anomaly Using Mixture Model
   Tomoharu Iwata, Kazumi Saito ............................................................624

Obtaining Shape from Scanning Electron Microscope Using Hopfield Neural Network
   Yuji Iwahori, Haruki Kawanaka, Shinji Fukui, Kenji Funahashi ...............632
Neural Networks as Universal Approximators and Paradigms for Information Processing – Theoretical Developments and Applications

Speech Recognition for Emotions with Neural Network: A Design Approach
Shubhangi Giripunje, Anshish Panat.................................................................640

Neuro-Genetic Approach for Bankruptcy Prediction Modeling
Kyung-shik Shin, Kyoung Jun Lee.................................................................646

Design of a Robust and Adaptive Wavelet Neural Network for Control of Three Phase Boost Rectifiers
Farzan Rashidi, Mehran Rashidi.................................................................653

The Comparison of Characteristics of 2-DOF PID Controllers and Intelligent Tuning of a Gas Turbine Generating Plant
Dong Hwa Kim............................................................................................661

Bankruptcy Prediction Modeling Using Multiple Neural Network Models
Kyung-shik Shin, Kyoung Jun Lee.................................................................668

Interpreting the Output of Certain Neural Networks as Almost Unique Probability
Bernd-Jürgen Falkowski..............................................................................675

A Stochastic Model of Neural Computing
Paolo Crippa, Claudio Turchetti, Massimiliano Pirani...............................683

Theoretical Developments and Applications of Fuzzy Techniques and Systems

Classification of Fuzzy Data in Database Management System
Deval Popat, Hema Sharda, David Taniar......................................................691

An Efficient Fuzzy Method for Handwritten Character Recognition
Romesh Ranawana, Vasile Palade, G.E.M.D.C. Bandara.............................698

The GA_NN_FL Associated Model for Authentication Fingerprints
Le Hoai Bac, Le Hoang Thai........................................................................708

Fuzzy Modeling of Zero Moment Point Trajectory for a Biped Walking Robot Dongwon Kim, Nak-Hyun Kim, Sam-Jun Seo, Gwi-Tae Park..............................716

Adaptive Resource Scheduling for Workflows Considering Competence and Preference
Keon Myung Lee..........................................................................................723

Analysis of Chaotic Mapping in Recurrent Fuzzy Rule Bases
Alexander Sokolov, Michael Wagenknecht................................................731

Highly Reliable Applications of Fuzzy Engineering

Damping Enhancement in Power Systems Using a Robust Fuzzy Sliding Mode Based PSS Controller
Farzan Rashidi, Mehran Rashidi.................................................................738
| Title                                                                 | Page |
|----------------------------------------------------------------------|------|
| Design a Robust and Adaptive Reinforcement Learning Based SVC Controller for Damping Enhancement in Power Systems | 745  |
| Farzan Rashidi, Mehran Rashidi                                       |      |
| A Rule-Based Approach for Fuzzy Overhaul Scheduling                  | 753  |
| Hongqi Pan, Chung-Hsing Yeh                                          |      |
| Fuzzy Kolmogorov’s Network                                           | 764  |
| Vitaliy Kolodyazhniy, Yevgeni Bodyanskiy                             |      |
| Fuzzy Selection Mechanism for Multimodel Prediction                  | 772  |
| Y. Bodyanskiy, S. Popov                                              |      |
| Efficient Approximate Reasoning with Positive and Negative Information | 779  |
| Chris Cornelis, Martine De Cock, Etienne Kerre                      |      |
| Chance Discovery                                                     |      |
| Chance Discovery as Novel Empathy with TV Programs                   | 786  |
| Masashi Taguchi, Yukio Ohsawa                                        |      |
| Enhancing Chance Discovery: Dimensions, Strategies and Tools         | 793  |
| Daniel Howard, Mark A. Eduards                                       |      |
| Consumer Behavior Analysis by Graph Mining Technique                 | 800  |
| Katsutoshi Yada, Hiroshi Motoda, Takashi Washio, Asuka Miyawaki      |      |
| A Chance Discovery Process to Understanding Spiral Behaviors of Consumers | 807  |
| Noriyuki Kusiro, Yukio Ohsawa                                        |      |
| Nursing Risk Prediction as Chance Discovery                          | 815  |
| Akinori Abe, Kiyoshi Kogure, Norihiro Hagita                        |      |
| Exploring Collaboration Topics from Documented Foresights of Experts  | 823  |
| Yumiko Nara, Yukio Ohsawa                                            |      |
| Condensation and Picture Annotations of Scenario Map for Consensus in Scenario Mining | 831  |
| Kenichi Horie, Takashi Yamaguchi, Tsuneki Sakakibara, Yukio Ohsawa   |      |
| Emergence of Product Value from On-line Communications               | 839  |
| Koichi Takahashi, Yukio Ohsawa, Naohiro Matsumura                    |      |
| Emerging Scenarios by Using DDM: A Case Study for Japanese Comic Marketing | 847  |
| Hiroshi Tamura, Yuichi Washida, Yukio Ohsawa                         |      |
| Intelligent Cooperative Work                                          |      |
| A Mobile Clickstream Time Zone Analysis: Implications for Real-Time Mobile Collaboration | 855  |
| Toshihiko Yamakami                                                   |      |
| Title                                                                 | Authors                                                                 | Page |
|----------------------------------------------------------------------|-------------------------------------------------------------------------|------|
| Interpretation of Emotionally Expressive Characters in an Intercultural Communication | Tomodo Koda                                                              | 862  |
| Development and Evaluation of an Intercultural Synchronous Collaboration System | Takashi Yoshino, Tomohiro Shigenobu, Shinji Maruno, Hiroshi Ozaki, Sumika Ohno, Jun Munemori | 869  |
| A Proposal of Knowledge Creative Groupware for Seamless Knowledge     | Takaya Yuizono, Jun Munemori, Akifumi Kayano, Takashi Yoshino, Tomohiro Shigenobu | 876  |
| comDesk: A Cooperative Assistance Tool Based on P2P Techniques       | Motoki Miura, Buntaoru Shizuki, Jiro Tanaka                             | 883  |
| Development of an Emotional Chat System Using Sense of Touch and Face Mark | Hajime Yoshida, Takashi Yoshino, Jun Munemori                           | 891  |
| Dual Communication System Using Wired and Wireless Correspondence in a Small Space | Kunihiro Yamada, Yoshihiko Hirata, Yukihisa Naoe, Takashi Furumura, Yoshio Inoue, Toru Shimizu, Koji Yoshida, Masanori Kojima, Tadanori Mizuno | 898  |
| The Beijing Explorer: Two-way Location Aware Guidance System         | Jun Munemori, Daisuke Kamisaka, Takashi Yoshino, Masaya Chiba           | 905  |
| Development of a System for Learning Ecology Using 3D Graphics and XML | Satoru Fujii, Jun Iwata, Yuka Miura, Kouji Yoshida, Sanshiro Sakai, Tadanori Mizuno | 912  |
| Practice of Linux Lesson in Blended Learning                         | Kazuhiro Nakada, Tomonori Akutsu, Chris Walton, Satoru Fujii, Hiroshi Ichimura, Kunihiro Yamada, Kouji Yoshida | 920  |
| Requisites for Talented People in Industry and the Method of Education | Teruhisa Ichikawa                                                      | 928  |
| Logic Based Intelligent Information Systems                             |                                                                        |      |
| Para-Fuzzy Logic Controller                                          | Jair Minoro Abe                                                        | 935  |
| Paraconsistent Artificial Neural Networks: An Introduction           | Jair Minoro Abe                                                        | 942  |
| The Study of the Effectiveness Using the Expanded Neural Network in System Identification | Shigenobu Yamawaki, Lakhmi Jain                                        | 949  |
A Paraconsistent Logic Program Based Control for a Discrete Event Cat and Mouse
Kazumi Nakamatsu, Ryuji Ishikawa, Atsuyuki Suzuki..............................................954

EVALPSN Based Railway Interlocking Simulator
Kazumi Nakamatsu, Yosuke Kiuchi, Atsuyuki Suzuki..............................................961

Learning by Back-Propagating Output Correlation in Winner-takes-all and Auto-associative Networks
Md. Shahjahan, K. Murase.....................................................................................968

Similarity Measures for Content-Based Multimedia Retrieval

Content-Based Video Retrieval Using Moving Objects’ Trajectories
Choon-Bo Shim, Jae-Woo Chang........................................................................975

Content-Based Image Retrieval Using Multiple Representations
Karin Kailing, Hans-Peter Kriegel, Stefan Schöner.............................................982

Similarity of Medical Images Computed from Global Feature Vectors for Content-Based Retrieval
Thomas M. Lehmann, Mark O. Güld, Daniel Keysers, Thomas Deselaers,
Henning Schubert, Berthold Wein, Klaus Spitzer...........................................989

Similarity: Measurement, Ordering and Betweenness
Walter ten Brinke, David McG. Squire, John Bigelow........................................996

Engineering of Intelligent Systems-Components and Activities

Qualitative Model for Quality Control in Production
Marjan Družovec, Tatjana Welzer........................................................................1003

A Functional Language for Mobile Agents with Dynamic Extension
Yasushi Kambayashi, Munehiro Takimoto.......................................................1010

Verifying Clinical Criteria for Parkinsonian Disorders with CART Decision Trees
Petra Povalej, Gregor Štiglic, Peter Kokol, Bruno Stiglic, Irene Litvan,
Dušan Flisar........................................................................................................1018

Improving Classification Accuracy Using Cellular Automata
Petra Povalej, Mitja Lenič, Gregor Štiglic, Tatjana Welzer, Peter Kokol........1025

Using Web Services and Semantic Web for Producing Intelligent Context-Aware Services
Kimmo Salmenjoki, Tatjana Welzer........................................................................1032

Internationalization Content in Intelligent Systems – How to Teach it?
Tatjana Welzer, David Riaño, Boštjan Brumen, Marjan Družovec......................1039
Intelligent System Design

Recognizing Frontal Faces Using Neural Networks  
Stephen Karungaru, Minoru Fukumi, Norio Akamatsu ...........................................1045

Identification of the Multi-layered Neural Networks by Revised GMDH-Type Neural Network Algorithm with PSS Criterion  
Tadashi Kondo, Abhijit S. Pandya ..............................................................................1051

Detection of Transition of Various Time Series Model Using BP Neural Networks  
Takahiro Emoto, Masatake Akutagawa, Hirofumi Nagashino, Yohsuke Kinouchi ..................................................................................................................1060

A Pattern Generator for Multiple Periodic Signals Using Recurrent Neural Networks  
Fumihiko Takahashi, Masatake Akutagawa, Hirofumi Nagashino, Yohsuke Kinouchi ..................................................................................................................1068

Identification of Number of Brain Signal Sources Using BP Neural Networks  
Hirofumi Nagashino, Masafumi Hoshikawa, Qinyu Zhang, Masatake Akutagawa, Yohsuke Kinouchi ........................................................................................1074

Knowledge–Based Intelligent Systems for Health Care

Development of Coronary Heart Disease Database  
Machi Suka, Takumi Ichimura, Katsumi Yoshida ..........................................................1081

Extraction of Rules from Coronary Heart Disease Database Using Automatically Defined Groups  
Akira Hara, Takumi Ichimura, Tetsuyuki Takahama, Yoshinori Isomichi .................1089

Immune Multi Agent Neural Network and Its Application to the Coronary Heart Disease Database  
Shinichi Oeda, Takumi Ichimura, Katsumi Yoshida ....................................................1097

FESMI: A Fuzzy Expert System for Diagnosis and Treatment of Male Impotence  
Constantinos Koutsojannis, Ioannis Hatzilygeroudis ..................................................1106

Disease Diagnosis Support System Using Rules, Neural Network and Fuzzy Logic  
Le Hoai Bac, Nguyen Thanh Nhi ................................................................................1114

Partial Merging of Semi-structured Knowledgebases  
Ladislau Bölöni, Damla Turgut ....................................................................................1121

Emotion Oriented Intelligent System for Elderly People  
Kazuya Mera, Yoshiaki Kurosawa, Takumi Ichimura ................................................1128

Multi-modal Data Fusion: A Description  
Sarah Coppock, Lawrence J. Mazlack ........................................................................1136
Multiagent Systems: Ontologies and Conflicts Resolution

Null Values and Chase in Distributed Information Systems
Agnieszka Dardzinska Glebocka.................................................................1143

Soft Implementations of Epistemic Satisfaction Relations in
Communicative Cognitive Agents
Radosław Piotr Katarzyniak.........................................................................1150

Multi-agent Web Recommendation Method Based on Indirect
Association Rules
Przemysław Kazienko................................................................................1157

Migration Mechanisms for Multi-class Objects in Multiagent Systems
Dariusz Król......................................................................................................1165

A Distributed Model for Institutions in Open Multi-agent Systems
Marcos De Oliveira, Martin Purvis, Stephen Cranefield,
Mariusz Nowostawski......................................................................................1172

Deriving Consensus for Conflict Situations with Respect to Its Susceptibility
Ngoc Thanh Nguyen, Michal Malowiecki.................................................................1179

A Collaborative Multi-agent Based Workflow System
Bastin Tony, Roy Savarimuthu, Maryam Purvis..................................................1187

A Subjective Logic-Based Framework for Aligning Multiple Ontologies
Krzysztof Juszczyszyn..........................................................................................1194

Operations Research for Intelligent Systems

When to Stop Range Process – An Expanded State Space Approach
Kazuyoshi Tsurusaki, Seiichi Iwamoto...............................................................1201

A Nondeterministic Dynamic Programming Model
Toshiharu Fujita, Takayuki Ueno, Seiichi Iwamoto...............................................1208

Toward The Development of an Auto-poietic Multi-agent Simulator
Katsumi Hirayama..............................................................................................1215

A Mean Estimation of Fuzzy Numbers by Evaluation Measures
Yuji Yoshida........................................................................................................1222

An Objective Function Based on Fuzzy Preferences in Dynamic Decision Making
Yuji Yoshida, Masami Yasuda, Jun-ichi Nakagami, Masami Kurano,
Satoru Kumamoto..............................................................................................1230

Intelligent Data Analysis and Application

An Efficient Clustering Algorithm for Patterns Placement in Walkthrough System
Shao-Shin Hung, Ting-Chia Kuo, Damon Shing-Min Liu........................................1237
| Title                                                                                           | Authors                                      | Page |
|------------------------------------------------------------------------------------------------|----------------------------------------------|------|
| Distance Preserving Mapping from Categories to Numbers for Indexing                            | Huang-Cheng Kuo, Yi-Sen Lin, Jen-Peng Huang  | 1245 |
| An Evolutionary Clustering Method for Part Family Formation with Multiple Process Plans         | Sheng-Chai Chi, In-Jou Lin, Min-Chuan Yan    | 1252 |
| Design the Hardware of Genetic Algorithm for TSP and MSA                                        | Wen-Lung Shu, Chen-Cheng Wu, Wei-Cheng Lai   | 1260 |
| Robust Bayesian Learning with Domain Heuristics for Missing Data                                | Chian-Huei Wun, Chih-Hung Wu                 | 1268 |
| OLAM Cube Selection in On-Line Multidimensional Association Rules Mining System                | Wen-Yang Lin, Ming-Cheng Tseng, Min-Feng Wang| 1276 |
| Mining Fuzzy Association Rules with Multiple Minimum Supports Using Maximum Constraints       | Yeong-Chyi Lee, Tzung-Pei Hong, Wen-Yang Lin | 1283 |
| Author Index                                                                                    |                                              | 1291 |
Implementing EGAP-Based Many-Valued Argument Model for Uncertain Knowledge

Taro Fukumoto, Takehisa Takahashi, and Hajime Sawamura

Department of Information Engineering and Graduate School of Science and Technology, Niigata University, 8050, Ninocho, Ikarashi, Niigata, 950-2181 Japan
{fukumoto, tekehisa, sawamura}@cs.ie.niigata-u.ac.jp

Abstract. We studied the many-valued argumentation frameworks. They allow agents to make arguments with self or other agents under uncertain knowledge which is to be represented in the expressive EGAP (Extended Generalized Annotated Logic Programming). In this paper, we describe the implementation of the EGAP-based multi-valued argument models. The versatility of our many-valued argument models is shown through convincing argument examples.

1 Introduction

Argumentation is one of the intelligent activities among humans, and is a useful computational apparatus for ill-formed problem domains for which modeling is difficult to build. Much work has been devoted to two-valued argument modes so far [1]. However, there has been no devotion to the multi-valued cases despite the fact that knowledge is usually uncertain. This paper is a continuation of our former paper on a theoretical argumentation framework (AF) [5]. It allows agents to make arguments under the multi-valuedness of knowledge using the EGAP (Extended Generalized Annotated Logic Programming). It is an expressive knowledge representation language that is syntactically GAP [3] extended by default negation, and is semantically guaranteed to have the well-founded semantics. The many-valuedness of GAP allows agents to assert their arguments under uncertainty such as vagueness, incompleteness and inconsistency. The default negation, on the other hand, allows them to represent incomplete knowledge or beliefs.

On the basis of our theoretical results, in this paper, we describe the implementation of the EGAP-based Multi-Valued Argument Model (Section 3) and its evaluation through meaningful argument examples (Section 4). Specifically, we show that the many-valued argument model can deal with various and versatile arguments for agents. In the next section, we outline our argumentation theory, so that the paper is to be self-contained.
2 Overview of EGAP and AF

The argumentation framework for the extended generalized annotated logic programs have been studied in [5]. One is a basic argumentation framework in which an agent makes arguments within his own knowledge base, and the other a multi-agent argumentation framework in which agents are to be involved in arguing issues with multiple knowledge bases.

2.1 Basic Argumentation

The basic argumentation (BA) framework deals with argumentation within a single EGAP.

Definition 1 (Extended Generalized Annotated Logic Programs). We assume a complete lattice of truth values \((T, \geq)\). An extended generalized annotated logic programs (EGAP) on \((T, \geq)\) is a set of ground instances of rules of the form:

\[ A_0: \mu_0 \leftarrow B_1: \mu_1 & \ldots & B_n: \mu_n & \textbf{not} \ (B_{n+1}: \mu_{n+1}) & \ldots & \textbf{not} \ (B_m: \mu_m). \]

Where each annotation \( \mu_i \) \((1 \leq i \leq m)\) is an element of \((T, \leq)\). \textbf{not} is the default negation symbol. \(A_0: \mu_0, \text{ and } B_1: \mu_1, \ldots, B: \mu_m\) are annotated atoms, and \textbf{not} \((B_i: \mu_i)\) \((n + 1 \leq i \leq m)\) are annotated default atoms. An EGAP with no annotated default atom coincides with a generalized annotated logic program (GAP [3]).

Definition 2 (Arguments). Let \(P\) be an EGAP. An argument in \(P\) is a finite sequence \(\text{Arg} = [r_1, \ldots, r_n]\) of minimal reductants [5] in \(P\) such that:

1. For every \(i \ (1 \leq i \leq n)\) and for every annotated atom \(A_j: \mu_j\) in the body of \(r_i\), there exists a minimal reductant \(r_k\) such that \(A_j: \mu_k \ (\mu_k \geq \mu_j, \ n \geq k > i)\) is head of \(r_k\).

2. There exists no proper subsequence of \([r_1, \ldots, r_n]\) which meets the first condition, and includes the minimal reductant \(r_1\).

The heads of minimal reductants in \(\text{Arg}\) are called conclusions of \(\text{Arg}\), and it called assumptions of \(\text{Arg}\) that is the annotated default atoms in the body of minimal reductants in \(\text{Arg}\). We write \(\text{concl}(\text{Arg})\) for the set of conclusions and \(\text{assm}(\text{Arg})\) for the set of assumptions of \(\text{Arg}\). We denote the set of all arguments in \(P\) by \(\text{Args}_P\).

Definition 3 (Undercut). \(\text{Arg}_1\) undercut \(\text{Arg}_2 \Leftrightarrow \) there exists \(A: \mu_1 \in \text{concl}(\text{Arg}_1)\) and \textbf{not} \((A: \mu_2) \in \text{assm}(\text{Arg}_2)\) such that \(\mu_1 \geq \mu_2\).

Definition 4 (BA-Dialogue Tree). Suppose \(P\) is EGAP. A BA-dialogue is a finite nonempty sequence of moves \(\text{move}_i = (\text{Player}_i, \text{Arg}_i) \ (i \geq 1)\) such that:

1. \(\text{Arg}_i \in \text{Args}_P \ (i \geq 1)\).

2. \(\text{Player}_i = \text{P} \ (\text{Proponent}) \Leftrightarrow i \text{ is odd;}
   \text{ and } \text{Player}_i = \text{O} \ (\text{Opponent}) \Leftrightarrow i \text{ is even.}\)
3. If $\text{Player}_i = \text{Player}_j = P$ ($i \neq j$) then $\text{Arg}_i \neq \text{Arg}_j$.

4. $(\text{Arg}_i, \text{Arg}_{i-1}) \in \text{undercut}$ ($i \geq 2$).

A BA-dialogue tree is a tree of moves such that every branch is a BA-dialogue, and for all moves $\text{move}_i = (P, \text{Arg}_i)$, the children of $\text{move}_i$ are all those moves $(O, \text{Arg}_j)$ such that $(\text{Arg}_j, \text{Arg}_i) \in \text{undercut}$. 

**Definition 5 (BA-Justified Argument).** There exists BA-dialogue tree of $\text{Arg}$ whose every termination is a move of proponent $\iff$ $\text{Arg}$ is BA-Justified argument.

We denote the set of BA-justified argument for an EGAP $P$ to be $J_P$. In [5], the equivalence of WFS and BA semantics of EGAP has been shown. That is, we have

1. $\text{BA}(P) \models A: \mu \iff$ there exists a justified argument $\text{Arg} \in J_P$ such that for some $\rho \geq \mu$, $A: \rho \in \text{concl}(\text{Arg})$;
2. $\text{BA}(P) \models \text{not}\ (A: \mu) \iff$, for every arguments $\text{Arg} \in \text{Args}_P$, if there exists $\rho \geq \mu$ such that $A: \rho \in \text{concl}(\text{Arg})$, $\text{Arg}$ is overruled.

**Theorem 1.** Let $P$ be an EGAP, and $\text{WFS}(P)$ be a well-founded model of $P$. Then $\text{WFS}(P) \models A: \mu \iff \text{BA}(P) \models A: \mu$ and $\text{WFS}(P) \models \text{not}\ (A: \mu) \iff \text{BA} \models \text{not}\ (A: \mu)$.

### 2.2 Multi-agent Argumentation

MAA is an argumentation framework in which agents argue each other about what each agent believes to be right (i.e., justified arguments for each agent in terms of BA). In MAA the rebuttal relation is introduced so as to reflect a skeptical view of agents to other knowledge bases (refer to [5] for the details).

**Definition 6 (Maximal Arguments).** Let $\text{Args}$ be a set of argument, and $\text{Arg}$ be an argument in $\text{Args}$. We define the set of conclusions whose annotations are maximal as follows. $\text{max}_{\text{concl}}(\text{Arg}) = \{A: \mu \in \text{concl}(\text{Arg}) \mid \text{for any } \rho, \text{if } A: \rho \in \text{concl}(\text{Arg}), \mu \geq \rho\}$. Then $\text{Arg}$ is called a maximal argument (m-argument) if for all $A: \mu \in \text{max}_{\text{concl}}(\text{Arg})$, there is no $\text{Arg}' \in \text{Args}$ such that for some $\rho > \mu$, $A: \rho \in \text{max}_{\text{concl}}(\text{Arg}')$.

**Definition 7 (Rebut and Defeat).** $\text{Arg}_1$ rebutts $\text{Arg}_2 \iff$ there exists $A: \mu_1 \in \text{max}_{\text{concl}}(\text{Arg}_1)$ and $A: \mu_2 \in \text{max}_{\text{concl}}(\text{Arg}_2)$ such that $\mu_1 \nRightarrow \mu_2$. $\text{Arg}_1$ defeats $\text{Arg}_2 \iff \text{Arg}_1$ undercuts $\text{Arg}_2$, or $\text{Arg}_1$ rebutts $\text{Arg}_2$ and $\text{Arg}_2$ does not undercut $\text{Arg}_1$.

Suppose $\text{MAS} = \{KB_1, \ldots, KB_n\}$ is a set of EGAPs. We denote the set of all arguments in $\text{MAS}$ by $\text{Args}_{\text{MAS}} = \bigcup_i \{\text{Arg} \mid \text{Arg} \text{ is a m-argument of } J_{KB_i}\}$

**Definition 8 (MAA-Dialogue Tree).** Suppose $\text{KBS} = \{KB_1, \ldots, KB_n\}$ is a set of EGAPs. A MAA-dialogue is a finite nonempty sequence of moves $\text{move}_i = (\text{Player}_i, \text{Arg}_i)(i \geq 1)$ such that:
1. $\text{Arg}_i \in \text{Arg}_{\text{MAS}} (i \geq 1)$.

2. $\text{Player}_i = \text{P} (\text{Proponent}) \iff i \text{ is odd};$
   and $\text{Player}_i = \text{O} (\text{Opponent}) \iff i \text{ is even}.$

3. If $\text{Player}_i = \text{Player}_j = \text{P} (i \neq j)$ then $\text{Arg}_i \neq \text{Arg}_j$.

4. If $\text{Player}_i = \text{P} (i > 1)$ then $(\text{Arg}_i, \text{Arg}_{i-1}) \in \text{undercut};$ and if $\text{Player}_i = \text{O}$ then $(\text{Arg}_i, \text{Arg}_{i-1}) \in \text{defeat}$.

The definition of a MAA-dialogue-tree is a tree of moves such that every branch is a MAA-dialogue, and for all moves $\text{move}_i = (\text{P}, \text{Arg}_i)$, the children of $\text{move}_i$ are all those moves $(\text{O}, \text{Arg}_j)$ such that $(\text{Arg}_j, \text{Arg}_i) \in \text{defeat}$.

**Definition 9 (MAA-Justified Argument).** There exists a MAA-dialogue tree for $\text{Arg}$ in which every leaf is a move of proponent $\Leftrightarrow \text{Arg}$ is a MAA-Justified argument.

### 3 Implementation

Each arguing agent has own knowledge base and reasoning capability for making arguments and counterarguments. The EGAP allows using arbitrary complete lattice of truth values, therefore, A user has to input the order of truth values with the knowledge base. The reasoning capability of each agent is realized following the BA argument framework described in the previous section. We introduced a special mediator agent, who directs the agent communication, keeping track of the argument flow, as illustrated in Figure [1]. The task and role of the mediator are similar to those of the judge agent in the contract net protocol (CNP) [4]. The generic communication framework of CNP is suitable to argument-based agent systems as well, and in fact makes the implementation easy and reliable. CNP can also be used to invite agents who attend to arguments. The communication part of the implementation was realized in Java, and the reasoning part in Prolog. Since the equivalence of WFS and BA semantics of EGAP, according to Theorem [1] the reasoning part of BA is the EGAP interpreter.

MAA is an argumentation framework in which agents with argue each other about what they believes to be right. An issue to be argued is first passed to the mediator agent by a user. Then, he asks an appropriate agent to ask to construct an argument for it according to Definition [2] that is, by means of SLD-like proof.

![Fig. 1. Communication flow of EGAP-based Many-Valued Argument Model](image-url)
procedure with reductants if necessary. BA is an argument framework for a single agent with a knowledge base. Put it differently, it produces a set of justified arguments (agents’ beliefs) by self-argumentation, or argumentation by monologue. BA is to be used in MAA argumentation framework for multi-agents. It should be noted that in MAA, arguments that agents cast each other are confined to m-arguments of Definition 6. According to the multi-agents argumentation dialogue tree (Definition 8), we introduce the following argumentation protocol to implement the argumentation scenario based on MAA, which coincides with the dialectical proof theory [5]. Figure 1 depicts an overall image of argumentation by multi-agents.

**Step 1.** A user passes the mediator an issue in be argued.

**Step 2.** The mediator picks up an appropriate agent who seems to be able to make an argument on it, and asks him to argue on it. In the meantime, the mediator receives the argument from him.

**Step 3.** The mediator broadcasts the argument to other agents, asking them to make undercutting or defeating arguments for it.

**Step 4.** Each agent tries to rebut or undercut the argument, and replies with those counterarguments if any (where counterargument means undercutting or defeating arguments).

**Step 5.** If there are counterarguments put forwarded to the mediator, then for each of them, he calls **Step 3** to seek further counterarguments for it. Otherwise, go to the next **Step 6**.

**Step 6.** Here there are two cases to be considered. If the mediator has no more counterarguments left to be considered, then go to **Step 7**. Otherwise, return to **Step 3**.

**Step 7.** If every terminal of a MAA-dialogue tree with first argument about issue is a proponent, then the argument is justified and the dialogue stops here. Otherwise, it is not justified and terminate.

### 4 Argument Examples and Evaluation

In order to show the variety and versatility of the expressiveness of EGAP in the BA and MAA argumentation, we illustrate two argument examples.

**Example 1. (BA Argument)** Suppose a complete lattice $T = [0, 1]^2$ where $(\mu_1, \rho_1) \leq (\mu_2, \rho_2) \iff \mu_1 \leq \mu_2$ and $\rho_1 \leq \rho_2$, and in $(\mu, \rho) \in T$, Then, each element of the ordered pair $(\mu, \rho)$ represents the independent ratios of an affirmative and a negative information in agents’ belief respectively. For instance, in terms of an annotated atom, $\text{agree}($Ptolemaic system$):(0.8, 0.2)$, an agent can express such a stand that he rather strongly believes the Ptolemaic system is true although he still has a bit of doubt on it. In other words, agents can assert not only absolute but also relative recognition of propositions or things with such a lattice of truth-values. (Note that two-valued argument models like for ELP only allow for a definite information such as affirmative or negative.)

Let us consider the following $KB_\alpha$ collecting agent $\alpha$’s beliefs on the Ptolemaic system.
\( KB_\alpha = \{ \)
\begin{align*}
&\text{agree}(\text{Ptolemaic System}):(0.8, 0.2) \leftarrow \\
&\text{move}(\text{Sun}):(0.9, 0.2) \land \textbf{not} \ (\text{move}(\text{Earth})): (0.9, 0.3), \\
&\text{agree}(\text{Ptolemaic System}):(0.5, 0.5) \leftarrow \textbf{not} \ (\text{right}(\text{book})): (0.9, 0.3), \\
&\text{agree}(\text{Ptolemaic System}):(0.1, 0.9) \leftarrow \text{move}(\text{Earth}):(0.2, 0.9), \\
&\text{move}(\text{Earth}):(0.9, 0.3) \leftarrow \textbf{not} \ (\text{right}(\text{book})): (0.9, 0.3), \\
&\text{move}(\text{Sun}):(0.9, 0.3) \leftarrow, \quad \text{right}(\text{book}):(0.9, 0.3) \leftarrow \}
\end{align*}

As can be seen from this belief base, EGAP allows agents to have apparently contradictory beliefs from different perspectives without having fears of committing to inconsistency. From this, we have the following well-formed arguments:

\[
\text{Arg}_1 = \{ \text{agree}(\text{Ptolemaic System}):(0.8, 0.2) \leftarrow \\
\text{move}(\text{Sun}):(0.9, 0.2) \land \textbf{not} \ (\text{move}(\text{Earth})): (0.9, 0.3), \\
\text{move}(\text{Sun}):(0.9, 0.3) \leftarrow \}
\]
\[
\text{Arg}_2 = \{ \text{move}(\text{Earth}):(0.9, 0.3) \leftarrow \textbf{not} \ (\text{right}(\text{book})): (0.9, 0.3) \}
\]
\[
\text{Arg}_3 = \{ \text{right}(\text{book}):(0.9, 0.3) \leftarrow \}
\]

Then, it can be seen that by Definition 4 for BA, \( \text{Arg}_1 \) is justified since although \( \text{Arg}_2 \) undercuts \( \text{Arg}_1 \), \( \text{Arg}_2 \) is undercut by \( \text{Arg}_3 \).

**Example 2. (MAA Argument).** Let us consider an example with a somewhat odd use of annotations. Let the complete lattice be a power set of the set \( T \) which consists of symposium venues A, B, and C as elements, with the set-inclusion order. Then, for instance, an annotated atom, \( \text{symposium}:(A, B, C) \) means that the symposium venue candidates are A, B, and C. Four agents \( \alpha, \beta, \gamma, \delta \) are assumed to attend a discussion to decide the symposium venue candidates, in \( MAS = \{ KB_\alpha, KB_\beta, KB_\gamma, KB_\delta \} \), where

\[
\text{KB}_\alpha = \{ \text{symposium}:(A, B, C) \leftarrow \text{scenic}:(A, B) \land \text{safe}:(A, C), \\
\text{scenic}:(A, B) \leftarrow, \text{easy access}:(C) \leftarrow, \text{safe}:(A, C) \leftarrow \},
\]
\[
\text{KB}_\beta = \{ \text{symposium}:(\neg A) \leftarrow \text{last venue}:(A) \land \textbf{not} \ (\text{tasty food}):(A), \\
\text{last venue}:(A) \leftarrow \}
\]
\[
\text{KB}_\gamma = \{ \text{symposium}:(\neg C) \leftarrow \textbf{not} \ (\text{easy access}):(C) \}
\]
\[
\text{KB}_\delta = \{ \text{symposium}:(A, B, C) \leftarrow \text{tasty food}:(A, B, C). \\
\text{tasty food}:(A, B, C) \leftarrow \}
\]

In the knowledge bases, \( \neg \) stands for the explicit negation, such that \( \neg : T \rightarrow T \), and \( \neg A : \mu = A : \neg \mu \). Then for instance, \( \text{symposium}:(\neg A) \) means “A is not a venue candidate for the symposium”, and defined to be \( \text{symposium}:(\neg A) = \text{symposium}:(T - A) = \text{symposium}:(B, C) \).

The flow of the argument in Figure 2 shows that \( agent_\alpha \) first puts forward such an argument that the symposium venue candidates are A, B, and C, then \( agent_\beta \) and \( agent_\gamma \) defeat this argument in the two ways, however \( agent_\delta \) and \( agent_\alpha \) finally undercut those two. After all, the \( agent_\alpha \)’s argument results in being justified. This use of annotations brings us a compact and easy to understand
Implementing EGAP-Based Many-Valued Argument Model

Fig. 2. Argument tree in Example 2

the description of knowledge and arguments, compared with other orthodox ones such as $T = \{\bot, t, f, \top\}$ and $T = [0, 1]$.

5 Concluding Remark

We have described the implementation of the EGAP-based multi-valued argument models: BA and MAA, and evaluated them through two argument examples with unusual annotations. Lessons learned may be summarized into two points:

(1) EGAP is not only a highly expressive knowledge representation language but also a computationally feasible one, and;
(2) the versatility of the many-valued argumentation is advantageous to develop it to many application domains.

The implementation of EGAP through argumentation brings us an interpreter of EGAP as a byproduct, since it is known that BA semantics is equivalent to the well-founded semantics of EGAP [5]. In other words, we may say we have given a sound and complete way to compute EGAP through argumentation, although this idea originates from Dung’s one [2].

References

1. C. I. Chesnevar, A. G. Maguitman, R, P, Loui: Logic models of argument, ACM Computing Surveys, Vol. 32, pp. 337-383, 2000.
2. P. M. Dung: An argumentation theoretic foundation of logic programming, J. of Logic Programming, Vol. 22, No. 2 pp. 151-177, 1995.
3. M. Kifer and V .S. Subrahmanian: Theory of generalized annotated logic programming and its applications, J. of Logic Programming, Vol. 12, pp. 335-397, 1992.
4. H. Sawamura, and S. Maeda: An Argumentation-Based Model of Multi-Agent Systems, Information Modeling and Knowledge Base XII, pp. 137-150, 2001.
5. T. Takahashi, Y. Umeda, and H. Sawamura: Formal Argumentation Frameworks for the Extended Generalized Annotated Logic Programs, Proc. of the 7th Int. Conference on Knowledge-Based Intelligent Information & Engineering System, LNAI, Vol. 2773, Springer-Verlag, pp. 28-38, 2003.
Ontology Revision Using the Concept of Belief Revision

Seung Hwan Kang and Sim Kim Lau

Information Systems Discipline,
School of Economics and Information Systems,
University of Wollongong,
Northfields Avenue, Wollongong, NSW, 2522, Australia
{sk33, simlau}@uow.edu.au

Abstract. One of the problems identified in the development of ontology is the difficulty in maintaining ontology that often faces on issues of changes in knowledge or perception about things within the community of practice. When new information is added, consistency needs to be maintained to ensure it does not cause inconsistency within existing concepts in ontology. This paper discusses the feasibility of using the concept of belief revision as a basis for ontology revision. It is an effort to the use of expansion, revision and contraction operators of belief revision to revise ontology.

Keywords: Ontology Revision, Belief Revision, Ontology, the Semantic Web.

1 Introduction

Ontology provides a very useful way to structure and define the meaning of metadata of the Web documents. The emergence of the Semantic Web also provides a way to bring structure to the content of Web pages and create an environment that allow agents to be used and deployed to perform tasks for the users [2]. One of the problems identified in the development of ontology is the difficulty in maintaining ontology when there is a change in knowledge or perhaps a change in the perception about things within the community of practice. When the system accepts new information or knowledge, or when people change perception about certain things, this new information may contradict with what was initially agreed or defined in the ontology. When this happens, the ontology needs to be revised to reflect the changes. Ontology revision, defined as “a change in the components of ontology”, is closely related to the concept of belief revision [11]. This paper discusses the feasibility of using the concept of belief revision as a mechanism for ontology revision.

The paper is organized as follows. Section 2 discusses the needs for ontology revision. Section 3 presents the concept of belief revision. An illustration to use the belief revision concept to revise ontology is given in Section 4, and conclusion follows in Section 5.
2 Motivation

There are various definitions of ontology in the literature, commonly used definitions include: “ontology is a science or study of being” [13]; “ontology is a particular theory of the nature of being or existence” [17]; “ontology is a formal explicit specification of a shared conceptualization” [9], [10]. Ontology is derived from cognitive semantics which relate to expressions of conceptual structures [8], [20]. In particular, ontology is essential in order to achieve the vision of Semantic Web Ontology [12].

There are different ontology application scenarios [14]. Each of the ontology application scenarios has the underlying assumption that ontology remains unchanged in a static way. Once the ontology has been written and integrated to applications, little or no consideration is given about changing the ontology. However, a specification or conceptualization of domain knowledge tends to change when we learn something new or when the system accepts new information. When an individual or a community learns something new, a change of knowledge might occur through some form of belief changes. This is closely related to the concept of representation adjustment and presentation adjustment.

Ontology can evolve over time as a result of extension from previous ontologies or revision over time. When this occurs, problems such as ontology inter-operability problem and handling of multiple ontologies need to be addressed. A possible approach to tackle the ontology maintenance issues is to use ontology versioning or ontology library system [3], [15]. The concept of ontology versioning is used to reduce the inter-operability problem caused by the evolution of ontology [15]. It allows comparability issues to be taken into consideration when new knowledge is added to the system over time. Ontology Web Language (OWL) resolves the ontology versioning problem using a standard tag to provide consistency in terms of version control [19]. Ontology library is used to handle ontology maintenance. Dynamic concept sets has also been used in ontology revision [5].

3 Belief Revision

Belief revision deals with inconsistency when the new knowledge base or database needs to be revised. This process is important to ensure that new information do not cause inconsistent beliefs and contradict with the existing belief [7], [18]. There are two approaches to describing belief revision: the foundation theory and the coherence theory. The foundation theory focuses on keeping track of justifications for one’s belief [7], [8]. The coherence theory highlights the logical structure of things in a “world” which are semantics in a form of logically consistent structure.

We will briefly introduce three belief revision operators as proposed by the AGM (Alchourrón, Gärdenfors and Makinson) model [1]. Let a belief set $K$ be represented by a set of sentences in the logical language $L$. The language $L$ contains the standard logical connectives: negation ($\neg$), conjunction ($\land$), disjunction ($\lor$) implication ($\rightarrow$), and two truth values of truth ($T$) and falsity ($\bot$). In a consistent belief set $K$, there are three possible epistemic states towards a sentence $\alpha$: accepted, rejected and unknown.
\( \alpha \) is accepted (\( \alpha \in K \))
\( \alpha \) is rejected (\( \neg \alpha \in K \))
\( \alpha \) is unknown (\( \alpha \notin K \) and \( \neg \alpha \notin K \))

It is worth pointing out that “\( \alpha \) is unknown” means that both \( \alpha \) and \( \neg \alpha \) are accepted is inconsistent. Thus modeling that epistemic state is not allowed.

Consider the following set of sentences in the belief set \( K \).

\( \alpha \): All cameras are electronics.
\( \beta \): The camera displayed in the shop is DSCV1.
\( \gamma \): The camera displayed in the shop is a Sony product.
\( \delta \): Sony is a part of electronics’ industry.

Thus using \( \alpha - \delta \) the following fact is derived, \( \varepsilon \): The camera displayed in the shop is electronics.

Assume that the shop owner discovers that the camera (DSCV1) displayed in the shop is a digital camera. The owner believes that the digital camera should belong to computer peripherals. Therefore, \( \varepsilon \) is no longer consistent in his belief set, and there is a need to add negation of \( \varepsilon (\neg \varepsilon) \) to the belief set. This kind of change is called an expansion of belief set. The belief set that results from expanding \( K \) by a sentence \( \phi \) is denoted by \( K^+ \phi \). In the above example the new sentences added to the belief set as a result of expansion are: \( \varepsilon \) and \( \phi \) (where \( \phi \) is \( \neg \varepsilon \)).

Now consider the following scenario to describe the revision operator. Assume that the shop owner does not want to lose valuable information which describes the belief “All cameras are electronics”. In this case, the shop owner needs to revise the belief based on the given evidence that contradicts with what had previously agreed to accept. We denote the result of revising \( K \) by a sentence \( \phi \) as \( K^* \phi \). As a result of revision, the belief set is now made up of: \( \alpha, \beta, \gamma, \delta, \varepsilon, \phi \) and \( \alpha' \) (where \( \alpha' \): All cameras except the one displayed in the shop are electronics).

The third example illustrates the case when some beliefs are found to be invalid. In this case the belief is to be given up (contract) to allow new beliefs to be accepted. The belief set that results from contracting \( K \) by a sentence \( \phi \) is denoted by \( K^- \phi \). Contraction occurs when as a result from expanding \( K \) by a sentence \( \phi \), results in an inconsistency such as \( \alpha \). When this happens, \( \alpha \) needs to be removed because it is no longer consistent in the belief set \( K \). Thus a contraction occurs when some sentences in the belief set is retracted without adding any new beliefs. In order that the resulting belief set be closed under logical consequences some other sentences from the belief set may need to be given up [6]. In the belief system, it involves a step known as dependency-directed backtracking to make assumptions that admits the possible contractions [4].

One of the concerns of the underlying idea of revision and contraction methods is removing potentially useful information in the process of removing conflicting beliefs [6], [16], [18]. Using the same example as demonstrated above, the sentences in the belief set as a result of contraction are: \( \beta, \gamma, \delta, \varepsilon, \phi \) and \( \alpha' \).
4 Illustrations

We will use a scenario, an online buying of a digital camera, to illustrate the application of belief revision concept on ontology revision. Assume that a buyer agent is triggered to buy a camera in an e-commerce environment. Based on the specifications that are provided, the buyer agent roams to a seller website to find the “right” product to buy. The concepts related to the camera such as “all cameras are electronics”, “Sony is a manufacturer of electronic products”, “Sony manufactures cameras”, “Sony is a brand name”, “DSCV1 is a camera produced by Sony” are stored in the ontology of the buyer agent. We use concept hierarchy to illustrate the conceptual relationship, in which the relationships of different concepts are shown using parent-child relationship. We attempt to use the belief revision concept to illustrate the updates of ontology of the buyer agent as a result of encountering new information from the ontology of the seller agent.

Fig. 1 shows a part of the ontology of the buyer agent, which describes a brief concept of a camera that includes industry, manufacturer and brand. Fig. 2 shows a segment of the ontology of the seller agent, which includes the concept of a digital camera as a form of a computer and is described using the concept resolution. Fig. 3 shows partial codes of a representation of ontology for buyer agent in OWL format.

<?xml version="1.0"?>
<rdf:RDF
  xmlns="http://myhost/ontology#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xml:base="http://myhost/ontology">
  ...
  <owl:Class rdf:ID="Camera">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Electronics"/>
    </rdfs:subClassOf>
  </owl:Class>
  ...
</rdf:RDF>

Fig. 1. An example of ontology concept hierarchy of the buyer agent

Fig. 2. An example of ontology concept hierarchy of the seller agent

Fig. 3. Partial codes to show ontology of buyer agent
The first step is to represent the conceptualization from ontology to the belief set. Sentences of ontology for buyer agent are denoted as $\alpha$, $\beta$, $\gamma$, $\delta$, and that of seller agent are denoted as $A$, $B$, $C$, $D$. Fig. 4 shows the belief sets of the ontology for buyer and seller agents.

$\alpha$: A Camera is a subclass of Electronics.

$\beta$: Sony is a subclass of Electronics.

$\gamma$: DSCV1 is a subclass of Sony.

$\delta$: DSCV1 is a subclass of a Camera.

$A$: A Computer is a subclass of a Thing.

$B$: A Digital Camera is a subclass of a Computer

$C$: DSCV1 is a subclass of a Digital-Camera

$D$: A Resolution is a subclass of DSCV1

Fig. 4. An example of sentences from ontology belief set

We will use the following scenario to illustrate a process of expansion of new concepts. When the buyer agent contacts the ontology of the seller agent, it comes across the concept of “resolution”, which is not included in its own ontology. The buyer agent is not able to recognize the concept resolution. If the seller agent provides information and explanation of what the concept of resolution is (for example the concept of resolution is explained as a characteristic of the digital camera and is measured by number of pixels), then the buyer agent knows that it has learnt a new concept that can be used to describe digital camera. When this occurs the belief system of the buyer agent can check and deduce the new truth value of the new information in its ontology through the application of expansion operator in the belief revision model. In this case, the belief set $K$ is expanded by a new sentence $A (K^+)$.

The following describes the revision process. After an expansion of the concept of resolution is applied, the revised ontology may not be consistent. Following from the previous example, as a result of expansion, the concept DSCV1 is no longer consistent in the belief set. When this happens, revision operator is applied to ensure consistency is achieved. Fig. 5 shows postulates based on the AGM model for revision, which fully defined a rational revision function [7]. To satisfy the corresponding revision function, the basic sets of postulates ($K^*1$) – ($K^*6$) are required. Revising $K$ with sentences $A$ and $B$ are the same as the revising $K$ with respect to $A$ then expanding $K^+_A$ by $B$ (see $K^*7$). If $B$ does not contradict the beliefs in $K^+_A$, then $(K^+_A)^+_B$ is the same as $K^+_A \wedge B$ (see $K^*8$). Thus, when the seller agent provides a concept a Digital Camera is a subclass of a Computer ($B$), the buyer agent uses this new information and new perception about a camera to revise its own concept on camera. Consequently, the sentence $B'$: A Digital Camera is a subclass of a Computer is added to the belief set $K$ of buyer agent:

The next step of ontology revision is the application of contraction function [7]. To satisfy the corresponding contraction function, the basic sets of postulates ($K-1$) – ($K-6$) are required (see Fig. 6). Then contracting $K$ with sentences $A$ and $B$ cannot
remove any sentence unless the contraction has at least A and B (see K-7). (K-8) is another complementary postulate, if A does not belong to contraction of A and B, then the removal of A and B is also a subset of removal A.

(K*1) For any sentence A, and any belief set K, $K_A^*$ is a belief set (Closure)

(K*2) $A \in K_A^*$ (Success)

(K*3) $K_A^* \subseteq K_A^*$ (Expansion 1)

(K*4) If $\neg A \notin K$, then $K_A^* \subseteq K_A^*$ (Expansion 2)

(K*5) $K_A^* = K_\bot$ if and only if $\not\vdash \neg A$ (Consistency Preservation)

(K*6) If $\vdash A \leftrightarrow B$, then $K_A^* = K_B^*$ (Extensionality)

(K*7) $K_{A \land B}^* \subseteq (K_A^*)^*_{B}$ (Conjunction 1)

(K*8) If $\neg B \notin K_A^*$, then $(K_A^*)^*_{B} \subseteq K_{A \land B}^*$ (Conjunction 2, Rational Monotony)

Fig. 5. Postulates of revision function based on AGM model

(K-1) For any sentence A, and any belief set K, $K_A^-$ is a belief set (Closure)

(K-2) $K_A^- \in A$ (Inclusion)

(K-3) If $A \notin K$ then $K_A^- = K$ (Vacuity)

(K-4) If $\not\vdash A$, then $A \notin K_A^-$ (Success)

(K-5) If $A \in K$, then $K \subseteq (K_A^-)^*_{A}$ (Recovery)

(K-6) If $\vdash A \leftrightarrow B$, then $K_A^- = K_B^-$ (Extensionality)

(K-7) $K_A^- \cap K_B^- \subseteq K_{A \land B}^-$ (Conjunction 1)

(K-8) If $A \notin K_{A \land B}^-$, then $K_{A \land B}^- \subseteq K_A^-$ (Conjunction 2)

Fig. 6. Postulates of contraction function based on AGM model

In our example, when a new sentence C (DSCV1 is a subclass of a digital camera) is introduced to the belief set K, then it becomes inconsistent with the sentence $\gamma$ (DSCV1 is a subclass of SONY). To ensure consistency, we need to make a reasonable choice on which sentence is to be given up. Let assume C meets the postulates of (K-1) – (K-6), then $\delta$ (DSCV1 is a subclass of Camera) is a reasonable choice to give up by the contraction function.

The final step is to create new ontology that is based on revised belief set by the agent. In our example, ontology includes the following sentences: $\alpha$, $\beta$, $\gamma$, A, B', C, and D.

5 Conclusions

The notion of ontology revision means that there is a need to make adjustment of new concepts, rules and relations of ontology when an agent learns new information or perception changes in the community of practice. In this paper, the concept of belief revision is based on the AGM model. The expansion, revision and contraction operators of the AGM model provide a tool to achieve ontology revision. One of the issues
worth further investigation is the concern of maintaining the ontology versioning system and ontology library to enable management and handling of comparability issues in ontology as a result of ontology revision.

References

1. Alchourron, C., Gärdenfors, P. and Makinson, D. (1985) "On the Logic Theory Change: Partial Meet Contraction and Revision Functions", The Journal of Symbolic Logic 50: 510-530
2. Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. Scientific American. <http://www.sciam.com/2001/0501issue/0501berners-lee.html>. 1 Apr. 2002
3. Ding, Y. and Fensel, D. (2001) Ontology Library Systems: The key to successful Ontology Re-use, In Proceedings of International Semantic Web Working Symposium (SWWS) on Ontology and Ontology Maintenance. 30 Jul. - 1 Aug. Stanford University, California, USA. <http://www.semanticweb.org/SWWS/program/full/>
4. Doyle, J. (1979) A Glimpse of Truth Maintenance, In Proceedings of the Sixth International Joint Conference on Artificial Intelligence. 20-23 Aug. Tokyo.
5. Foo, N. (1995) Ontology Revision, In Proceedings of the 3rd International Conference on Conceptual Structures. 14-18 Aug. University of California, Santa Cruz. <http://www.cse.unsw.edu.au/~ksg/Abstracts/Conf/ontology_rev.html>
6. Gärdenfors, P. (1990) In Revue Internationale de Philosophie. Vol. 44 (Eds, Brennan, G. and Walsh, C.) Cambridge University Press, Cambridge. pp.24-46.
7. Gärdenfors, P. (1992) In Belief Revision. Cambridge University Press. pp.1-20.
8. Gärdenfors, P. (1995) In Mindscapes: Philosophy, Sciences, and the Mind (Eds, Carrier, M. and Machamer, K. P.) Pittsburgh University Press. pp.61-86.
9. Gomez-Perez, A. (1999) Applications of Ontologies and Problem-Solving Methods, In Proceedings of Workshop at Thirteenth Biennial European Conference on Artificial Intelligence. 1 - 4 Oct. Berlin, Germany.
10. Gruber, R. T. (1993) A Translation Approach to Portable Ontology Specifications
11. Heflin, D. J. and Hendler, J. (2000) Dynamic Ontologies on the Web, In Proceedings of the Seventeenth National Conference on Artificial Intelligence. 30 Jul. - 3 Aug. Austin, Texas. <http://www.aaai.org/Conferences/National/2000/aaaai00.html>
12. Hendler, J. (2001) Agents and the Semantic Web. Department of Computer Science. University of Maryland. <http://www.cs.umd.edu/users/hendler/AgentWeb.html>
13. Hornby, S. A. (1995) "Oxford English Dictionary". Oxford University Press. Oxford, New York.
14. Jasper, R. and Uschold, M. (1999) A Framework for Understanding and Classifying Ontology Applications, In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence Workshop on Ontology. 31 Jul. - 6 Aug. City Conference Center, Stockholm, Sweden. <http://www.sdv.su.se/ijcai-99/> 
15. Klein, M. and Fensel, D. (2001) Ontology versioning on the Semantic Web, In Proceedings of International Semantic Web Working Symposium. 30 Jul. - 1 Aug. Stanford University, California, USA. <http://www.semanticweb.org/SWWS/>
16. McAllester, A. D. (1990) Truth Maintenance, In Proceedings of AAAI-90.
17. Russell, S. and Norvig, P. (2003) "Artificial Intelligence: A Modern Approach". Prentice Hall.
18. Segal, R. (1994) Belief Revision. Department of Computer Science and Engineering, FR-35. University of Washington. <http://citeseer.nj.nec.com/segal94belief.html>.
19. W3C (2003) *OWL Web Ontology Language Use Cases and Requirements*. World Wide Web Consortium. <http://www.w3.org/TR/webont-req/>. 27 Sep. 2003

20. Zúñiga, L. G. (2001) Ontology: Its Transformation From Philosophy to Information Systems, In *Proceedings of the International Conference on Formal Ontology in Information Systems*. 17 - 19 Oct. Ogunquit, Maine, USA. <http://www.fois.org/fois-2001/>
A Robust Rule-Based Event Management Architecture for Call-Data Records

C. W. Ong and J. C. Tay

Center for Computational Intelligence,
Nanyang Technological University
asjcmtay@ntu.edu.sg

Abstract. Rules provide a flexible method of recognizing events and event patterns through the matching of CDR data fields. The first step in automatic CDR filtering is to identify the data fields that comprise the CDR format. In the particular case of the Nortel Meridian One PABX, five different call data types can be identified that are critical for call reporting. The architecture we have proposed will allow for line activity analysis while continuously publishing action choices in real-time. For performance evaluation of serial-line CDR data communications, an approximation to the CDR record loss rate at different simulated call traffic intensities was calculated. Here, the arrival process represents the arrival of newly generated CDR to the output buffer and the service process represents the process of transmitting the CDR over the serial connection. We calculate the CDR loss rate at different arrival intensities and observed that the CDR loss rate is negligible when the CDR arrival rate is less than 4 CDR per second.

Keywords. Rule-based system, CDR loss rate, event filtering and correlation.

1 Introduction and Motivation

Telecommunications monitoring is usually conducted by maintaining raw data logs produced by the Private Automatic Branch Exchange (or PABX). There is usually very little meaningful correlation to the user directory to allow the management to view user call details before monthly bills are invoiced from the service provider.

Call data records (or CDRs) are typically produced during or after calls are made and need to be filtered and identified before they can be used meaningfully. In this manner, we can classify CDRs as events that can be processed. One common method of call data filtering is essentially a set of ‘if-then’ control structures used to systematically decipher each CDR according to predefined vendor-specific CDR formatting. Such an implementation implies little room for variation in call data formatting, and imposes brittleness on the design of the filtration and correlation function. Another factor to motivate automatic filtration of CDR information is the incidence of telecommunications fraud, typified by overseas calls and prolonged phone usage during work hours. This is a costly issue which can be prevented through the proactive use of rules to ensure suspicious activity is recognized and that management is alerted.

Rules provide a flexible method of recognizing events and event patterns through the matching of CDR data fields. A basic architecture must provide a sufficient
subset of rules that can match most usage patterns and data formats while allowing unknown patterns to be learnt via user-intervention. Such a system would save the tedious and impossible task of categorizing each and every vendor-specific CDR format and having to hard-code for all present and future usage patterns. Although there is commercially available software on the market like Hansen Software’s CASH Call Accounting Software or TelSoft Solutions [1], these are closed-source and are not ideal platforms for customizing to the user’s needs.

In this paper, we propose an effective architecture for filtering and learning CDRs from correct, partially-correct, and unknown PABX data formats through the use of an embedded forward-chaining rule-based engine. In addition, the proposed architecture also provides web-based customizable reports for trend and historical analysis of phone-line usage. This research is based on our experience in implementing similar systems for Banks, Credit Collection Agencies and Manufacturing research centers [2].

2 Overview of CDR Filtering and Correlation

The CDR transaction model we assume in our study is one in which the CDR is produced only after a call has been placed. On one hand, this model makes the job of collecting call data for completed phone calls simpler but on the other, this implies that a complete picture of the state of current phone-line usage is unknown. However, it remains realistic as most phone systems issue CDRs only during or after a call is made.

The first step in automatic CDR filtering is to identify the data fields that comprise the CDR format and which can be used to identify the type of CDR being produced. In the particular case of the Nortel Meridian One PABX [3], five different call data types [4] can be identified that are critical for call reporting. They are; Normal Calls, Start/End Calls, ACD Call Connections, Authorization and Internal.

![Fig. 2.1. CDR Generation for Normal Calls](image)

A more complex scenario occurs when calls are transferred and CDRs have to be correlated together to form a single record in the system. This is shown in Fig 2.2. After the variables are identified, they can be classified and used within rules for filtering CDR and identification.

3 Simple Fraud Alert Rules Development

In traditional fraud monitoring practices, the administrator would only be aware of the case after it has occurred and by then it would be too late to take measures against the
Fig. 2.2. CDR Generation for Start/End Calls

perpetrator. Usually the reports will have to be viewed by the administrator and flagged for suspicious activity, all of which is time-consuming and prone to errors.

The model presented here is intended as a first step towards improving fraud detection efficiency and effectiveness. The expert system model once developed and introduced will allow for line activity analysis while continuously publishing action choices in real-time. Specifically, after every interval of CDR arrival, it can be checked against fraud rules and reports can immediately be sent to the administrator. The administrator can then take further action by flagging specific lines for further analysis and monitoring. The action will usually be recommended (based on similar past actions) by the system to minimize the administrative load.

3.1 Rule-Based Fraud Modeling

Ideally the process of modeling fraud activity involves collecting historical line activity data, and then applying statistical techniques (such as discriminant analysis) on the dataset to obtain a predictive model that is used to distinguish between fraud and non-fraud activity [5][6]. In our case however, it would be more difficult to distinguish fraud and non-fraud activity as CDRs are only issued after a call is made. Instead, a rule-based approach is used to correlate variables which form conjunctive patterns in fraudulent phone-line usage. A fraud variable datum (or FVD) is a data field in a CDR which is used as part of a rule condition to detect fraudulent phone-line usage.

Some examples of FVDs (in order of significance) are; Duration of Call, Frequency of Call, Time of Call, Day of Call and Destination of Call. The FVDs represent the signs which a human operator should take note of when detecting fraud. Duration of Call is often the prime suspect since an international call (or IDD) that is placed for more than a certain duration incurs great cost and is unlikely to be used for meetings, is cause for alarm. Frequency of Call can also indicate fraud activity since a high frequency detected within a certain time period could be indicative of redial activity in an attempt to gain unauthorized access to systems. By monitoring Time and Day of Call, unusual activities can also be caught during periods when calls are not expected to be placed for most extensions. The last category, Destination of Call, could be used to monitor lines in which IDD calls are not allowed. Some examples of Fraud Detection Rules are:
- If DurationOfCall > 3600 then email Administrator;
- If FrequencyOfCall > 10 AND TimeOfCall = 5 then email Administrator;
- If TimeOfCall > OFF_PEAK then monitor line;

Actions or Jobs can be specified to be executed when rules are triggered to facilitate reporting and alert creation. Actions include using HTML alert, E-Mail and short message service (or SMS).

4 An Event Management Architecture

By classifying incoming CDR as events to be managed, we design a system that is able to process incoming events using JESS rules. In this case, events are generated from the Nortel PABX. The CDRTool system consists of modules to intercept events, process them with the help of the Rule Framework and store processed events in a database. This database can then be processed offline by a Report Framework for statistical analysis and report generation.

The CDRTool System consists of various submodules to intercept events, process and store them into a database. The architecture shown in Fig 4.1 allows for component failure and recovery with little intervention from the administrator. A Supervisor Module ensures each component is running by using a Publish/Subscribe message service. Clients address messages to a topic. Topics retain messages only as long as it takes to distribute them to current subscribers. Publishers and subscribers have a timing dependency. A client that subscribes to a topic can consume only messages published after the client has created a subscription, and the subscriber must continue to be active in order for it to consume messages.

Each component that makes up the workflow from getting the raw CDR to a processed CDR for rule matching communicates using a point-to-point messaging system also via JMS (Java Messaging System). A point-to-point (or PTP) product or application is built around the concept of message queues, senders, and receivers. Each message is addressed to a specific queue, and receiving clients extract messages from the queue(s) established to hold their messages. Queues retain all messages sent to them until the messages are consumed or until the messages expire.

The application business model allows a component to send information to another and to continue to operate without receiving an immediate response. Using messaging for these tasks allow the various components to interact with one another efficiently, without tying up network or other resources. The PABX Client module is a Web-based communications program in charge of reading raw CDR from the PABX.

The CDR Processor Module will retrieve from the Attributes and Schema database, those attributes that make up a CDR and attempt to create a well-formed CDR from the raw record. If it is able to do so, it will then proceed to queue it for processing by the CDR Manager. If not, errors are reported and tagged for batch processing to identify those CDR sets which will be used by the administrator to take corrective action (for example adding it to the database so that future CDR can be identified).

The CDR Manager uses JESS [7] to identify the CDRs that are possibly alert cases based on rules set down by the administrator and also logs them into a central database for report generation. CDRs that arrive at the CDR Manager are assumed to
be well-formed (as verified by CDR Processor); however, the rule engine still performs a verification check against the schema database for any data range changes if they have been specified by the administrator. This two-check process ensures that the CDRs that enter into the database are valid. Fraud rules are also applied at this step and any fraud triggers are then queued for job execution. CDRs which trigger the fraud rules are tagged upon insertion into the CDR database and the appropriate alert action (using E-Mail or SMS) is taken.

Fig. 4.1. The CDRTool System Architecture

The Attribute Manager is a web-based application that allows user specification of CDR schema for the specific PABX that is in use. The various attributes and variables specified here will be used by the system to correlate CDRs and also identify unknown CDRs for system learning. The schema is stored in XML form for portability.

The rule builder is where users can enter new rules. The rule builder retrieves conditions from the Attributes and Schema database and performs data range verification. The information is grouped and displayed so that rules can be added quickly and all the system rules viewed at a glance. Job types can be specified when rules are matched (Job options are Email, SMS or HTML reports). The results of rule executions are sent via the Report Framework.

The rules which are entered via the Rule Builder Form are stored into a JESS rule-base. The rule engine is used primarily by the CDR Manager and also for dispatching of alerts to the Report Framework when fraud rules are triggered. Rules are created by the Rule Builder and stored in a SQL database called the Rule Database which is used by the rule engine. A wizard-guiding system for adding rules was developed to ease the administrative burden of manually going through error logs to identify new rules.
A reporting framework provides the system with a means to generate different user specified reports in HTML, E-Mail or Excel spreadsheet formats. The system also allows report templates to be customized using the web-based GUI. The major modules in the Report Framework are the Template Builder, the User Query Interface and the Report Presenter. Each module represents the work required from creating reports to gathering user queries and finally presenting the results in HTML or Excel. These results can then be viewed either in a browser or sent as E-mail as well as SMS alerts.

The Template Builder gathers attributes from the Attributes and Schema database and provides a web-based interface for building report templates. This allows customisation to suit departmental needs and data analysis requirements. Each user-query is then manipulated for each report. Each query is built as a SQL statement whose results can be in graphical format or raw data format. A test SQL function is provided to ensure the query executes correctly against the database. The User Query Interface obtains Report Templates from the Report Template Database and builds the user interface for presentation to the user (using HTML for display on a browser). Finally, from the raw data received from the results of User Query, Report Presenter will then format reports to suit user needs. Drill-down reporting allows more detailed view of data.

5 Performance Testing

From Fig 5.1, there were a total of 103574 records for the month of January 2003. The implementation of hard coded IF-Else statements shown in Fig 1.1 produced 5867 error records which meant there was a 5.36% error rate. The rule-based approach through the use of wizards to modify rules produces 1352 error records even after rule adjustment due to inability to filter the CDR. This translates to a 1.28% error rate. This shows a slight improvement over the old system. The disadvantages of naive approach are that the hard coded rules are difficult to change and usually can only be modified by shutting down the server and examining the error logs. A rule-based system does not require a shutdown of the system since the rules can be compiled by the CDR Processor immediately when new rules are added. CDRs with recurring errors are also accumulated and presented to the user with the option to add in new CDR filter rule based on closest rule match.

In this section an approximation to the CDR record loss rate at different simulated call traffic intensities will be calculated. This approximation is made to investigate the limitations of using the serial interface for output of CDR data. The approximation is based on a simple model of a queuing system; a M/M/1*B system. This system assumes exponentially distributed interarrival times and exponentially distributed service times, using only one server and having a limited buffer space of B buffers.

Here, the arrival process represents the arrival of newly generated CDR to the output buffer and the service process represents the process of transmitting the CDR over the serial connection. Exponentially distributed interarrival times are often used in telephony based queuing systems to model the arrival process and has shown to often be a very good approximation (see [8]). The time to transmit the data over the serial line is intuitively constant in this case, since the size of each CDR and the
transmitting rate are constant. However, as also mentioned in [8], systems with
general and deterministic service times can often be very closely approximated using
exponentially distributed service times in the model. By using exponential
distributions rather than general, the calculations can be simplified but still be
accurate enough to be used for approximating the limitations of the system.

| Type of Call      | Number of Calls |
|-------------------|-----------------|
| Outgoing Call     | 53310           |
| Incoming Call     | 39800           |
| Initial Connection| 4106            |
| End               | 1789            |
| Internal Call     | 1726            |
| Start             | 1438            |
| Outgoing T2T      | 945             |
| Incoming T2T      | 460             |
| Total             | 103574          |

Fig. 5.1. Typical Records from PABX for one month

The CDR loss rate was calculated for different arrival intensities and plotted in a
graph (see Fig 5.2). From the graph it can be determined that the CDR loss rate may
be neglected when the CDR arrival rate is less than close to 4 CDR per second. When
the arrival rate reaches 4 CDR per second, the output buffer starts to fill up and CDRs
are lost. At stress testing, call traffic generates a maximum arrival intensity of
approximately 1 CDR record per second, which is far lower than the critical arrival
intensity when call information records begin to get lost. Even if the traffic load
increases to three times the traffic load of today, there’s no immediate risk of losing
CDRs due to saturated output buffers.

From Fig 5.3 we can see that this arrival rate; at the point when the output buffer
starts to fill up, corresponds to a traffic intensity of about 80%.

6 Summary and Conclusion

The CDR transaction model we have assumed in our study is one in which the CDR is
produced only after a call has been placed. The first step in automatic CDR filtering is to
identify the data fields that comprise the CDR format and which can be used to identify
the type of CDR being produced. In the particular case of the Nortel Meridian One PABX
[5], five different call data were identified that are critical for call reporting. The
architecture that is proposed will allow for line activity analysis while continuously
publishing action availabilities in real-time. For performance evaluation, an approximation
to the CDR record loss rate at different simulated call traffic intensities was calculated.
From the results, we observe that the CDR loss rate is negligible when the CDR arrival
rate is less than 4 CDR per second. At stress testing, call traffic generates a maximum
arrival intensity of approximately only 1 CDR record per second, which is far lower than
the critical arrival intensity when call information records begin to get lost.
Fig. 5.2. CDR Record Rate Loss

Fig. 5.3. Traffic Intensity

References

[1] TelSoft Solutions for Call Accounting, http://telsoftsolutions.com/callaccount.html (verified on 15 Jan 2004).
[2] Nguyen A. T., J. B. Zhang, J. C. Tay, “Intelligent management of manufacturing event & alarms”, technical report, School of Computer Engineering, Nanyang Technological University, Jan 2004.
[3] Reference for Nortel Meridian 1 PBX/Meridian Link Services, http://callpath.genesyslab.com/docs63/html/norts11/brsl1m02.htm#ToC (verified on 15 Jan 2004).
[4] Nortel Networks. (2002) CDR Description and Formats. Document Number: 553-2631-100 Document Release: Standard 9.00.
[5] Nikbakht, E. and Tafti, M.H.A, Application of Expert Systems in evaluation of credit card borrowers. Managerial Finance 15/5, 19-27, 1989.
[6] Peter B., John S., Yves M., Bart P., Christof S., Chris C., Fraud Detection and Management in Mobile Telecommunications Networks, Proceedings of the European Conference on Security and Detection ECOS 97, pp. 91-96, London, April 28-30, 1997. ESAT-SISTA TR97-41.
[7] Java Expert System Shell or JESS, website at http://herzberg.ca.sandia.gov/jess/
[8] Jain, Raj, The Art of Computer Systems Performance Analysis. ISBN 0-471-50336-3, USA: John Wiley & Sons, Inc., 1991.
Adaptive Agent Integration in Designing Object-Based Multiagent System

Jaya Sil

Computer Science & Technology Department,
Bengal Engineering College (Deemed university),
711103 Howrah, India
jayasil@hotmail.com

Abstract. The aim of the paper is to design a multiagent system in an object oriented paradigm where agents jointly contribute to solve a problem based on their expertise and the information available in a blackboard system. The problem knowledge base is represented by ‘If-Then’ rules and a scheduler integrates the rules with the agents act as objects. The problem has been formalized into a tree structure. A tree traversal algorithm has been presented in the paper that determines the scope of contribution of the agents for solving the problem. If the required agent is not available or its contribution is not satisfactory, the system with the help of Kohonen network identifies the winner agent for replacement resulting overall better system performance.

1 Introduction

The features of object-oriented programming (OOP) are exploited in multiagent system development [3] to solve the real world problems. Though OOP is weak in inference generation process but has the central idea of encapsulation [2] and reuse, which encourage modular programming development [5]. On the other hand, rule-based programming expresses relationships between objects very explicitly, however, they don’t express updating clearly. The proper integration of agents (objects) with logic rules provides an extremely flexible and powerful environment, as rule-based components facilitates deductive retrieval and pattern matching while object-oriented components act as agents, bind data structure and their operations that can be applied together into one package. Two different strategies are reported till date for integrating object-oriented and rule-based programming paradigms. The first strategy [6], logical emulation of objects emphasizes the process view of the system by logical rules. The second strategy [8], object emulation of predicates describes the class methods as a set of logical clauses where predicates are specified within clauses. But none of these approaches deal with the degree of ability of the agents to contribute through interaction with others in order to face the increasing complexity of applications. Further, the availability of the agents is not always known a priori in the dynamic environment or its contribution is not satisfactory enough even if it is available. These issues are also not tackled in the current literatures.
The multiagent system describes in the paper exploits expertise of different agents and their ability to contribute with respect to teamwork while executing the tasks to solve the problem, taking blackboard [1] as their work place. Using standard software engineering approaches, a problem is decomposed into different sub-problems where each sub-problem consists of several tasks. The smallest executable unit is termed as task represented by ‘If-Then’ rules. A scheduler based on the expertise of agents integrates the object-based agents with logic rules while contribution of the agents are obtained with the consultation of the domain experts. Once the tasks are embedded into different agents, a tree is constructed to formalize the problem.

The paper suggests a tree traversal algorithm based on depth first search strategy to identify the agents, which are not available in practice though embedded in the rules by the scheduler. Further, the available ones may not contribute satisfactorily to implementing the solution. To tackle the situation, Kohonen network [4] is employed here to substitute or improve the contribution of the agent and thus, the overall performance of the system has been improved without invoking any proxy agent.

The paper has been organized into six sections. Problem definition using tree structure and formation of knowledge object has been described in section 2. In section 3, the tree traversal algorithm is presented. The multiagent learning using Kohonen’s competitive learning rule is focused in section 4. Section 5 illustrates the application of such a system while conclusions are arrived at section 6.

2 Knowledge-Object

Using AND-OR graph search strategy a problem is decomposed into sub-problems, each sub-problem further decomposed and the process continues until some simple tasks can be attained by means of the basic methods. The tasks are expressed in terms of logic rules using ‘If-Then’ clauses. The left hand side (LHS) of the rule represents the preconditions while its right hand side (RHS) represents the effect (consequent) of execution of the rule by the appropriate agent. The domain expert assigns contribution to the agent to executing a particular task after analyzing its expertise. Agents try to perform using blackboard as their work place where the rules and initial information are written. An agent independently or jointly acts whenever an opportunity arises to apply the expertise based on the available information and activity of the agents generate new information, which is recorded to the blackboard system. This additional information may trigger other agents to apply their expertise and thus the process continues until the goal state is achieved.
The problem has been formalized (Fig. 1) into a tree structure where the root node of the tree represents the goal state. In Fig. 1, \( c_1 \) and \( c_2 \) denote the percentage of contributions of Agent 1 and Agent 2 respectively for attaining the problem state \( s_1 \), represented by the node of the tree. The state \( s_1 \) is achievable provided the sum of \( c_1 \) and \( c_2 \) is 100. The tree can be extended up to any level depending on the problem under consideration.

An agent node is mapped as child of the respective goal state/state nodes. Contribution of an agent towards achieving a particular state is mapped at the arc connecting the agent node and the respective goal state/state nodes. A scheduler integrates a rule to the agents provided the agents could apply their expertise based on the information available in the environment. If more than one agent is available for the same task then the agent with highest contribution is embedded into the rule. Thus, the concept of knowledge-object has been utilized in the paper.

3 Tree Traversal Algorithm

Once the tasks are distributed over the agents by the scheduler, it is now turn of the agents to perform the tasks. For all practical purposes the agents should cooperate with one another because there is a need to meet the global constraints and no agent

---

**Fig. 2.** Flowchart for tree traversal algorithm
has sufficient competence to achieve the goal by its own. The algorithm traverses the
tree to ensure the availability of the appropriate agents along with their performance
for evaluating the solution of the problem. Whenever an agent scheduled to perform
the task is not available, the algorithm stores the agent in a list. Moreover, during
traversal the sum of contribution of the agents to achieving a particular state is
checked and in case it is less than 100, stored in a separate list. Assume an agent
node and a state node of the tree reside at the agent level (AL) and the state level (SL)
respectively. The flowchart shown in Fig. 2 describes how the tree is traversed at a
particular depth (level) and on receiving only the leaf nodes (agent nodes) the
algorithm backtracks to the immediate lower AL and examines other agent nodes.
The traversing process continues recursively till SL becomes the root node.

3.1 Tree Traversal Program

Program Tree_Traversal (input)
T //tree representing the problem
A //list of available agents in the environment
c_i //contribution of Agent i (ag_i)
L_i //ag_i situated at L_i
(output) List S //store agents, which are not available
List Q //contribution not satisfactory
begin
  take the root node;
  if (the root node has no child) report failure
  else
    begin
      AL=-1; depth_first_search;
      repeat
        AL=AL-1; ag_i=RHS(ag_i);
        if (ag_i≠null)depth_first_search;
      until (AL=0)
    end;
  end.
depth_first_search
begin
  down_to_leaf; ag_i=RHS(ag_i); check_upto_null;
end.
down_to_leaf;
begin
  repeat
    AL=AL+1; select ag_i; //extreme LHS agent node
    sum=0; check_contribution (ag_i, c_i);
    until (child(ag_i) = leaf_node);
    return (ag_i);
  end.
check_upto_null (ag_i)
begin
  repeat
if (child(\(ag_i\)) \(\neq\) leaf_node) down_to_leaf
else
    \(ag_i\) = RHS(\(ag_i\));
until (\(ag_i\) = null);
return (LHS(\(ag_i\)));
end.

check_contribution (\(ag_i\), \(c_i\))
begin
    if (\(ag_i\) \(\in\) A) then
        begin
            sum = sum + \(c_i\);
            if (sum \(\neq\) 100) then
                begin
                    \(ag_k\) = RHS(\(ag_i\));
                    if (\(ag_k\) \(\neq\) null) check_contribution (\(ag_k\), \(c_k\))
                    else
                        begin
                            store \(ag_i\) in Q; return (Q);
                        end;
                end;
            else
                begin
                    store \(ag_i\) in S; return (S);
                end;
        end;
    else
        begin
            store \(ag_i\) in S; return (S);
        end;
end.

4 Multiagent Learning

Assume, the scheduling process suggests to employ \(m\) specialized agents to determine the solution of the problem while each agent is characterized by \(p\) attributes, belonging [0, 1]. Thus, the input vector \(X_{(px1)}\) has been formed and \(m\) such input vectors are available to train the Kohonen network. It is worth to mention here that in order to cluster more than one agent into a single output node, the network should have output nodes \(n\) at least one less than \(m\).

The output \(y\) is computed by \(y = a(W_i'X)\), where \(a(.)\) is a continuous activation function and \(W_i=(w_{i1}, w_{i2}, \ldots, w_{im})'\), connection strength between input and output neurons. The training set \{ \(X_{(px1)}^1, X_{(px1)}^2, \ldots, X_{(px1)}^m\) \} represents \(n\) clusters, but there is no information regarding which input vector belongs to which cluster. Hence the Kohonen network classifies \(m\) input training patterns into one of the specified \(n\) categories according to the clusters detected in the training set. The learning rule is described by a two-stage computation, similarity matching that detects the winning neurons and updating of weights of the winning neurons. The learning process stops when two consecutive weight changes are small enough. After training the agents in list \(S\) (output of the tree traversal algorithm) are likely to be substituted by the available agents in the environment, clustered in the same class and the contribution of the selected agent of list \(Q\) is modified so that the sum of contribution is 100.
5 Application

This section illustrates the distributed approach of the multiagent system in the domain of flood control management. The current state of sensors and communication technology provide information about the wind state, rainfall, water level of the watersheds etc. The problem has been decomposed into sub-problems such as local emergency management, dam control management and relief management while each sub-problems consists of several tasks to evaluate the current situation and suggests potential action plans that can adequately control a real-time risk management problem. Different types of agents along with their contributions are identified to tackle the situation described in Fig. 3.

![Fig. 3. Tree structure representing flood control management system](image)

Local agent (LA) raises alarm to the dam_control agent (DCA) who is responsible for dam control management. Similarly the fire_brigade agent (FBA) is for population evacuation and resource protection while the transport agent (TA) is for the viability in the road transport network. The health agent (HA) is responsible for providing medicine and the food management agent (FA) is to feed the affected people. Finally, the volunteer agent (VA) takes part in rescue operation. Each agent is characterized by several attributes say, knowledge, experience, communication_skill, machinery and rate_of_success belonging to [0,1], which together form the input vector to train the Kohonen network. After training seven (no. of agents) input vectors are mapped into five clusters; cluster 1: FA and VA, cluster 2: TA and LA, cluster 3: HA, cluster 4: FBA and cluster 5: DCA.
Result: At any particular instant of time FA is not available and the contribution of the TA is not satisfactory, identified by the tree traversal algor. FA is replaced by VA (clustered together) while LA (20) contributes along with the TA (20) to cope up the situation. Thus, the system dynamically tries to cope up the situation until normalcy is returned in the environment.

6 Conclusions

The paper aims at developing a multiagent system by utilizing the concept of contribution with respect to teamwork where agents play a significant role in distributed problem solving environment. The agents are implemented as objects using Java [2] while the domain knowledge base is designed in MS-Access and the connection between these two has been established with the help of the ODBC driver. A tree traversal algorithm has been designed for identifying the non-available and less credible agents from the system. The Kohonen network is used to replace the agents with available and more efficient ones at the time of emergence. The multiagent system has been illustrated using an environmental emergency management problem where agents operate in a highly dynamic and unpredictable [6] environment.

References

1. Corkill, D. D., Gallagher, K.Q., Murray, K.E., GBB: A generic blackboard development system, Proc. AAAI-86, Philadelphia, PA, (1986) 1008-1014
2. Freeman, A., Ince, D.: Active Java, reading, MA, Addison-Wesley Publishing (1996)
3. Jennigs, N.R., Sycara, K., Wooldridge, M.: A roadmap of agent research and development, Vol. 1, Autonomous Agents and Multi-Agent Systems, (1998) 7-38
4. Kohonen, T.K.: Self-organization and associative memory, (3rd. Eds), NY, Spinger-Verlag, (1989)
5. Maes, P. (ed.): Designing autonomous agents, The MIT Press: Cambridge, MA, (1990)
6. Martin, D.L., Cheyer, A.J., Moran, D.B.: The open agent architecture: A framework for building distributed software systems, Vol. 13(1-2), Applied Artificial Intelligence, (1999) 92-128
7. Singh, M.P.: Considerations on Agent Communication, presented at FIPA workshop, (1997)
8. Sycara, K., Decker, K., Pannu, A., Williamson, M., Zeng, D.: Distributed Intelligent agents, Vol. 11, IEEE Expert, (1996) 36-46
Ontological Representations of Software Patterns

Jean-Marc Rosengard and Marian F. Ursu
Department of Computing, Goldsmiths College, University of London
London, SE14 6NW, UK
jm@syronex.com, m.ursu@gold.ac.uk

Abstract. This paper is based on and advocates the trend in software engineering of extending the use of software patterns as means of structuring solutions to software development problems (be they motivated by best practice or by company interests and policies). The paper argues that, on the one hand, this development requires tools for automatic organisation, retrieval and explanation of software patterns. On the other hand, that the existence of such tools itself will facilitate the further development and employment of patterns in the software development process. The paper analyses existing pattern representations and concludes that they are inadequate for the kind of automation intended here. Adopting a standpoint similar to that taken in the semantic web, the paper proposes that feasible solutions can be built on the basis of ontological representations.

1 Introduction

Software patterns are proven solutions to recurring software construction problems in a given context. They describe the knowledge refined from experienced practitioners about a particular aspect of the domain of discourse. The concept of design pattern was formulated in [1] and has since been generalised in for example [2]. The practical employment of patterns in software development has continuously grown [3].

Patterns are generally intended for human/manual use, as structured but informal documentations. Their primary aim is to guide software engineers, by presenting and explaining solutions regarding software construction applicable to a particular context. They are means of structuring solution descriptions. From this angle, therefore, they could be regarded as codes of good practice. This is the perspective that we take in this paper. In this context, patterns are presented in printed catalogues, employing mainly natural language, but also examples of code and diagrammatic representations.

Patterns can occur at different levels of abstraction with respect to the specification of a software solution. Thus, there may be architectural patterns, design patterns in a specific design paradigm (e.g. observer and composite [4] in the OO paradigm) and language-specific patterns (e.g. counted body [5] in C++, and marker interface in Java). The latter can also be called idioms. Furthermore, software patterns may have different degrees of generality. Some may
be application or domain specific (e.g., contract and portfolio in trading applications), whereas others may be general design patterns, applicable across application domains (e.g. observer and composite).

In recent years, software development using patterns has become common practice in the practitioners’ community. Subsequently, the amount of refined patterns is growing, irrespective of their category—from general language-specific patterns to application specific patterns. However, it is towards the application-domain end that a high rate of growth is expected. The amount of printed documentation, thus, too, is increasing, to the extent that it becomes difficult for it to be effectively used. The problems that appear in this context are similar to the problems faced by engineering designers who have to comply with large codes of regulations and good design practice, which we have already discussed in [7].

We aim to develop tools for intelligent dissemination of patterns to software practitioners. We propose a general model that is applicable to patterns disregarding their level of abstraction (specification) and generality (application domain). On its basis we will implement specific solutions for different categories of patterns. We are here adapting some of the solutions we have proposed previously for the dissemination and enforcement of regulatory engineering-design knowledge (e.g., [7]) to the domain of software engineering and software patterns.

Research has been investigating the possibility of automatic code generation from formal representations of software patterns [8]. The goal, according to this approach, is to reduce as much as possible the involvement of the human agent from the design and implementation process. This may be feasible for restricted areas. However, our standpoint is to develop tools that empower rather than replace the software practitioners; “patterns should not, cannot and will not replace programmers” [2]. This is consistent with our previous work in intelligent design [9].

2 Software Patterns: Evolution

In their early years, patterns have been mainly used within the community close to the group that described them. A number of fundamental patterns have been refined, in particular at the level of design, and are now widely used among software engineers. They are involved in the construction of most medium-size and large object-oriented systems. Some have also been integrated in programming platforms, such as Java, becoming thus readily available for application programming.

As the result of almost a decade of pattern mining, a large quantity (hundreds) of patterns have been described, reviewed, and catalogued. However there have been few initiatives to structure and organise this knowledge into a consistent representation framework.

The rate of growth varies with respect to the level of abstraction—with reference to the specification of a solution—but more so with the level of generality—with reference to the reusability across application domains. General or core pat-
terns tend to be considered as fundamental abstractions and, in time, become integrated into programming languages and tools. Their number is limited and essentially constant. A rate of growth is displayed by patterns specified at the level of middleware. This is because software applications are increasingly complex and, thus, have to be developed around middleware platforms (e.g. J2EE).

A higher rate can be predicted at the level of particular application domains or within particular software development companies. Patterns can naturally describe expertise regarding a specific software development application. Furthermore, they can also naturally express specific policies regarding software development within different organisations. The focus, here, is on promoting the use of explicit, locally defined constructs, motivated by concerns like quality, security, performance or code management.

Domain specific patterns is the category that strongly motivates our work. Because they represent a natural way for the formulation of accumulated expertise and policies, we consider that they will become the means for the representation of such knowledge. Consequently, large knowledge repositories of domain specific patterns will be created (both within organisations and for general use). Furthermore, domain specific patterns form a dynamic pool of knowledge. They are expected to evolve more rapidly than the more generic ones, because the requirements within application domains are under continuous change, and their review and publication process can be expected to be less rigorous.

At this end, manual use of patterns is not an effective solution anymore. Their expected development depends on the existence and motivates the development of tools for their automatic organisation, retrieval and explanation. By development we mean both refinement/statement and employment/use. The latter term encapsulates all the various cognitive activities involved in software development—such as understanding whether to use or not a pattern applicable to a given context, choosing a particular pattern suitable to a specific context and understanding how to generate code in accordance to a particular chosen pattern—and sharing.

3 Existing Pattern Representations

This section discusses existing representations of patterns and their suitability to automatic organisation, retrieval and provision of explanations.

3.1 Informal Representation

Patterns are most generally represented in natural language, and are typically published in printed catalogues. The term “presentation” seems more suitable for this type of description. Such documents are loosely structured, in what we call canonical forms. Such a structure consists of a series of fields, each having a meaning introduced via an informal definition or description. An example of a canonical form is that proposed in [4]. A fragment of this is illustrated in Table 1 below.
Table 1. Fragment of a canonical form for pattern representation [4]

| Field             | Explanation / Definition                                                                                                                                 |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Name              | Ideally a meaningful name that will be part of the shared design vocabulary. Many existing patterns do not satisfy this requirement for historical reasons. |
| Also known as     | Other names of the pattern.                                                                                                                                   |
| Intent            | A short specification or rationale of the pattern, used as a principal index for goal-oriented pattern search.                                              |
| Applicability     | An outline of the circumstances in which the pattern may be applicable and, perhaps more importantly, when it should not be applied.                       |
| Structure         | A diagrammatic representation of the pattern.                                                                                                               |
| Consequences      | Discusses the context resulting from applying the pattern. In particular, trade-offs should be mentioned.                                                   |
| Implementation    | Advices on how to implement the patterns, and other language specific issues. The implementation will depend on the abstractions (objects, parameterised types, ...) supported by the target language. |
| Known uses        | Patterns are by essence derived from existing systems. It is therefore important that they be justified by their use in several real systems.             |
| Related patterns  | Patterns are often coupled or composed with other patterns, leading to the concept of pattern language; e.g. a visitor may be used to apply an operation to the closed structure provided by a composite. |

Consider, for example, the most common situation when a software developer is within a specific situation and wants to identify whether there exists a particular pattern useful to the situation at hand. A search within a repository of patterns would, most probably, involve the intent and applicability descriptors. Assuming that the catalogue exists in an electronic format that preserves the structure of the printed catalogue, as described above, then the best option available to him is a keyword search; intent and applicability have no internal structures. This means that the software engineer attempts to retrieve documents describing relevant patterns based on phrases that he would have to guess are used in the descriptors. Each time a document/pattern is retrieved, he would have to read it thoroughly—since no summarisation or explanatory features would be supported by the discussed representation—and decide upon its suitability. Obviously, this is a cumbersome process.

The drawbacks of such a retrieval process are well known. They are more critical if the agent who carries out the process does not have at least some knowledge of the descriptions’ jargon or of the possible expected results; in our case, if the software engineer is a novice. Note that by novice, we mean without much software development experience, but also inexperienced with a particular technology, or new to a company and not familiar with its policies and codes.

These drawbacks have been identified and well described in the context of the web and represent a motivating factor for the development of the semantic web [10].
Although they have the same nature, the scale of the problem in the context of software patterns is obviously not as large as in the context of the web. However, the effects can be similarly drastic, under the founded assumption that the pattern repository has a substantial size. Furthermore, missing or misusing a pattern with respect to a particular situation could have severe implications if the patterns represent company policies or codes of best practice. The above argument was implicitly carried out at the level of application/domain specific patterns. However, it is equally valid in the context of domain independent, but language-specific patterns (idioms). A good example for this is the Javadoc documentation of the Java platform. This knowledge base is significantly large and finding relevant solutions to a specific problem is a cumbersome process for non-expert Java programmers.

Another major drawback of this representation is the fact that it does not readily support knowledge management and sharing, also necessarily required, in particular for application-domain patterns (refer to Section 2). Informal representations based on canonical forms cannot support the level of automation at which we aim. For this, we need better-structured representations.

3.2 Patterns in UML

Patterns are represented in UML using the familiar class/object and interaction diagrams, and also using the more specific parameterised collaboration model \[11\]—allowing the variation of roles around a given collaboration. While these representations are useful for understanding a pattern and guiding through its implementation, they only express the structural aspects of the pattern. They do little to help the engineer understand its higher-level concerns, like its intent, applicability and tradeoffs. Unsurprisingly, UML is not suitable for pattern representation for the purpose stated in this paper. As a simple example, consider the strategy and state patterns. Although their intents \[4\] are very different, they exhibit a similar structure.

3.3 Formal Representations

Although patterns primarily constitute a body of knowledge for human consumption, several initiatives have been made to formalise some aspects of their representation, opening the way to some automated support for pattern-based development.

Formalisation is applied to some of the essential properties of patterns (pattern invariants) by means of specification languages, like the Object Constraint Language (OCL) \[12\]. On the instantiation of a pattern or the modification of an existing occurrence of a pattern, its implementation may be automatically validated using the structural and behavioural/temporal constraints specified in OCL expressions. Such representations, although useful in such situations, cannot express all the knowledge (that a pattern encapsulates) required for instantiations or modifications. Furthermore, aspects like pattern intent, motivation and applicability, cannot be expressed in OCL.
Constraint languages and ontologies (proposed here) are complementary in pattern representation. The former are focussed on automatic code generation, whereas the latter are focussed on the provision of intelligent advice to software developers. Also, constraints and ontological representations operate at different stages—expressions represented in constraint languages become applicable after the engineer has made solution decisions.

4 Ontological Representation

There are various meanings that the term ontology can have in AI [13]. We adopt the same view as [14] and take ontology first to mean a specification of a conceptualisation, and second—pragmatically—to define a (standard) vocabulary with which queries and assertions are exchanged among agents. Ontologies are particularly useful in knowledge sharing and reuse. If two agents make the same ontological commitment, then they can exchange knowledge. Alternatively, any knowledge base represented within an ontology can be accessed by agents who committed to the respective ontology. The latter viewpoint is relevant to our proposal. If an ontology for the representation of patterns is in place, then pattern repositories (represented in the respective ontology) become accessible by various tools—for intelligent organisation, retrieval and provision of explanations—provided they committed to the ontology.

An ontology becomes standard within a community when sufficient (or sufficiently powerful) agencies from that community commit to it. The quality of an ontology being standard is only required if knowledge reuse and sharing is an issue within the community. sharing and reuse should be understood, in the context of software patterns, with respect to the type of the patterns. Idioms should be shareable across application domains, whereas application specific patterns may need to be shared only at the level of an institution.

As a method of work, we started with the development of a basic ontology for design patterns. They are of a manageable size and their generality is implicitly transferred to the basic ontology. Thereafter, we shall enhance the basic ontology with language-specific concepts and domain-specific concepts, when we move towards the representation of the respective software patterns.

Although we do not necessarily intend that the deployment of documentation based on patterns be made within the web, our work, here, is strongly connected with that carried out within the semantic web [15]. The use of ontologies was proposed in software engineering, but in the context of component based development. The focus of these efforts (e.g., [16]) is on automatic retrieval and assembly. Our focus is on the provision of intelligent advice to software engineers.

5 Conclusions

In this paper, we introduced the idea of combining software patterns with ontological representations, with a view to developing tools for the automatic
organisation, retrieval and explanation of reusable solutions to software development, codes of good practice and company policies.

References

1. Gamma, E., Helm, R., Vlissides, J., Johnson, R.E.: Design patterns: Abstraction and reuse of object-oriented design. In Nierstrasz, O., ed.: Proceedings ECOOP '93. Volume 707 of LNCS., Springer-Verlag (1993) 406–431
2. Coplien, J.O.: Software Patterns. SIGS, New York (1996)
3. Gamma, E.: Design patterns ten years later. In Broy, M., Denert, E., eds.: Software Pioneers: Contributions to Software Engineering, Springer-Verlag (2001) 689–699
4. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1995)
5. Coplien, J.: Advanced C++: Programming Styles and Idioms. Addison-Wesley (1992)
6. Fowler, M.: Analysis Patterns: Reusable Objects Models. Addison-Wesley (1997)
7. Ursu, M.F., Hammond, P.: Representing expressed design knowledge. In Benbasat, I., ed.: Association for Information Systems, AIS’98. (1998) 393–395
8. Eden, A., Yehudai, A., Gil, J.: Precise specification and automatic application of design patterns. In: 1997 International Conference on Automated Software Engineering, IEEE Press (1997) 143–152
9. Ursu, M.F., Hammond, P.: Expressing regulatory design knowledge for critiquing intelligent design assistants - spacial representations. In Gero, J., ed.: Artificial Intelligence in Design, AID’00, Kluwer (2000) 103–126
10. Helsinki Institute for Information Technology (HIIT): Semantic web kick-off in Finland: Vision, technologies, research and applications (2002)
11. Rumbaugh, J., Jacobson, I., Booch, G.: Unified Modeling Language Reference Manual. Addison-Wesley (1998)
12. OMG: Object Constraint Language Specification, version 1.1. (1997)
13. Uschold, M., King, M.: Towards a methodology for building ontologies. In: Basic Ontological Issues in Knowledge Sharing, IJCAI’95, Montreal, Canada (1995)
14. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5 (1993) 199–220
15. Wahlster, W., Lieberman, H., Hendler, J.: Spinning the Semantic Web: Bringing the WWW to Its Full Potential. MIT Press (2002)
16. MIT: Project oxygen (2004) http://oxygen.lcs.mit.edu/Software.html.
Dynamic Traffic Grooming and Load Balancing for GMPLS-Centric All Optical Networks

Hyuncheol Kim¹, Seongjin Ahn², and Jinwook Chung¹

¹ Dept. of Electrical and Computer Engineering, Sungkyunkwan University, 300 Chunchun-Dong Jangan-Gu, Suwon, Korea, 440-746
{hckim, jwchung}@songgang.skku.ac.kr

² Dept. of Computer Education, Sungkyunkwan University, 53 Myungryun-Dong Jongro-Gu, Seoul, Korea, 110-745
sjahn@comedu.skku.ac.kr

Abstract. Fast and resource optimized lightpath restoration strategies are urgent requirements for the near future Optical Transport Networks (OTN) with a Generalized Multi-Protocol Label Switching (GMPLS) control plane. In this paper, we propose a lightpath selection scheme with Shared Risk Link Group (SRLG) constraints that guarantees fast and load balanced recovery of lightpaths in GMPLS-centric all optical networks. To this, we propose a lightpath allocation and recovery strategy called “Enhanced Reverse Shared Risk Link Group (E-RSRLG)” and demonstrate how this concept can be applied to minimize recovery blocking probability and network convergence time. A 32-node network topology has been set up in OPNET™ to analyze the effects of E-RSRLG handling in terms of two performance factors: the number of recovery contention and the number of extra hops.

1 Introduction

Along with development of various transmission technologies including Dense Wavelength Division Multiplexing (DWDM), it is self-evident that the future integrated services network will include various switching techniques at various levels of the capacity hierarchy, from the wavelength up to the packet level. GMPLS is intended for such multipurpose signaling paradigm that can be used at various switching levels. It supports not only packet switching devices, but also devices that perform switching in the time, wavelength, and space domains [1]. In optical networks, GMPLS provides essentially routing, signaling and link management protocols for dynamic setup and recovery of lightpaths. In GMPLS terminology this lightpath is called Label Switched Path (LSP). Recently, there has been many detailed investigations toward establishing a survivability framework for GMPLS. However, the survivability framework considered the relationship only between one originator and one end, and did not consider other originators that can exist in a network [2]. Also, it did not even mention the detailed methodology of network load balancing [3][4][5].
This paper takes the aim at investigating the lightpath allocation strategies for fast and load balanced restoration in GMPLS, focusing on new specific information, the Reverse Shared Risk Link Groups (RSRLGs). To this, we had proposed a restoration scheme called "Reverse Shared Risk Link Group" [3]. In this study we propose enhanced RSRLG and demonstrate how E-RSRLG concept can be applied to minimize reconfiguration cycle and recovery time.

The rest of this paper is organized as follows. In section 2, we describe our Enhanced Reverse Shared Risk Link Group path selection scheme for survivable lightpaths. In section 3, the optical network model implemented in OPNET is described and the performance is evaluated and compared with the conventional approach. Finally, the paper concludes in Section 4.

2 Enhanced RSRLG

Diverse working and recovery paths with SRLG are used in GMPLS to increase survivability and to minimize the failure interference against various network faults. However, only a few algorithms have addressed in finding maximum fully disjoint recovery lightpaths with considering SRLG information [6]. That is, the number of fully disjoint lightpaths is limited by few. Moreover, all disjoint LSPs must start and end at the same nodes as shown in Fig. 1. When the source nodes are different, which is the typical post failure scenario for lightpath restoration, coordination among the source nodes is difficult, if not impossible [2][7]. It can cause significant performance degradation in case of a fiber failure because lots of requests are generated nearly simultaneously. The most influential solution is alleviating the restoration contention probability with considering load balance at connection setup stage.

In E-RSRLG, an originator takes into account the number of ultimate source-destination pair that appropriating the link. In E-RSRLG, a link that has many same ultimate source-destination pair with a lightpath setup request is preferentially selected as the shortest path after all the disjoint lightpaths are consumed.

Consider an example of the network topology as shown in Fig. 2. In E-RSRLG, SRLG scheme is still applied without considering load balance until the entire disjoint lightpaths are dissipated as shown in Fig. 2. However diversity of connections can cause failure indication flooding at a failure. As shown in Fig. 3 to alleviate the recovery contention probability, in the case in which there is no
more fully disjoint lightpath that satisfy the setup requirements, the E-RSRLG scheme is selecting a path that is used often by the originator if it satisfies load balance constraints.

![Diagram of SRLG](image)

**Fig. 2.** Provisioning Connections - SRLG

![Diagram of E-RSRLG](image)

**Fig. 3.** Provisioning Connections - E-RSRLG

To describe the E-RSRLG, we define the following parameters:

- $C_{\text{comp}}(i, j)$ : Link cost for $L(i, j)$ that is used to find shortest path.
- $K(s, d)$ : Number of obtained (partially) disjoint lightpaths between node $s$ and $d$. $K(s, d)$ is calculated with taking SRLG into account.
- $L(i, j)$ : A fiber link between node $i$ and node $j$.
- $T$ : Maximum number of lightpaths that can be allocated between one ultimate source-destination node pair in $L(i, j)$. Load balancing factor.
- $rsrlg(i, j, s, d)$ : RSRLG information. $rsrlg(i, j, s, d) = 1$ means that $L(i, j)$ has a lightpath from $s$ to $d$ and $rsrlg(i, j, s, d) = 0$ means that it does not have.
Dynamic Traffic Grooming and Load Balancing

- \( \alpha \): Weight factor for RSRLG and can be used hybrid lightpath selection.
- \( G \): Number of ultimate source-destination pair in a network.
- \( M(s,d) \): Number of members for source-destination pair \((s,d)\) in a network.

E-RSRLG is described as:

\[
\text{do SRLG-based lightpath selection until all disjoint lightpaths are exhausted}
\]

\[
\text{for each path do}
\]

\[
\text{if } T_{\text{path}}(s,d) \text{ satisfy the request do}
\]

\[
\text{if } C_{\text{path}}(s,d) \text{ satisfy the request do}
\]

\[
\text{then Select a lightpath which have maximum } C_{\text{max}}(s,d)
\]

\[
\text{if more than one lightpath exist do}
\]

\[
\text{then Select a lightpath at random}
\]

\[
\text{return the path}
\]

where,

\[
M(s,d) = \sum_i \sum_j r_{srlg}(i,j,s,d)
\]

(1)

\[
RSRLG(i,j) = \frac{G}{g} M(s,d) r_{srlg}(i,j,s,d)
\]

(2)

\[
RSRLG_{\text{max}} = \max_{i,j} RSRLG(i,j)
\]

(3)

\[
C_{\text{path}}(s,d) = \sum_{(i,j) \in \text{path}} C_{\text{comp}}(i,j)
\]

(4)

\[
T_{\text{path}}(s,d) = \sum_{(i,j) \in \text{path}} T(i,j)
\]

(5)

\[
C_{\text{comp}}(i,j) = \frac{\alpha}{RSRLG_{\text{max}}} \max\{RSRLG(i,j), 1\}
\]

(6)

3 Performance Evaluation

In this paper, a distributed optical network architecture is considered. The reference network shown in Fig. 4 has been modeled under OPNET\textsuperscript{TM} Modeler. As shown in Fig. 4 the 9 Label Edge Routers (LERs) and 14 Label Switching Routers (LSRs) are located, which are supposed to be sources and destinations of huge amounts of traffic. The internal finite state machine (FSM) for the LSR is shown in Fig. 3 we mainly focus on connection request contention issues with an inaccurate network information that affects restoration performance.

The connection requests are uniformly distributed across all ingress-egress node pairs. They arrive at the network according to an independent Poisson
process and the connection holding time is exponentially distributed. For calculating the explicit route, ingress routers need to know the current topology and available capacities. The topology is known by link state routing protocols and failure of LSPs due to link failures is detected from signaling protocols. We also assume the following conditions:

- Lightpaths are assumed to be bidirectional.
- Fully capable of wavelength conversion on each of their ingress/egress port.
• There are 20 wavelengths per link.
• Message processing time at each node is constant (ignored).

The simulation results, as shown in Fig. 6 and Fig. 7, shows that when a fault occurs at the connections which request different bandwidths, instead of using the method of requesting connection with just the bandwidth first (BF) or of allocating channel randomly, it would be much better to use E-RSRLG which can reduce connection setup contention and recovery channel setup time in the entire network standpoint [2].
4 Conclusions

In this paper we introduce the concept of E-RSRLG and demonstrate how this concept can be applied to achieve diversity in optical networks and to assess the risks associated with a path. Conventional protection scheme does provide quick recovery time, but it has disadvantage of using up too much bandwidth and lack of ability to find sufficient disjoint paths. This paper proposes a new enhanced lightpath selection algorithm that overcomes these problems. The great advantage of the proposed algorithm is that it provides very fast recovery path compared to the conventional recovery methods. This approach to distributed provisioning and restoration will speed up provisioning and will pave the way for dynamic reconfigurability of the entire optical network. It is found that only a additional Open Shortest Path First-Traffic Engineering (OSPF-TE) message is needed to support the E-RSRLG scheme and a very fast restorability can be accomplished for a single failure.

References

1. Ayan Banerjee, John Drake, et al.: Generalized Multiprotocol Label Switching: An Overview of signaling Enhancements and Recovery Techniques, IEEE Communications Magazine, Vol. 7. (2001) 144–151
2. Wang Jianxin, Wang Weiping, et al.: A Randomized Qos Routing Algorithm On Networks with Inaccurate Link-state Information, GLOBECOM ’99, vol. 3, (1999) 1694–1698
3. Hyuncheol kim, Junkyun Choi, et al.: Analysis of End to End Recovery Algorithms with Preemptive Priority in GMPLS Networks, LNCS 2662, (2003) 118–122
4. Timucin Ozugur, Myung-Ah Park, et al.: Label Prioritization in GMPLS-Centric All-Optical Networks, ICC ’03 Vol. 2 May (2003) 1283–1287
5. Chunsheng Xin, Yinghua Ye, et al.: A Joint Working and Protection Path Selection Approach in WDM Optical Networks, GLOBECOM ’01, vol. 4, Nov. 1999 (2165–2168)
6. Eiji Oki, Nobuaki Matsuura: A Disjoint Path Selection Schemes With Shared Risk Link Groups in GMPLS Networks, IEEE Communications Letters, Vol. 6, (2002) 406–408
7. Yinghua Ye, Sudhir Dixit, et al.: On Joint Protection/Restoration in IP-Centric DWDM-Based Optical Transport Networks, IEEE Communications Magazine, Vol. 6. (2000) 174–183
Probabilistic Model of Traffic Breakdown with Random Propagation of Disturbance for ITS Application

Bongsoo Son¹, Taewan Kim², Hyung Jin Kim¹, and Soobeom Lee³

¹Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul, Korea
{hyungkim, sbs}@yonsei.ac.kr
²Chung-Ang University, 72-1 Nae-Ri, Ansung-Si, Kyunggi-Do, Korea
twkim@cau.ac.kr
³University of Seoul, 90 Jeonnnong-Dong, Dongdaemun-Gu, Seoul, Korea
mendota@uos.ac.kr

Abstract. In this paper, a probabilistic model of vehicular traffic breakdown applicable to Intelligent Transportation Systems (ITS) is presented. When traffic demand exceeds freeway capacity, the so-called breakdown occurs and the congestion begins. While preventing the breakdown is a major concern of traffic operation, the mechanism of breakdown is not thoroughly explained and most of the research regarding traffic breakdown rely on empirical analysis. To further our understanding of traffic breakdown, this paper explains the phenomenon of traffic breakdown in terms of random propagation of traffic disturbance and proposes a probabilistic model of breakdown. A Monte-Carlo simulation is also conducted to investigate the characteristics of the proposed model.

1 Introduction

Freeway management systems, one of ITS implementations, make use of traffic flow control strategies, and operational activities such as incident management and information dissemination. A freeway management system consists of infrastructure elements utilized to accomplish the goals and objectives of freeway management. These things include field hardware, communications equipment, a traffic management center, the people who staff the center, the policies and procedures established to deal with various transportation-related events that impacts the freeway system. Besides, fundamental understanding for the traffic flows and the models representing the reality are essentially required to integrate all of the above-mentioned facilities and functions. The emphasis of this paper is on developing a freeway traffic flow model that can be employed in ITS application.

Two different states of vehicular traffic are generally established to explain the movements of vehicles on a freeway. In free-flow state, interactions among vehicles are small and vehicles usually travel at the free-flow speed, which is close to the speed limit. However, as the traffic flow increases, the transition from free-flow state to the congested state occurs. Because traffic congestion brings about reduction of travel speed and annoying vehicular maneuver such as stop-and-go traffic, it is a major concern of traffic operation to prevent the transition, which is often called traffic breakdown.
In deterministic aspect, traffic breakdown is believed to occur when traffic demand exceeds the capacity. However, some empirical studies have found that the breakdown could occur when the demand is under the capacity and the occurrence of the breakdown follows a probabilistic procedure [1,2]. The breakdown of a freeway is believed to be triggered by a disturbance (temporary speed down) generated by merging vehicles or spontaneously [3,4]. However, the generation and development of the disturbance are quite stochastic and it is not easy to model the procedure of traffic breakdown. In this paper, we employ the wave propagation model, which is recently developed by Kim and Zhang [5] and develop a probabilistic model of breakdown. A numerical test of the model is conducted through Monte-Carlo simulation.

2 Traffic Breakdown Models

Due to the complex features of traffic breakdown, research on the breakdown is not rich. The first probabilistic model of breakdown was suggested by Bullen [1]. He suggested that the probability of breakdown (PB) is a monotonically increasing function of the freeway flow. Banks [3], from traffic data of four bottlenecks in San Diego, has observed that when a large number of vehicles in a cluster merge onto the freeway, breakdowns could occur at under capacity. He mentioned that the ramp flow rate (or the cluster size) is an important element that causes breakdowns. Elefteriadou et al. [2] have confirmed Banks’ finding and described the mechanism of breakdowns as 1) the entrance of a large vehicle cluster to the freeway 2) subsequent small speed drop at the merging point 3) the speed drop spreads upstream and creates a breakdown. Based on this breakdown mechanism a PB model is proposed, which is a smooth S-curve and a function of freeway and ramp flow rates (Fig. 1). Persaud et al. [6] have examined the PB from the traffic data of three sites on Canadian freeways. According to the results the PB drastically increases when the freeway flow is above the queue discharge flow (approximately 2800 vph). Another approach for the estimation of PB is using the Master Equation of Markov processes [7]. Based on the fact that the cluster formation in traffic flow is a Markov process, the Master Equation describes the time evolution

![Fig. 1. Probability of breakdowns in 15min by Elefteriadou et al. [2], reproduced](image-url)
of the probabilistic distribution of cluster size. In Kühne’s model, the PB also drastically increases at a certain density, however it does not exhibit the S-shape. Due to various traffic and geometric factors involved and their stochastic nature, modeling traffic breakdowns is not straightforward. Most PB models rely on empirical studies or over-simplified assumptions.

3 Traffic Wave Propagation

To develop the probabilistic model of breakdown, we introduce the wave propagation model recently developed by Kim and Zhang [5]. This wave propagation model explains the stochastic movements of traffic waves and amplification/decay of disturbances with the variations of gap time. For the wave propagation model, several variables to be used throughout this paper are defined. Fig. 2 (a) represents a snapshot of two successive vehicles at time $t$. Assuming the $(n-1)$th vehicle is ahead of the $n$th vehicle, we have four variables that describe the spatial and temporal separations between the two vehicles.

**Fig. 2. Notations**

- Spacing ($s_n(t)$) refers to the distance between the head of the leader and the head of the follower at time $t$.
- Gap Distance ($d_n(t)$) refers to the distance between the rear of the leader and the head of the follower at time $t$. $s_n(t) = d_n(t) + L_{n-1}$.
- Headway ($h_n(t)$) refers to the time taken for the follower to travel the spacing.
- Gap Time ($\gamma_n(t)$) refers to the time taken for the follower to travel the gap distance.
Let us assume the \( n \)th vehicle changed its speed from \( v^{(k+1)} \) to \( v^{(k)} \). After time \( \tau_{n+1}^{(k)} \), which represents the reaction time of the driver, the \((n+1)\)th vehicle also changed the speed from \( v^{(k+1)} \) to \( v^{(k)} \). Due to the speed change, the gap time of the \((n+1)\)th vehicle also changes from \( \gamma_{n+1}^{(k+1)} \) to \( \gamma_{n+1}^{(k)} \). Then the following equation holds:

\[
v^{(k+1)} \gamma_{n+1}^{(k+1)} + L_n + v^{(k)} \tau_{n+1}^{(k)} = v^{(k+1)} \tau_{n+1}^{(k)} + v^{(k)} \gamma_{n+1}^{(k)} + L_n
\]  

(1)

From (1), \( \tau_{n+1}^{(k)} \) is computed as

\[
\tau_{n+1}^{(k)} = \frac{v^{(k+1)} \gamma_{n+1}^{(k+1)} - v^{(k)} \gamma_{n+1}^{(k)}}{v^{(k+1)} - v^{(k)}}
\]  

(2)

The distance separation of the speed change between \( n \)th and \((n+1)\)th vehicle, \( x_{n+1}^{(k)} \), is

\[
x_{n+1}^{(k)} = v^{(k+1)} \tau_{n+1}^{(k)} - v^{(k+1)} \gamma_{n+1}^{(k+1)} - L_n
\]  

(3)

Suppose the first speed change occurred at time \( t=0 \) and location \( x=0 \). The time and location at which the \( n \)th vehicle changes its speed can be computed as:

\[
T_{n}^{(k)} = t_1^{(k)} + t_2^{(k)} + \cdots + t_n^{(k)}
\]

\[
= \frac{1}{v^{(k+1)} - v^{(k)}} \left[ v^{(k+1)} \sum_{i=1}^{n} \gamma_{i}^{(k+1)} - v^{(k)} \sum_{i=1}^{n} \gamma_{i}^{(k)} \right]
\]  

(4)

\[
X_{n}^{(k)} = x_1^{(k)} + x_2^{(k)} + \cdots + x_n^{(k)}
\]

\[
= \frac{v^{(k+1)} \gamma_{n}^{(k)}}{v^{(k+1)} - v^{(k)}} \left[ \sum_{i=1}^{n} \gamma_{i}^{(k+1)} - \sum_{i=1}^{n} \gamma_{i}^{(k)} \right] - \sum_{i=1}^{n-1} L_{i-1}
\]  

(5)

4 Probabilistic Model of Traffic Breakdown

On a freeway system, the breakdown of freely flowing traffic usually occurs at the on/off ramps where traffic disturbances are frequent. While some cases of spontaneous breakdown (the breakdown that is triggered by a disturbance spontaneously generated on a freeway) are observed, we limit our model for the breakdown triggered by merging vehicles. Here, we apply the aforementioned wave propagation model to derive the probability of breakdown on the one-lane freeway with on-ramp, a first step in obtaining a PB model for a multi-lane freeway. When a vehicle merges onto a freeway with free-flow traffic, it will travel at a lower speed before accelerating to free flow speed (see Fig. 3).

Responding to the movement of the merging vehicle, the following freeway vehicles would decelerate and then accelerate. This disturbance generates a deceleration and an acceleration wave, which propagates in a stochastic manner according to the distribution of gap time. If the deceleration is not eliminated before the second on-ramp vehicle merges onto the freeway, the second merging vehicle and the following
freeway vehicles would apply stronger deceleration than before. Finally free flow traffic would break down as more vehicles merge onto the freeway. On the other hand, if the disturbance by the first on-ramp vehicle has dissipated before the second on-ramp vehicle merges onto the freeway, there would be no breakdown. Let the average time headway of on-ramp vehicles be $\bar{H}_f$. If there exists $n$ th vehicle whose deceleration and acceleration time is the same, in other words it does not need to decelerate before the second on-ramp vehicle merges, there would be no breakdown (Fig. 3 (b)). Let the free flow speed be $v_f$, and the speed and duration of the initial disturbance of a merging vehicle be $v_0$ ($>0$) and $T_0$ ($>0$). The merging movement of a vehicle is usually stochastic depending on many factors such as freeway traffic headway, merging vehicle speed, and geometry of the merging section. We assume $v_0$ and $T_0$ follow probabilistic distributions $p(v_0)$ and $p(T_0)$, respectively. The breakdown is avoided when the acceleration and deceleration wave generated by a merging vehicle meet ($T_n^1 = T_n^2$) before the time next vehicle merges ($\bar{H}_f$). The PB is represented by

$$PB = 1 - \int_0^\infty \int_0^\infty p(v_0) p(T_0) \Pr \{T_n^1 = T_n^2 \cap T_n^1 \leq \bar{H}_f \} dT_0 dv_0$$

(6)

where, $T_n^1$ and $T_n^2$ are obtained from (4) producing

$$T_n^{(1)} = T_0 + \frac{1}{v_0 - v_f} \left[ v_0 \sum_{i=1}^n \gamma_i^{(2)} - v_f \sum_{i=1}^n \gamma_i^{(1)} \right]$$

$$T_n^{(2)} = \frac{1}{v_f - v_0} \left[ v_f \sum_{i=1}^n \gamma_i^{(3)} - v_0 \sum_{i=1}^n \gamma_i^{(2)} \right]$$

(7)

and $\gamma$ represents the gap time of the $i$th freeway vehicle, before meeting the disturbance $\gamma_i^{(3)}$, during the disturbance $\gamma_i^{(2)}$, and after recovering from the disturbance $\gamma_i^{(1)}$.

**An Example (Monte-Carlo Simulation)**

To obtain the general features of how the PB changes with respect to the flow rate of the freeway and on-ramp, we solve (6) for an example. It is not straightforward to

**Fig. 3. Breakdown on a freeway**
obtain an analytical solution of (6) and a Monte Carlo simulation is carried out. Since the traffic data from which we can get the parameters for (6) is not available, we will simply assume their values. We take 30 m/sec for the free flow speed $v_f$ and 20 m/sec and 10 sec for $p(v_0)$ and $p(T_0)$, respectively. For the probabilistic density function of gap time, we exploit the results of the empirical study conducted by Kim [8]. For $\gamma$, which is gap time during the disturbance, we use the gamma distribution $\Gamma(1.8, 0.9, 0.4)$, following Kim [8]. For the free-flow traffic at the speed of 30 m/sec, we assume gap time distribution follows a gamma distribution $\Gamma(\alpha, \beta, \lambda)$. The mean gap time of the distribution $\Gamma(\alpha, \beta, \lambda)$ is $\alpha \beta + \lambda$ and we can obtain the mean flow rate by $3600/((\alpha \beta + \lambda + 5/30)$ (vph) assuming a vehicle length of 5 m. We assume the shape parameter $\alpha$ and shift parameter $\lambda$ are constant at 2.4 and 0.4 sec (from the empirical study of gap time) and apply different scale parameter $\beta$ to represent different flow rate. In other words, the gamma distribution corresponding to a flow rate $q$ is $\Gamma(2.4, (3600/q - 5/30 - 0.4)/2.4, 0.4)$. The flow rate out of the disturbance is assumed to be 2000 vph and we get $\Gamma(2.4, 0.51, 0.4)$ for $\gamma$ distribution. The flow rate of the on-ramp varies according to the values of $r_H$. For example, $r_H = 20$ sec corresponds to a ramp flow rate of 180 vph. For a given set of $\beta$ and $r_H$, we conducted 1000 simulation runs and determined whether breakdown occurs or not according to (6) for each simulation. The number of the simulations in which breakdown occurred divided by the total number of simulations (1000) is taken as the probability of breakdown.

Fig. 4 represents the simulation results. At higher ramp flow rate, the PB increases smoothly and at lower ramp flow rate the PB takes a S-shape, in which the PB rapidly increases around the freeway flow 1500 ~ 2500 vph. For freeway flow over 3000 vph, the PB is over 0.8 and does not differ significantly with respect to the ramp flow. On the contrary, for the freeway flow under 2000 vph, the PB is much affected by the ramp flow. For example, the difference of PB between ramp flow rate 240 vph and 100 vph is more than 0.35 at the freeway flow 1000 vph. This finding indicates that the metering of ramp vehicles may not be effective if the freeway flow is high and its effectiveness is nonlinearly related to the metered rate of ramp traffic. To complete

![Fig. 4. Simulation result of probability of breakdown](image-url)
the PB model, we need to develop a PB model that can be applied to a multi-lane freeway. However, the breakdown of a multi-lane freeway is much more complicated due to the inhomogeneous characteristics across lanes and interactions among lane-changing vehicles. The development of Multi-lane PB model is remained for future study.

5 Conclusion

A probabilistic model of breakdown on a freeway is developed based on the wave propagation model. The proposed PB model innovatively explains the breakdown mechanism in an analytical manner with vehicular dynamics. It is shown that the stochastic nature of the breakdown could be soundly explained by the random distribution of gap time and the function of PB is a S-shaped curve, in which the PB rapidly increases around the freeway flow 1500 ~ 2500vph. However, due to the lack of traffic data available, the validation of the proposed model is remained for future study and more general PB model, which can be applied to a multi-lane freeway is yet to be developed to complete the proposed PB model.

References

1. Bullen, A. G. R. Strategies for practical expressway control, Transportation Engineering Journal, Proceedings ASCE, Vol. 98, pp. 599-605, 1972
2. Elefteriadou, L., Roess, R. P., and McShane W. R. Probabilistic nature of breakdown at freeway merge junctions, Transportation Research Record 1484, pp. 80-89, 1995
3. Banks, J. H. Two-capacity phenomenon at freeway bottleneck: A basis for ramp metering ?, Transportation Research Record 1320, pp. 83-90, 1991
4. Kerner, B. S. Experimental features of self-organization in traffic flow, Physical Review Letters Vol.81, No. 17, pp. 3797-3800, 1998
5. Kim, T and Zhang, H. M. Development of a stochastic wave propagation model, Submitted to the Transportation Research Part B, 2004
6. Persaud, B., Yagar, S., and Brownlee R. Exploration of the breakdown phenomenon in freeway traffic, Transportation Research Record 1634, pp. 64-69, 1998
7. Kühne, R. D., and Anstett N. Stochastic methods for analysis of traffic pattern formation, Proceedings of the 14th Intl. Symp. on Transportation and Traffic Flow Theory, Jerusalem, Ceder ed., pp. 177-204, 1999
8. Kim, T Modelling congested traffic on a freeway through the study of time gaps, Ph. D. Dissertation, University of California, Davis, 2003
Novel Symbol Timing Recovery Algorithm for Multi-level Signal

Kwang Ho Chun¹ and Myoung Seob Lim²

¹ IITA, 52, Eoeun-Dong, Yuseong-Gu, DaeJeon, Korea
khchun@iita.re.kr

² Faculty of Electronic & Information Eng., Chonbuk National Univ., Korea
mslim@hslab.chonbuk.ac.kr

Abstract. A new algorithm for detection of symbol timing error for Multi-level PAM signal, where two samples per symbol, \( x_{n-1/2} \) and \( x_{n+1/2} \), around the \( n \) th symbol sample, \( x_n \), are weighted respectively by the gradient between \( x_n \) and \( x_{n+1} \), and between \( x_n \) and \( x_{n-1} \), is proposed. The S-curve to verify the timing error control using this newly proposed algorithm is drawn. Also, the mean and variance of timing error function are derived for the performance analysis and it is shown that the new algorithm outperforms the Gardner’s one.

1 Introduction

In Multi-level PAM systems, it is essential that the demodulator receives accurate information indicating the proper symbol timing instants. From the view of the receiver’s decisions about the transmitted data symbols for producing a timing estimate, symbol timing recovery algorithms can be divided into two groups: Decision Directed (DD) algorithm and Non Decision Directed (NDD) algorithm [1]. The first algorithm is the M&M methods. It considered a one sample per symbol decision directed algorithm, which will be called the DD algorithm. It is expressed as:

\[
e_n(t) = \frac{1}{2} \left[ x_n(t) a_{n+1}(t) - x_{n+1}(t) a_n(t) \right]
\]  

(1)

where \( x_n \) is the \( n \) th symbol sample and \( a_n \) is the estimate value of the \( n \) th transmitted symbol, hence it is the “Decision Directed” nature of the algorithm. Normally, \( a_n \) is estimated by choosing the symbol close to the current \( x_n \). This approach allows a very efficient implementation. This method needs many samples for better performance of Multi-level PAM signals. Considering the fact that there are many variables in the M&M method. It is difficult for the efficient design.

The second algorithm is the NDD(Non Decision Directed) algorithm. The samples can be subdivided into so-called main samples \( x_{n-1} \), \( x_n \) which are finally used also for data decision. And intermediate samples \( x_{n-1/2} \), \( x_{n+1/2} \) which are taken halfway in between the main samples. The subsequent main samples \( x_{n-1} \) and \( x_n \) can be used to
estimate the gradient of the signal curve in between this two samples. The mean deviation of the intermediate sample \( x_{n-1/2} \) is evaluated depending on this gradient between the main samples. The following equation is the timing phase detector function using the product of the signal slope which is simply the difference \( x_n - x_{n-1} \) with the intermediate samples \( x_{n-1/2} \).

\[
e_n(t) = x_{n+1/2}(t) \cdot [x_n(t) - x_{n+1}(t)]
\]

Because this algorithm is independent of carrier phase, it is suitable for both tracking and acquisition modes of operation. Therefore, the timing lock can be achieved without depending upon the prior carrier phase lock. However, this algorithm completely fails for very small roll-off factor, because of its S-curve slope is small, and it can be applied to only two-level signals, such as BPSK/QPSK modulated signal.[2][3]

Therefore, it is very important to develop a new symbol-timing algorithm for multi level signal. The new algorithm compensates the slope between a forward sample and an intermediate sample using the forward sample and a backward sample.

2 The Description of Novel Algorithm

The NDD method timing error detection algorithm is efficient for binary level signal[2]. However, the NDD method causes the jitter to have larger value because the intermediate sample at the transition point between two symbol samples does not have zero value. Hence, this paper will propose the newly efficient and robust timing error detection algorithm for not only QPSK signal but also Multi-level PAM signal. Fig.1 shows the oversampled signals for explanation of the new timing error detection algorithm in case of multi-level signal. Even when the timing is correct in the multi-level signal, the difference between the two samples \( x_{n-1/2} \), \( x_{n+1/2} \) will not be zero.

![Fig. 1. Multi-level Signals](image)

In case of Multi-level signal, the values of the two intermediate samples, \( x_{n-1/2} \) and \( x_{n+1/2} \) around the symbol sample \( x_n \), are different at even correct sample instant(on time). Therefore, the newly proposed algorithm compensates the difference
by using gradients of intermediate values, respectively, \( x_n - x_{n-1} \) and \( x_n - x_{n+1} \). Though, this method is similar to Gardner’s one which uses two samples per symbol, it is different that \( n \)th symbol is made by level decision in \( n \)th point. In case of the various alterative binary signal, without phase error \( \tau \), the intermediate sample value \( x_{n-1/2} \) has zero value. But Multi-level PAM signal is not always zero values because two symbols on front symbol \( x_n \) and back symbol \( x_{n-1} \) have variable levels.

Now, we explain the proposed algorithm. First step in algorithm is that the received \( n-1 \), \( n \), \( n+1 \)th symbols made by signal decision level. Then Second step is obtaining the slope between \( n \)th symbol and \( n-1 \)th symbol, the slope between intermediate sample \( n-1/2 \)th and \( n \)th symbol. Next step compensates the later slope by dividing the former slope. Similarly the slope between intermediate sample \( n+1/2 \)th and \( n+1 \)th symbol is compensated by the slope between \( n \)th symbol and \( n+1 \)th symbol. In the end, the former compensated value is subtracted by the later compensated value. The compensated slope’s difference is applied to the timing error function. If the phase error is zero, the compensated slope's difference is zero. But if the phase error is none zero, the error function \( e_{M\text{-level}}(t) \) of the timing detector is the negative value in fast timing error or the positive value in slow timing error. Illustrated by numerical formula, in symbol period \( T \), the slope between the symbol \( x_n(t) \) and the forward symbol \( x_{n-1}(t) \), and the slope between symbol \( x_n(t) \) and back symbol \( x_{n+1}(t) \) are as follows.

\[
\frac{x_n(t) - x_{n-1}(t)}{T}, \quad \frac{x_{n+1}(t) - x_n(t)}{T}
\] (3)

Intermediate sample slopes obtained by over-sampling of twice baud rate are as follows.

\[
\frac{x_n(t) - x_{n-1/2}(t)}{T/2}, \quad \frac{x_{n+1/2}(t) - x_n(t)}{T/2}
\] (4)

The error function of the timing detector \( e_{M\text{-level}}(t) \) is given the difference equation form of the compensated slope divided by intermediate slope.

\[
e_{M\text{-level}}(t) = \frac{x_n(t) - x_{n-1/2}(t)}{x_n(t) - x_{n-1}(t)} \frac{x_{n+1/2}(t) - x_n(t)}{x_{n+1}(t) - x_n(t)}
\] (5)

Therefore those two samples need to be weighted by the gradient between the neighboring samples. The difference between those two samples after the gradient weighting process is expressed. Figure 2 is the S-curve of timing detector’s error function for showing tracking performance on timing recovery circuit by the equation (5). The S-curve of the proposed method is superior the slope characterize of timing error function when roll-off factor \( \alpha \) is bigger than 0.4.
3 Performance Evaluation

If transmitted signal $a_n$ is independent identically distributed, and noise signal $v_n$ is Additive White Gaussian Noise, the systematic timing phase jitter is determined by the variance of the timing phase detector output at the equilibrium point in the control loop. However, it is difficult to calculate the variance of timing error[4]. Therefore this paper derives equation (6) for the special case when $\tau = 0$, in this case $g_n$ is cancelled.
in \( t=nT \) on Nyquist condition. If the baseband signal of removed carrier is expressed as \( x_n(t)=a_n(t)+v_n(t) \), then the intermediate sample \( x_{n-1/2}(t) \) is as follows:

\[
x_{n-1/2}(t) = \sum_{i} a_i g_{n-1/2-i}(t) + v_{n-1/2}(t)
\]

(6)

where the sequence left \( \{a_n\} \) is taken from the binary sequence where \( a_n=\pm 1 \). We will assume that the \( a_n \) is zero mean and uncorrelated each other. The function \( g_n \) is the shape of the filtered signal pulse. The \( g_n \) of this paper is Raised Cosine Filter.

\[
g(t) = \frac{\sin(\pi t/T) \cos(\alpha \pi t/T)}{\pi t/T (1-(2\alpha \pi t/T)^2)}
\]

(7)

The \( v_n \) is AWGN and \( E\{v(t)v(t+t_1)\} = \sigma^2 g(t_1) \).

These terms are substituted into Gardner’s timing error function. The result for the variance is the equation (8).[3][4] In the equation, \( \sigma^2_e \) is variance of thermal noise \( v_n \)

\[
\sigma^2_v / E^2\{a^2\} = E\{ e_{G,n}(\tau) \cdot e_{G,n-m}(\tau) \} / E^2\{a^2\} = 2 \left\{ \sum_n g^2_{n-1/2} - 2 g^2_{1/2} + g^2_{1/2} \left[ \frac{E\{a^4\}}{E^2\{a^2\}} - 1 \right] \right\} + 2 \frac{\sigma^2_n}{E\{a^2\}} \left( 1 + \sum_n g^2_{n-1/2} \right) + 2 \left[ \frac{\sigma^2_N}{E\{a^2\}} \right]^2
\]

(8)

The \( E\{a^2\} \) and the \( E\{a^4\} \) are the 2nd and 4th moments of each data signal \( a_n \).

The \( g^2_{1/2} \) is the main lobe energy of Raised Cosine Filter, the \( \sum_n g^2_{n-1/2} - 2 g^2_{1/2} \) is the side lobe energy. The first term is due to the self noise and the remaining parts are the thermal noise components of the detector output. The self noise (pattern noise) is related to the data dependent noise and the thermal noise is related to Signal to Noise Ratio(SNR). The BPSK/QPSK signal was little problem for self noise amplitude, but the Multi-level PAM signal is seriously affected by the self noise.[3][4] The equation (5) is the timing error value \( e_{M-level}(t) \) of the proposed method and the equation (9) is the variance \( \sigma^2_{M-level} \) for \( e_{M-level}(t) \) after convergence.

\[
\sigma^2_{M-level} / E^2\{a^2\} = E\left\{ e_{M-level,n}(\tau) \cdot e_{M-level,n-m}(\tau) \right\} / E^2\{a^2\} = \left[ \left\{ E\{a^4\} + 6 E\{a^4\} g^2_{1/2} + E^2\{a^2\} - 14 E^2\{a^2\} \right\} \cdot g^2_{1/2} - 4 E\{a^4\} g^2_{1/2} + 4 E^2\{a^2\} \cdot \sum_n g^2_{n-1/2} \right] + \left\{ 12 E\{a^4\} + 2 E\{a^2\} g^2_{1/2} + 4 E^2\{a^2\} g^2_{1/2} - 32 E\{a^2\} g^2_{1/2} + 4 E\{a^2\} \cdot \sum_n g^2_{n-1/2} \right\} \cdot \sigma^2_N + \left\{ 8 + 6 g^2_{1/2} - 10 g_{1/2} \right\} \cdot \sigma^2_N \right\} / \left[ E\{a^4\} + 3 E^2\{a^2\} + 12 E\{a^2\} \cdot \sigma^2_N + 6 \sigma^2_N \right] \cdot E^2\{a^2\}
\]

(9)
With the similarity of equation (9), the first term is due to the self-noise and the remaining terms are the thermal noise components of the detector output. Compared with the Gardner’s timing function variance, the effect of self-noise about the proposed method is relatively small. Also, the effect of thermal noise is small. Because this algorithm treats the slope at the decision symbols, the effect for adjacent signal is small. Figure 4 shows the performance evaluation about the variance of the Gardner method and the proposed method according to the Eb/No, in condition $\alpha$ in 4, 8-level signal. In case Eb/No=0, the proposed method has the uniform performance according to the bandwidth($\alpha$). This means that although the effect of thermal noise is large, the effect of pattern jitter is little. But in case Eb/No=9, when the $\alpha$ is in-
creasing, the performance is increasing. This reason is that the performance affects the pattern noise when the thermal noise's is small. Totally compared, Eb/No=0, in 4 level, the proposed method is superior to Gardner method by 22dB and about 20dB in Eb/No=9.

Figure 5 is the performance evaluation when Eb/No is 0, 3, 9 and according to $\alpha$ of Raised Cosine Filter in 8 level-PAM signal.

Figure 6 is the performance evaluation when the $\alpha$ of Raised Cosine Filter is 0.1, 0.9, and according to Eb/No in 16 level-PAM signal. If the thermal noise is small, the proposed method is superior to Gardner method.

![Graph showing variance of timing error according to Eb/No](image)

**Fig. 6.** Variance of timing error according to Eb/No

4 Conclusions

In this paper we proposed the new method that the slope compensation method between forward sample and intermediate sample using forward sample and backward sample for the Multi-level signal. To verify the performance of proposed algorithm, we obtained the variance of timing function and its numerical analysis and simulation. The proposed algorithm is superior to any other algorithms when roll-off fact is low state. The performance of the thermal noise is mighty than Gardner's algorithm. We propose the new algorithm adequate for such applications as satellite and any other modem synchronizer technique because of the better jitter and noise performance.

References

[1] Gardner F. M, “A BPSK/QPSK timing error detector for sampled data receivers”. IEEE Trans, Commun. May 1986.
[2] Cowley W. G and Sabel L. P, "The Performance of two symbol timing recovery algorithm for PSK demodulators". IEEE Commun. June 1994.
[3] Kim Jeong Kwon and Lee Yong Hwang, "Timing Recovery Based on Zero-crossing Detection for multi-level PAM Signals". The journal of the Korean institute of communication sciences, VOL.22/No.10 pp. 2246 - 2255.

[4] Lankl B and Sebald G. "Jitter-reduced digital timing recovery for multi-level PAM and QAM systems" ICC '93 Geneva. Technical Program, Conference Record, IEEE International Conference on, Volume: 2, 1993.
Development Site Security Process of ISO/IEC TR 15504

Eun-ser Lee¹ and Tai-hoon Kim²

¹Chung-Ang University, 221, Huksuk-Dong, Dongjak-Gu, Seoul, Korea
eslee@object.cau.ac.kr
http://object.cau.ac.kr/selab/index.html
²KISA, 78, Garak-Dong, Songpa-Gu, Seoul, Korea
taihoon@kisa.or.kr

Abstract. The IT products like as firewall, IDS (Intrusion Detection System) and VPN (Virtual Private Network) are made to perform special functions related to security, so the developers of these products or systems should consider many kinds of things related to security not only design itself but also development environment to protect integrity of products. When we are making these kinds of software products, ISO/IEC TR 15504 may provide a framework for the assessment of software processes, and this framework can be used by organizations involved in planning, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. But, in the ISO/IEC TR 15504, considerations for security are relatively poor to other security-related criteria such as ISO/IEC 21827 or ISO/IEC 15408 [10-12]. In fact, security related to software development is concerned with many kinds of measures that may be applied to the development environment or developer to protect the confidentiality and integrity of the IT product or system developed. In this paper we propose some measures related to development process security by analyzing the ISO/IEC 21827, the Systems Security Engineering Capability Maturity Model (SSE-CMM) and ISO/IEC 15408, Common Criteria (CC). And we present a Process of Security for ISO/IEC TR 15504.

1 Introduction

ISO/IEC TR 15504, the Software Process Improvement Capability Determination (SPICE), provides a framework for the assessment of software processes [1-9]. This framework can be used by organizations involved in planning, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. But, in the ISO/IEC TR 15504, considerations for security are relatively poor to others. For example, the considerations for security related to software development and developer are lacked.

When we are making some kinds of software products, ISO/IEC TR 15504 may provide a framework for the assessment of software processes, and this framework can be used by organizations involved in planning, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. But, in the ISO/IEC TR 15504, considerations for security are relatively
poor to other security-related criteria such as ISO/IEC 21827 or ISO/IEC 15408 [10]. In fact, security related to software development is concerned with many kinds of measures that may be applied to the development environment or developer to protect the confidentiality and integrity of the IT product or system developed.

In this paper, we propose a process related to security by comparing ISO/IEC TR 15504 to ISO/IEC 21827 and ISO/IEC 15408. The proposed scheme may be contributed to the improvement of security for IT product or system. And in this paper, we propose some measures related to development process security by analyzing the ISO/IEC 21827, the Systems Security Engineering Capability Maturity Model (SSE-CMM) and ISO/IEC 15408, Common Criteria (CC). And we present a Process for Security for ISO/IEC TR 15504.

2 ISO/IEC TR 15504

2.1 Framework of ISO/IEC TR 15504

ISO/IEC 15504 provides a framework for the assessment of software processes. This framework can be used by organizations involved in planning, managing, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. ISO/IEC 15504 provides a structured approach for the assessment of software processes for the following purposes:

– by or on behalf of an organization with the objective of understanding the state of its own processes for process improvement;
– by or on behalf of an organization with the objective of determining the suitability of its own processes for a particular requirement or class of requirements;
– by or on behalf of one organization with the objective of determining the suitability of another organization’s processes for a particular contract or class of contracts.

The framework for process assessment:

– encourages self-assessment;
– takes into account the context in which the assessed processes operate;
– produces a set of process ratings (a process profile) rather than a pass/fail result;
– through the generic practices, addresses the adequacy of the management of the assessed processes;
– is appropriate across all application domains and sizes of organization.

The process assessment framework is based on assessing a specific process instance. A process instance is a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a manner that provides repeatable ratings. Each process instance is characterized by a set of five process capability level ratings, each of which is an aggregation of the practice adequacy ratings that belong to that level. Hence the practice adequacy ratings are the foundation for the rating system.
2.2 Process Dimension of ISO/IEC TR 15504

ISO/IEC TR 15504-5 defines the process dimension of the assess model. The process dimension is directly mapped to that of the reference model in ISO/IEC TR 15504-2, and adopts the same process definitions and structure given by the reference model. The three life cycles process groupings are:

- The primary life cycle processes consisting of the process categories Engineering and Customer-Supplier.
- The Supporting life cycle processes consisting of the process category Support.
- The Organizational life cycle processes consisting of the process categories Management and Organization.

The process dimension contains five process categories which are:

- CUS, Customer-Supplier
- MAN, Management
- ENG, Engineering
- ORG, Organization
- SUP, Support

The description of each process category includes a characterization of the processes it contains, followed by a list of the process names.

2.3 Assessment Model and Indicators of Process Performance

ISO/IEC TR 15504-5 describes the assessment model, and the assessment model expands the reference model depicted in the ISO/IEC TR 15504-2 by adding the definition and the use of assessment indicators which are defined to support an assessor’s judgment of the performance and capability of an implemented process.

Base practice, input and output work products and their associated characteristics relate to the processes defined in the process dimension of the reference model, and are chosen to explicitly address the achievement of the defined process purpose. The base practices and work products are indicators of a level 1 process performance. The presence of the work products with the existence of the characteristics of the work products, and evidence of performance of the base practices, provide objective evidence of the achievement of the purpose of the process.

3 A New Process for Development Site Security

3.1 Work Products of ISO/IEC TR 15504 Related to Development Security

As mentioned earlier, ISO/IEC TR 15504 provides a framework for the assessment of software processes, and this framework can be used by organizations involved in planning, managing, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. ISO/IEC TR 15504 does
not define any Process related to security, but the security-related parts are expressed in some Work Products (WP) as like.

| ID | WP Class | WP Type                          | WP Characteristics                                                                 |
|----|----------|----------------------------------|-------------------------------------------------------------------------------------|
| 10 | 1.3      | Coding standard                  | - Security considerations                                                           |
| 51 | 3.2      | Contract                         | - References to any special customer needs (i.e., confidentiality requirements, security, hardware, etc.) |
| 52 | 2.2      | Requirement specification        | - Identify any security considerations/constraints                                    |
| 53 | 2.3      | System design/architecture        | - Security/data protection characteristics                                           |
| 54 | 2.3      | High level software design       | - Any required security characteristics required                                     |
| 74 | 1.4/2.1  | Installation strategy plan       | - Identification of any safety and security requirements                             |
| 80 | 2.5      | Handling and storage guide       | - Addressing appropriate critical safety and security issues                         |
| 101| 2.3      | Database design                  | - Security considerations                                                            |
| 104| 2.5      | Development environment          | - Security considerations                                                            |

ISO/IEC TR 15504 may use these work products as input materials, and these may be the evidence that security-related considerations are being considered. But this implicit method is not good for the ‘security’ and more complete or concrete countermeasures are needed. Therefore, we propose some new processes which deal with the security.

### 3.2 A New Process for Development Site Security

For example, we want to deal the security for the site where the software is developed. In the ISO/IEC TR 15504-5, there is the Engineering process category (ENG) which consists of processes that directly specify, implement or maintain the software product, its relation to the system and its customer documentation. In circumstances where the system is composed totally of software, the Engineering processes deal only with the construction and maintenance of such software.

The processes belonging to the Engineering process category are ENG.1 (Development process), ENG.1.1 (System requirements analysis and design process), ENG.1.2 (Software requirements analysis process), ENG.1.3 (Software design process), ENG.1.4 (Software construction process), ENG.1.5 (Software integration process), ENG.1.6 (Software testing process), ENG.1.7 (System integration and testing process), and ENG.2 (Development process).

These processes commonly contain the 52nd work product (Requirement specification), and some of them have 51st, 53rd, 54th work products separately. Therefore, each process included in the ENG category may contain the condition, ‘Identify any security considerations/constraints’. But the phrase ‘Identify any security considera-
tions/constraints’ may apply to the ‘software or hardware (may contain firmware) development process’ and not to the ‘development site’ itself.

In this paper we will present a new process applicable to the software development site. In fact, the process we propose can be included in the MAN or ORG categories, but this is not the major fact in this paper, and that will be a future work. We can find the requirements for Development security in the ISO/IEC 15408 as like;

**Development security covers the physical, procedural, personnel, and other security measures used in the development environment. It includes physical security of the development location(s) and controls on the selection and hiring of development staff.**

Development security is concerned with physical, procedural, personnel, and other security measures that may be used in the development environment to protect the integrity of products. It is important that this requirement deals with measures to remove and reduce threats existing in the developing site (not in the operation site). These contents in the phrase above are not the perfect, but will suggest a guide for development site security at least.

The individual processes of ISO/IEC TR 15504 are described in terms of six components such as Process Identifier, Process Name, Process Type, Process Purpose, Process Outcomes and Process Notes. The style guide in annex C of ISO/IEC TR 15504-2 provides guidelines which may be used when extending process definitions or defining new processes. Next is the Development Security process we suggest.

(1) Process Identifier: ENG.3
(2) Process Name: Development Security process
(3) Process Type: New
(4) Process purpose:

The purpose of the Development Security process is to protect the confidentiality and integrity of the system components (such as hardware, software, firmware, manual, operations and network, etc) design and implementation in its development environment. As a result of successful implementation of the process:

(5) Process Outcomes:
- access control strategy will be developed and released to manage records for entrance and exit to site, logon and logout of system component according to the released strategy
- roles, responsibilities, and accountabilities related to security are defined and released
- training and education programs related to security are defined and followed
- security review strategy will be developed and documented to manage each change steps
(6) Base Practices:
ENG.3.BP.1: Develop physical measures. Develop and release the physical measures for protecting the access to the development site and product.
ENG.3.BP.2: Develop personnel measures. Develop and release the personnel measures for selecting and training of staffs.
ENG.3.BP.3: Develop procedural measures. Develop the strategy for processing the change of requirements considering security.
Development Site Security Process of ISO/IEC TR 15504

ENG.3 Development Security process may have more base practices (BP), but we think these BPs will be the base for future work. For the new process, some work products must be defined as soon as quickly. Next items are the base for the definition of work products.

| WP category number | WP category | WP classification number | WP classification | WP type |
|--------------------|-------------|--------------------------|-------------------|---------|
| 1                  | ORGANIZATION| 1.1                      | Policy            | Access control to site and so on |
|                    |             | 1.2                      | Procedure         | Entrance and so on |
|                    |             | 1.3                      | Standard          | Coding and so on |
|                    |             | 1.4                      | Strategy          | Site open and so on |
| 2                  | PROJECT     | Future work              | Future work       | Future work |
| 3                  | RECORDS     | 3.1                      | Report            | Site log and so on |
|                    |             | 3.2                      | Record            | Entrance record and so on |
|                    |             | 3.3                      | Measure           | Future work |

4 Conclusions

In this paper we proposed a new Process applicable to the software development site. Some researches for expression of Base Practice and development of Work Products should be continued. But the work in the paper may be the base of the consideration for security in ISO/IEC TR 15504. ISO/IEC TR 15504 provides a framework for the assessment of software processes, and this framework can be used by organizations involved in planning, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. Therefore, it is important to include considerations for security in the Process dimension.

References

1. ISO. ISO/IEC TR 15504-1:1998 Information technology – Software process assessment – Part 1: Concepts and introductory guide
2. ISO. ISO/IEC TR 15504-2:1998 Information technology – Software process assessment – Part 2: A reference model for processes and process capability
3. ISO. ISO/IEC TR 15504-3:1998 Information technology – Software process assessment – Part 3: Performing an assessment
4. ISO. ISO/IEC TR 15504-4:1998 Information technology – Software process assessment – Part 4: Guide to performing assessments
5. ISO. ISO/IEC TR 15504-5:1998 Information technology – Software process assessment – Part 5: An assessment model and indicator guidance
6. ISO. ISO/IEC TR 15504-6:1998 Information technology – Software process assessment – Part 6: Guide to competency of assessors
7. ISO. ISO/IEC TR 15504-7:1998 Information technology – Software process assessment – Part 7: Guide for use in process improvement
8. ISO. ISO/IEC TR 15504-8:1998 Information technology – Software process assessment – Part 8: Guide for use in determining supplier process capability
9. ISO. ISO/IEC TR 15504-9:1998 Information technology – Software process assessment – Part 9: Vocabulary
10. ISO. ISO/IEC 15408-1:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model
11. ISO. ISO/IEC 15408-2:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 2: Security functional requirements
Improving CAM-DH Protocol for Mobile Nodes with Constraint Computational Power

Yong-Hwan Lee¹, Il-Sun You², and Sang-Surm Rhee¹

¹Dept. of Electronics and Computer Engineering, Dankook Univ., Korea
hwany1458@empal.com
²Dept. of Information and Computer Science, Dankook Univ., Korea

Abstract. CAM-DH is a publickey-based protocol for secure binding updates in Mobile IP V.6 (MIPv6), which combines the BAKE/2 protocol with a digitally signed Diffie-Hellman key exchange. In MIPv6 environment, an important design consideration for publickey-based binding update protocols is to minimize asymmetric cryptographic operations in Mobile Nodes (MNs) with constraint computational power. In this paper we propose a novel approach that can resolve the current security problems in CAM-DH, by adopting Aura’s Cryptographically Generated Address (CGA) scheme with two hashes, which search for hash collisions in the CGA method in order to prevent brute-force attacks. By comparing with CAM-DH, our approach shows promise to minimize the computational overhead of MNs, as well as provide better management and stronger security than CAM-DH does.

1 Introduction

A MN in MIPv6 environment belongs to a home link and is always addressable by its Home of Address (HoA), regardless of its current point of attachment to the Internet [1, 2, 7]. While attached to some foreign link away from its home, each MN is addressable by one or more Care-of Addresses (CoA). The basic idea is to allow a Home Agent (HA) to work as a stationary proxy for the MN. Whenever the MN is away from home, the HA intercepts packets destined to the HoA of the MN, encapsulates them, and tunnels them to the registered CoA of the MN. When the MN wants to send packets to a Correspondent Node (CN), it sends them to the HA over the reverse tunnel. The HA un-encapsulates the packets and forwards them to the CN. Thus, MIPv6 enables MNs to have both mobility and reachability. However, it results in longer paths and degraded performance. In order to mitigate the performance problem, MIPv6 includes route optimization that allows the MN and its CN to directly exchange packets, excluding the initial setup phase. A binding is the association between a MN’s HoA and CoA. The MN initializes the route optimization by sending Binding Update (BU) messages including its current binding to the CN. Upon receiving the BU message, the CN learns and caches the MN’s current binding. After that, the CN can directly send packets to the MN using the MN’s CoA. The essential re-
quirement to address the security threats is for the CN to authenticate the MN by sending the BU message. Only after successfully authenticating the MN, the CN has to update its binding cache entries. Unfortunately, it is so difficult to achieve strong authentication between two previously unknown nodes (MN and CN) in which no global security infrastructure is available.

In this paper we propose a novel approach that improves the optimization of CAM-DH in order that the HA can prevent the denial of service attacks and off-load the expensive cryptographic operations of its MNs to itself. Furthermore, we adopt Aura’s CGA-based scheme with two hashes in order to prevent brute-force attacks by searching for hash collisions in the CGA method [5].

2 Related Work

Recently, the Return Rout-ability (RR) protocol has been accepted as a basic technique for securing BUs. Nevertheless, the RR protocol has some critical drawbacks, in aspect of its security properties and performance [2]. The protocols such as CAM, CAM-DH, SUCV (Statistics Uniqueness and Cryptographic Verifiability) and ABKs (Address-based Keys) have been based on public key [2-6]. These protocols attempted to associate the MN’s address with its public key to avoid additional Public Key Infrastructure (PKI), by using CGA and identity-based cryptosystems [2]. For performance, it is desirable to minimize the expensive cryptographic operations in MNs with constraint computational power. CAM-DH, SUCV and Deng-Zhou-Bao’s protocol provide an optimization to off-load the expensive cryptographic operation of the MN to its HA [2, 6, 7]. But, in CAM-DH, cryptographic operations of the MN are expensive and can not be handed over to its HA [4, 7]. In SUCV, managing the MN’s private key causes unwanted additional cost to the HA [6]. Deng-Zhou-Bao’s protocol needs, an additional security infrastructure, for handling Public Key Certificates (PKC) [2].

3 CAM-DH Protocol

CAM-DH is reviewed and its weaknesses are analyzed. Notation is as follows.

\( h() : \) a cryptographic secure one-way hash function

\( prf(k, m) : \) a keyed hash function. It accepts a secret key \( k \) and a message \( m \), and generates a pseudo random output.

\( \mathbb{P}_X/S_X : \) a public and private key pair of \( X \).

\( S_X(m) : \) node \( X \)’s digital signature on a message \( m \).

\( m|n : \) concatenation of two messages \( m \) and \( n \).

\( \text{CN} : \) CN represents both the correspondent node and its IP address.

Let assume that \( p \) and \( g \) are public Diffie-Hellman parameters, where \( p \) is a large prime and \( g \) is a generator of the multiplicative group \( \mathbb{Z}_p^* \). For simoke notation, \( g^r \mod p \) can be expressed as \( g^r \). It is assumed that the values of \( p \) and \( g \) are agreed upon before hand by all the parties concerned. Fig. 1 outlines CAM-DH with an optimiza-
tion for MNs with constraint computational power, such as PDAs and cellular phones. For the optimization, the HA intercepts the second message camdhpc2 and performs certain processing on it before forwarding it to the MN. Because communication between the MN and the HA is protected with pre-established security association in the MIPv6, such optimization is available [2,7].

Each CN generates a nonce, \( N_j \), at regular intervals, for example every few minutes. A CN uses the same \( K_{CN} \) and \( N_j \) with all MNs it is in communication with, so that it does not need to generate and store a new \( N_j \) when a new MN contacts it. Each value of \( N_j \) is identified by the subscript \( j \). Thus, the CN is not flooded with nonces. CGA is IPv6 address where the interface identifier is generated by hashing the address owner’s public key [3, 5]. The address owner can use the corresponding private key to assert address ownership and to sign messages sent from the address without any additional security infrastructure. In this protocol, each MN’s HoA is generated from its public key \( P_{MN} \) and used as a CGA. After this protocol, the subsequent BU messages from the MN are authenticated through the session key \( K_h \) established between the MN and the CN.

In spite of high-level security, CAM-DH has the following drawbacks. First, the optimization for low-power MNs results in the HA’s vulnerability to denial of service attacks, since the HA uses Diffie-Hellman key agreement to calculate a session key \( K_h \) without authenticating the CN. Thus, the HA is easy to be flooded with a storm of camdp2 messages. Second, the protocol does not unload all asymmetric cryptographic operations from the MN, since the HA just performs expensive cryptographic operations for a session key \( K_h \) instead of the MN. Therefore, the MN should compute \( SIG_{MN} \) with its private key \( S_{MN} \). Third, CAM-DH, a CGA-based protocol, is vulnerable to brute-force attacks searching for hash collisions, because of using only the 62 bits of the interface identifier as the hash value for the address owner’s public key.
4 Our Proposed Protocol

4.1 Two Hash Based CGA Scheme

The first hash value (Hash1) is used to produce the interface identifier (i.e., rightmost 64 bits) of the address. The purpose of the second hash (Hash2) is to artificially increase that computational complexity of generating new addresses and, consequently, the cost of brute-force attacks. In our protocol, a home link is associated with a public/private key pair $P_{HA}$ and $S_{HA}$ in a digital signature scheme. A HA in the home link keeps the public/private key pair, and derives a CGA from the public key $P_{HA}$. Each CGA is associated with an optimized parameter format including the HA’s public key information and CGA parameters [5]. The process of obtaining a new CGA is as follows.

1. Generate a public/private key pair $P_{HA}$ and $S_{HA}$ for a home link.
2. Generate a new CGA via the algorithm presented in [5].
3. Create an optimized parameter format. The format is simply the concatenation of the DER-encoded subjectPublicKeyInfo and CGAParameters data value.

subjectPublicKeyInfo and the format of CGAParameters are defined in [5].

4.2 Protocol Operation

In our protocol in the drawbacks of CAM-DH, first, to off-load the asymmetric cryptographic operations of the MN to the HA, our protocol allows the HA to perform the expensive operations on behalf of the MN. For that, the HA keeps the public/private key pair $P_{HA}/S_{HA}$ and uses CGA, derived from its public key $P_{HA}$, as its own MIPv6 address. The CN should validate the public key $P_{HA}$ with the HA’s CGA before verifying the signature $SIG_{HA}$. Such a mechanism enables our protocol to be more manageable and scalable than other public key based protocols where the MN binds its public key with its own address, in addition to unloading all asymmetric cryptographic operations from the MN. Second, to prevent denial of service attacks on the HA, a cookie $C_0$ is created and added to the first message $\{HoA, CoA\}$ sent by the MN. Only if the cookie is valid, the HA performs asymmetric cryptographic operations. For a cookie $C_0$, the HA, like the CN, generates a nonce, $N_c$, at regular intervals, preventing itself from not being flooded with nonces. Third, to overcome the limited length of the hash used in the CGA, our protocol uses Aura’s two hash based CGA, which enhances CAM-DH’s security by increasing the cost of brute-force attacks by a factor $2^{12*Sec}$.

Fig. 2 show our protocol where the HA functions as a security proxy for the MN, testifies the legitimacy of the MN’s HoA, facilitates authentication of the MN to the CN and establishes a session key for them. By sending ecamdp1, the MN tries to contact the CN. Upon receipt of ecamdp7, the CN firstly checks $MAC_j$ with $K_r$. It should attempt to verify $MAC_j$ only if $MAC_2$ is valid. If $MAC_2$ is valid, the CN creates a cache entry for the MN’s HoA and the key $K_h$, which will be used for authenticating subsequent BU messages from the MN. Especially, before computing $K_{BU}=prf(K_h, r_j)$,
the CN should verify the HA’s CGA and $SIG_{HA}$. The algorithm for verifying the HA’s CGA is defined in [5]. When the verification is positive, the CN can be confident that the MN’s $HoA$ is valid and the Diffie-Hellman public value $g^x$ is freshly generated by the HA.

5 Analysis of the Proposed Protocol

5.1 Security

5.1.1 Denial of Service Attacks
Since the MN-HA path is protected with pre-establish security association, we focus on denial of service attacks in the HA-CN path. By sending a storm of ecamdp2 messages, an intruder can try to attack the CN. Since our protocol uses the same $g^x$ as the CN’s Diffie-Hellman public value instead of generating a new one, it is not vulnerable to such an attack. To prevent a storm of ecamdp3 messages, the HA uses a cookie $C_0$. Also, the CN tests RR of the MN’s new care of address $CoA$ to protect itself against a storm of ecamdp7 messages.

5.1.2 The Cost of Brute-Force Attacks
During the address generation phase of our protocol, the input for the additional hash Hash2 is modified by varying the value of modifier until the leftmost $12 \times \text{Sec}$ bits of Hash2 are zero. This increases the cost of address generation approximately by a factor of $2^{12 \times \text{Sec}}$. It also increases the cost of brute-force attacks by the same factor (from $2^{59}$ to $2^{59+12 \times \text{Sec}}$). Therefore, our protocol is more secure than other CGA based approaches such as CAM-DH and SUCV, which require the cost of brute-force attacks, $O(2^{62})$. 

![Fig. 2. Our Secure Binding Update Protocol](image-url)
5.2 Performance and Manageability

We evaluate the performance of our protocol in terms of the cryptographic operations that each MN should perform. The costs can be expressed as follows.

| Proposed Protocol | CAM-DH |
|-------------------|--------|
| 1. two hash-based CGA | one hash-based CGA |
| 2. HA | MN |
| 3. $O(2^{59+2c_{sec}})$ | $O(2^{c})$ |
| 4. 0 | 1 |
| 5. High | Low |
| 6. O | X |
| 7. $C_{hash} + 2* C_{hmac}$ | $C_{sign} + C_{hash} + 3* C_{hmac}$ |

Table 1. The Comparison of the Protocols

- 1. Mechanism binding the public key with its owner
- 2. Node who generates and manages the private key/public key pair
- 3. Cost of brute force attacks
- 4. Asymmetric cryptographic operations the MN should perform
- 5. Manageability and Scalability
- 6. Ability to Prevent Denial of Service Attacks
- 7. Cost of the cryptographic operations that a MN should perform

\[
C_{Our-P-MN} = \text{the cost for computing } K_3 + \text{the cost for computing } K_{BU} + \text{the cost for computing } MAC_1 \\
= C_{hash} + 2* C_{hmac}
\]

\[
C_{CAMDH-MN} = \text{the cost for computing } K_3 + \text{the cost for computing } K_{BU} + \text{the cost for computing } MAC_1 + \text{the cost for computing } SIG_{MN} + \text{the cost for computing } MAC_2 \\
= C_{sign} + C_{hash} + 3* C_{hmac}
\]

$C_{Our-P-MN}$: the cost of cryptographic operations that a MN should perform in our protocol

$C_{CAMDH-MN}$: the cost of cryptographic operations that a MN should perform in CAM-DH protocol

$C_{sign}$: the cost for one signing operation

$C_{hash}$: the cost for one hash operation

$C_{hmac}$: the cost for one HMAC operation

In comparison to CAM-DH, our protocol needs an additional cost, $2* C_{hmac} + C_{hash}$, for a cookie $C_0$ and two hash-based CGA. But, as shown above the MN in our protocol just needs $C_{hash} + 2* C_{hmac}$ without any cost for asymmetric cryptographic operations, whereas the one in CAM-DH needs $C_{sign} + C_{hash} + 3* C_{hmac}$. Thus, it can minimize cryptographic operations that each MN should perform, while satisfying an important design consideration for public key based BU protocols [2]. Also, since the HA, instead of the MN, keeps the public/private key pair, and derives a CGA from the public key $P_{HA}$, our protocol is more manageable and scalable than CAM-DH where the MN binds its public key with its own address. The comparison of our protocol with CAM-DH is summarized in Table 1.
6 Conclusions

In this paper we propose a protocol for securing BUs in MIPv6, which improves the weaknesses of CAM-DH. Our protocol advances the optimization for MNs with constraint computational power in order that the HA can prevent the denial of service attacks and off-load the expensive cryptographic operations of its MNs to itself. Additionally, since our protocol allows the HA to generate the public/private key pair and derive a CGA from the generated public key, it is more manageable and scalable than CAM-DH where the MN binds its public key with its own address. Furthermore, our protocol uses Aura’s two hash-based CGA scheme to overcome the limitation of the CGA method. Because the two hash-based CGA scheme increases the cost of brute-force attacks by a factor of $2^{12\text{Sec}}$ (from $2^{59}$ to $2^{59+12\text{Sec}}$), our protocol can achieve stronger security than other CGA-based protocols.

References

1. Arkko J, “Security Framework for Mobile IPv6 Route Optimization,” <draft-arkko-mipv6ro-secframework-00.txt>, Nov. 2001.
2. Deng R, Zhou J, and Bao F, “Defending Against Redirect attacks in Mobile IP,” CCS’02, Nov. 2002.
3. O’Shea G and Roe M, “Child-proof authentication for MIPv6 (CAM),” ACM Computer Communications Review, April 2001.
4. Roe M, Aura T, O’Shea G, and Arkko J, “Authentication of Mobile IPv6 Binding Updates and Acknowledgments,” <draft-roe-mobileip-updateauth-02.txt>, Feb. 2002.
5. Aura T, “Cryptographically Generated Addresses (CGA),” <draft-aura-cga-00.txt>, Feb. 2003.
6. Montenegro G, Castelluccia C, “SUCV Identifiers and Addresses,” <draft-montenegro-sucv-02.txt>, Nov. 2001.
7. Johnson D, Perkins C and Arkko J, “Mobility Support in IPv6,” <draft-ietf-mobileip-ipv6-24.txt>, Jun. 2003.
Space Time Code Representation in Transform Domain

Gi Yean Hwang, Jia Hou, and Moon Ho Lee

Institute of Information & Communication, Chonbuk National University, Chonju, 561-756
infoman@mail.chonbuk.ac.kr, houjiastock@hotmail.com,
moonho@chonbuk.ac.kr

Abstract. In this paper, the space time block code is investigated in transform domain. We represent the orthogonal space time block codes by using a fast Hadamard transform. Additionally, the filter theory is considered to realize this transform.

1 Introduction

Space Time code has potential capacity in Multiple Input Multiple Output (MIMO) channel, especially, the recent research shows it could obviously improve the performance in fading channel. Since the multiple antennas system including a potential capacity, a wireless communication system with n transmitters and m receivers is considered. In this paper, we develop the space time block code in the transform domain, since exactly it is a matrix representation by using the orthogonal theory [1, 2, 3, 4]. In the proposal, the space time block codes based on orthogonal matrix can be decomposed to the operations, filter bank and puncture in the transform domain. By considering the I, Q channels information is directly calculated from the transform values, we separate the complex member to $x_I + y_Qj$, and then we construct them to a new transform to obtain the space time processing, therefore, the filtering is easily to be designed. As a consequence, the space time block code can be efficiently represented in transform domain and filtering, also, from the transform domain construction consideration, we can improve the space time processing and exploit the redundant matrix for symbol interleaved iterative decoding algorithm.

2 Transform for Orthogonal Space Time Block Codes

The space time block code efficiently exploits the orthogonal properties and diversity gain in fading channel to achieve the full rank. The main transmission model is based on matrix theory, therefore, it is possibly represented in transform domain. To Simplify the space time block code encoding algorithm [1, 2, 3], we firstly denote the transmitted symbols as $s_1, s_2$, and their conjugate are $s_1^*, s_2^*$. By considering the I, Q channel calculation in the transform domain, it represent the transmitted symbols as

\[ s_1 = x_1 + y_1j, \quad s_2 = x_2 + y_2j \]
where $x$ and $y$ denote the real and imaginary part of the transmitted symbols. Then the transform from I, Q channel to complex symbols can be denoted as

$$
\begin{bmatrix}
  s_1 & s_1^* \\
  s_2 & s_2^*
\end{bmatrix} =
\begin{bmatrix}
  x_1 + y_1 j & x_1 - y_1 j \\
  x_2 + y_2 j & x_2 - y_2 j
\end{bmatrix} =
\begin{bmatrix}
  x_1 & y_1 \\
  x_2 & y_2
\end{bmatrix}
\begin{bmatrix}
  1 & 1 \\
  j & -j
\end{bmatrix}
$$

(2)

where the (2) is drawn as a butterfly pattern, as shown in Fig. 1.

Fig. 1. Butterfly algorithm for I, Q channel calculation

To simplify the transform operations and improve the performance, we consider constructing a new structure similar to the fast Hadamard transform in this section. Firstly, let us define the two parallel I and Q channels information input as $d = [x_1 \ y_1 j \ x_2 \ y_2 j]$, then we have

$$
\begin{bmatrix}
  x_1 & y_1 \\
  x_2 & y_2
\end{bmatrix}
\begin{bmatrix}
  1 & 1 & 0 & 0 \\
  0 & 0 & 1 & 1 \\
  0 & 1 & 0 & 0 \\
  1 & 0 & 0 & 1
\end{bmatrix} =
\begin{bmatrix}
  x_1 + y_1 j & x_1 - y_1 j & x_1 + y_1 j & x_1 - y_1 j
\end{bmatrix}
$$

(3)

where $j = \sqrt{-1}$. Assume that a transform matrix is $T$, we can obtain the orthogonal space time block codes as

$$
\begin{bmatrix}
  s_1 & s_1^* & s_2 & s_2^*
\end{bmatrix} T =
\begin{bmatrix}
  s_1 & -s_2^* \\
  s_2 & s_1^*
\end{bmatrix}
$$

(4)

where $T$ is the transform matrix that we want to find. Let us define

$$
\begin{bmatrix}
  x_1 & y_1 j & x_2 & y_2 j
\end{bmatrix}
\begin{bmatrix}
  1 & 0 & 0 & r \\
  0 & r & D^{-1} & 0 \\
  0 & r & 1 & 0 \\
  -D^{-1} & 0 & 0 & r
\end{bmatrix}
$$

(5)

where $r$ presents random values. After a puncture function, we have

$$
\begin{bmatrix}
  s_1 - D^{-1}s_2^* \\
  D^{-1}s_1^* + s_2 \\
  R
\end{bmatrix} [Pu] =
\begin{bmatrix}
  s_1 - s_2^* \\
  s_2 \\
  s_1^*
\end{bmatrix}
$$

(6)

where the $R$ denotes any values from the transform matrix. And the transform matrix before the puncture can be decomposed as
\[
\begin{bmatrix}
1 & 0 & 0 & r \\
0 & r & D^{-1} & 0 \\
0 & r & 1 & 0 \\
-D^{-1} & 0 & 0 & r
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 0 & D^{-1} \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} +
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & r & 0 & 0 \\
0 & r & 0 & 0 \\
0 & 0 & 0 & r
\end{bmatrix}
\]
\tag{7}
\]

To easily take out the \(D^{-1}\), we can design a systematic structure matrix to substitute the random matrix by using
\[
\begin{bmatrix}
0 & 0 & 0 & r \\
0 & r & 0 & 0 \\
0 & 0 & 0 & r
\end{bmatrix}
= \begin{bmatrix}
0 & 0 & 0 & D^{-1} \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
\tag{8}
\]
thus we can have
\[
\begin{bmatrix}
s_1 & s_1^* & s_2 & s_2^*
\end{bmatrix}
= \begin{bmatrix}
0 & 0 & 0 & D^{-1} \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & s_1 & -D^{-1}s_2 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
s_1^* & s_1^* & s_2^* + D^{-1}s_1 & s_2
\end{bmatrix}
\]
\tag{9}
\]

Obviously, it is one orthogonal space time block code, but it is a conjugate format of original case. Therefore, we can develop one outer and inner space time block code which uses a quasi-orthogonal pattern \([5,6]\). Thus we obtain the complex symbols \(S_1, S_2\) as
\[
S_1 = s_1 - D^{-1}s_2^*, \quad S_2 = s_2 + D^{-1}s_1^*,
\]
\tag{10}
\]

then the conjugate of them can be denoted as
\[
S_1^* = s_1^* - D^{-1}s_2, \quad S_2^* = s_2^* + D^{-1}s_1,
\]
\tag{11}
\]

these symbols are all from the extending transmission matrix and original case is orthogonal, they have
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 0 & D^{-1} \\
0 & 0 & 1 & 0 \\
-D^{-1} & 0 & 0 & 0
\end{bmatrix}
= (-1) \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
\tag{12}
\]

The combination of (18-19) can be separately decoded, since they are same orthogonal structure. Thus the quasi-orthogonal transform space time block code can be shown as
\[
\begin{bmatrix}
s_1 & -D^{-1}s_2^* & 0 & 0 \\
s_2 & D^{-1}s_1^* & 0 & 0 \\
0 & s_3 & -D^{-1}s_4^* & 0 \\
0 & s_4 & D^{-1}s_3^* & 0
\end{bmatrix}
\tag{13}
\]
where the $s_3 = x_1 - y_1 j$, $s_4 = x_2 - y_2 j$, easily, we find these are the conjugates of original data as (10) (11), the new transform is not only transmitted the conventional space time information, also its conjugate information is delivered by a quasi-orthogonal interleaver case. As a consequence the proposed transform then can be written as

\[
[x_i, y_i, x_j, y_j] \begin{bmatrix}
1 & 1 & 0 \\
1 & -1 & 0 \\
1 & 1 & -1 \\
0 & 1 & 0
\end{bmatrix} \begin{bmatrix}
1 & 0 & 0 & D^{-1} \\
0 & 1 & D^{-1} & 0 \\
0 & -D^{-1} & 1 & 0 \\
-D^{-1} & 0 & 0 & 1
\end{bmatrix}
= [y_i, s_i^* s_i s_i]
\]

\[(14)\]

where the $D^{-1}$ time delay can be simply designed out of the transform. After that the transmitted data can be listed as a permutation function

\[
\begin{bmatrix}
S_1 \\
S_1^* \\
S_2 \\
S_2^*
\end{bmatrix}
\rightarrow
\begin{bmatrix}
Antenna 1 \\
Antenna 2
\end{bmatrix}
\]

\[
\begin{bmatrix}
s_1 \\
s_2 \\
0 \\
0
\end{bmatrix}
\begin{bmatrix}
-D^{-1}s_2^* \\
D^{-1}s_1^* \\
0 \\
0
\end{bmatrix}
= \begin{bmatrix}
S_1 \\
S_2 \\
S_1^* \\
S_2^*
\end{bmatrix}
\]

\[(15)\]

And the transmission matrix can be plotted, as shown in Fig. 5.

![Fig. 2. Proposed transform matrix for space time block codes](image)

As shown in the Fig.2, the transmitted matrix $T$ without time delay function can be easily written as

\[
T = \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & -1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & -1
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 \\
0 & -1 & 1 & 0 \\
-1 & 0 & 0 & 1
\end{bmatrix}
= \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 \\
-1 & -1 & 1 & 1 \\
1 & -1 & 1 & -1
\end{bmatrix}
\]

\[(16)\]

its inverse is its transpose and
we can obtain it from the Hadamard matrix after a special permutation matrix \( P_4 \)

\[
T = [P_4'] [H_4] \quad (18)
\]

where \( H_4 \) is the four by four Hadamard transform. The permutation matrix \( [P_4'] \) simply generalizes the time delay and permutation processing to one matrix, and it is shown in detail as Fig. 3.

**Fig. 3.** Analysis of the permutation matrix related to conventional design

In Fig. 3, to show the filter in time domain representation of whole analysis bank, we have

\[
\begin{bmatrix}
    \text{Time1} \\
    \text{Time2}
\end{bmatrix} = \begin{bmatrix}
    L \\
    B
\end{bmatrix} = \begin{bmatrix}
    1 & 0 & 0 & 0 \\
    0 & 0 & 0 & 1 \\
    0 & 0 & -1 & 0 \\
    0 & 1 & 0 & 0
\end{bmatrix}
\]

The combined matrix is invertible, its inverse is the transpose as

\[
\begin{bmatrix}
    L \\
    B
\end{bmatrix}^{-1} = \begin{bmatrix}
    L^T & B^T
\end{bmatrix} = \begin{bmatrix}
    1 & 0 & 0 & 0 \\
    0 & 0 & 0 & 1 \\
    0 & 0 & -1 & 0 \\
    0 & 1 & 0 & 0
\end{bmatrix}
\]

The second matrix \( [L^T \quad B^T] \) represents the synthesis bank. This is an orthogonal filter bank, because \( \text{inverse} = \text{transpose} \). The channels \( L \) and \( B \) of an orthogonal filter are represented in the time domain by a combined orthogonal matrix, it is shown as

\[
\begin{bmatrix}
    L^T & B^T
\end{bmatrix} = \begin{bmatrix}
    0 & 0 & 0 & -1 \\
    0 & 1 & 0 & 0
\end{bmatrix}
\]

The synthesis bank is the transpose of the analysis bank. When one follows the other we have perfect reconstruction as Fig. 4.
Based on the above introduction (16-21), easily we can construct the space time code with fast Hadamard transform and a permutation filter bank channel as Fig. 5.

$$[F'] = ([1 \ 0 \ 0 \ 1] \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + [D^{-1}]([0 \ 1 \ 0 \ 0] \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix})) [F']$$

To reduce the operation, an improved algorithm also can be given as

$$T = [P'] ([H_2 \otimes H_2] ([U_A \otimes U_B]) ([H_2 \otimes I_2])$$

$$= \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} [H]$$

$$= \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

(22)

where $U_A = [I_2], U_B = [1 \ 0 \ 0 \ 1 \ 0 \ 0]$. The amount of computation needed at each stage of this proposed is 8 real additions and 1 multiplication. Since the transform matrix is constructed by a quasi-orthogonal design, then we have

$$r_1 = h_0 s_1 + h_1 s_2 + n_1, \quad r_2 = -h_0 s_2^* + h_1 s_1^* + n_2$$
$$r_3 = h_0' s_3 + h_1' s_4 + n_3, \quad r_4 = -h_0' s_4^* + h_1' s_3^* + n_4$$

(23)  

(24)

the channel is denoted as

$$H = \begin{bmatrix} h_0 & h_1 & 0 & 0 \\ h_1^* & -h_0^* & 0 & 0 \\ 0 & 0 & h_0' & h_1' \\ 0 & 0 & h_1'^* & -h_0'^* \end{bmatrix}$$

(25)

and
\[ H^H H = \begin{bmatrix} h_0^* & h_0 & 0 & 0 \\ h_0 & h_1 & 0 & 0 \\ h_1^* & -h_0^* & 0 & 0 \\ 0 & 0 & h_0' & h_1' \\ 0 & 0 & h_1'^* & -h_0'^* \end{bmatrix} \begin{bmatrix} h_0^* \\ h_0 \\ h_1^* \\ -h_0^* \\ 0 \end{bmatrix} = \begin{bmatrix} |h_0|^2 + |h_1|^2 \end{bmatrix} I_2 \]

where \( h_0 \), and \( h_1 \) is the first transmitted path gain, the \( h_0' \), \( h_1' \) is the next time slot path gain.

Let

\[ B = \begin{bmatrix} s_1 - \bar{s}_1 & s_2 - \bar{s}_2 & 0 & 0 \\ -s_2^* + \bar{s}_2^* & s_1^* - \bar{s}_1^* & 0 & 0 \\ 0 & 0 & s_3 - \bar{s}_3 & s_4 - \bar{s}_4 \\ 0 & 0 & -s_4^* + \bar{s}_4^* & s_3^* - \bar{s}_3^* \end{bmatrix} \]

then

\[ B^* B = \begin{bmatrix} s_1^* - \bar{s}_1^* & -s_1 + \bar{s}_1 & 0 & 0 \\ s_2^* - \bar{s}_2^* & s_1 - \bar{s}_1 & 0 & 0 \\ 0 & 0 & s_3^* - \bar{s}_3^* & -s_3 + \bar{s}_3 \\ s_4^* - \bar{s}_4^* & s_3 - \bar{s}_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} |s_1 - \bar{s}_1| \\ -s_1 + \bar{s}_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (|s_1 - \bar{s}_1|^2 + |s_1 - \bar{s}_1|)I \end{bmatrix} \]

3 Conclusion

The paper propose a transform representation for space time block codes. The numerical results show that we can use several permutations and Hadamard transform to design a filtering system for diversity transmission.

References

[1] Siavash M. Alamouti, “A Simple Transmit Diversity Technique for Wireless Communications,” IEEE JSAC, Vol. 16, No. 8, Oct. 1998.
[2] Tarokh V, Seshadri N, and Calderbank A. R, “Space-time codes for high data rate wireless communication: performance criteria and code construction,” IEEE Trans. Inform. Theory, Mar. 1998.
[3] Tarokh T, Jafarkhani H and Calderbank A. R, “Space Time Block Codes From Orthogonal Designs,” IEEE Trans. Information Theory, vol.45, pp.1456-1467, July 1999.
[4] Radon J, “Lineare Scharen orthogonaler matrizen,” in Abhandlungen aus dem Mathematischen Seminar der Hamburgischen University, vol. I, pp. 1-14, 1922.
[5] Jafarkhani H, “A Quasi-orthogonal space time block code,” IEEE Trans. Commun., vol.49, No.1, pp. 1-4, 2001.
[6] Tirkkonen O, Boariu A and Hottinen A, “Minimal non-orthogonality rate one space time block code for 3+ Tx antennas,” ISSSTA 2000, pp.429-432, September 2000.
An Multimedia Database System Using Mobile Indexing Agent in Wireless Network

Jong-Hee Lee¹, Kwang-Hyoung Lee¹, Moon-Seog Jun¹, and Keun-Wang Lee²

¹School of Computing, Soong-sil University, Korea
²Dept. of Multimedia Science, Chungwoon University, Korea
multistar@freechal.com

Abstract. Indexing and retrieval techniques for various multimedia data have become more and more important as a diverse range of multimedia-based content services are increasingly provided in wireless network. This paper presents a semantics-based video retrieval system that lets the user retrieve the meanings of large capacity multimedia data in a various way that employs annotation- and feature-based retrieval mechanisms in wireless network. As extract comparison area of optimized after learn feature comparison area, proposed system could reduce number of comparison area remarkably. Also, showed experiment result that improve more recall and precision than previous research.

Keywords: Multimedia Database System, Mobile Indexing Agent, Semantics-based Indexing and Retrieval.

1 Introduction

An efficient management of multimedia data, on the one hand, requires the technology that enables the systematic classification and integration of large capacity multimedia data. On the other hand, multimedia data requires efficient and effective retrieval and storage in order to provide users with their desired information for multimedia data according to diverse environments around them like mobile hosts[1, 2].

However, a mobile host has many more limiting factors than a desktop computer in terms of CPU processing rate, battery capacity, and screen size[3, 4]. For the mobile host, lower bandwidth and CPU processing rate, in particular, are inhibitors to the seamless provision of multimedia services by the server side. For this reason, the effective retrieval and regeneration of multimedia data at a mobile host essentially requires improved host CPU performance and advances in network technology as well as systematic indexing technologies applicable to transmitted multimedia data. So far, a number of active research initiatives have been developed to address the limitations of mobile hosts[5, 6]. However, such video indexing techniques are merely a method for classifying different video genres or types, instead of translating user requirements into mobile hosts.

For annotation-based retrieval system as VideoSTAR[7], the user uses characters to directly embed annotations to the semantics information (which permits
little automatic recognition) into multimedia data storage so that he/her can use previously added annotations to conduct retrieval[8]. This approach allows the user to make annotations to video contents while watching a video. This may lead to the accurate representation and retrieval of the video contents. However, due to the resulting cumbersomeness of getting users to add annotations to all videos with characters, this content-based multimedia data retrieval not only takes too much time and effort, but also causes a significant increase in the amount of unnecessary annotations. This approach also cannot ensure the accuracy of the retrieval results because individual users as annotators attach their own annotations to video scenes.

Feature-based retrieval system as VisualSEEK[9] enables the user to extract low-level feature information from multimedia data such as color, texture, regional information, and spatial color distribution [10]. This retrieval technique highlights comparative retrieval through the analysis of the similarity among visual features extracted from the video. Accordingly, extraction of visual features is of critical importance. However, the accurate extraction of a lot of video feature information is quite challenging. Furthermore, it is not so easy to perform retrieval while matching the extracted feature information with large capacity multimedia data.

This paper presents a semantics-based video retrieval system using an indexing agent in a mobile environment, which is geared toward learning the result of user queries and automatically updating meta data on the multimedia database server in a continuous manner.

2 Mobile Indexing Agent System

2.1 System Architecture

The proposed system largely consists of three parts: video server, agent middleware, and mobile client. Figure 1 shows the overall architecture of the proposed system.

![Fig. 1. System Architecture](image-url)
The multimedia database server stores all information of annotation data, feature data, comparison area data, video data and its metadata. Once video data are inputted into the system, the video server extracts annotation/feature information while updating its metadata continuously. The agent middleware processes the queries WAP Gateway receives from users through a query processor while creating the queries accessing the video server. Additionally, it receives another response to queries, processes annotations to keyframes and images, and transmits both keyframe images and WAP responses to the client side. The mobile client accesses the agent middleware via wireless service provider’s wireless networks, their infrastructure, and WAP Gateway where mobile hosts are already built and in use.

2.2 Automatical Indexing Processing

The indexing agent extracts keywords from the queries submitted from the user for video retrieval, and matches them with basic annotation information stored in the metadata. Then, it detects the keyframes that has the same keywords as annotation information, and sends them to the user.

Figure 2 presents the architecture of annotation-based retrieval performed by the indexing agent in the Agent Middleware.

Once entered, queries submitted from the user are analyzed, and keywords are extracted. The user’s keywords extracted are matched with the annotation information of metadata stored in the annotation database. As a result of matching, the keyframes having exactly matched keywords are detected in the database, and then they are sent to the user. Additionally, the keywords that do not exactly match annotation information among queries received from the user are defined as potential keywords.

Whenever auto-annotation processing is performed by user queries, the keywords extracted from the queries sent by individual users are matched with the keywords of the example images. As a result, exactly matched keywords receive higher sim-
ilar weight while unmatched keywords receive lower similar weight. This results in more concrete semantics annotations to the corresponding keyframe.

3 Mobile Indexing Mechanism

3.1 Automatic Annotation Processing Mechanism

We had proposed algorithm about automatic annotation updating and feature detection processing in previous research [11]. Where annotation keywords in the keyframe are same keywords, new semantics weight is calculated as Formula 1:

$$W_{Keyword\_new} = W_{Keyword\_old} + \frac{1}{N_{Kframe\_SK}}$$

(1)

While $W_{Keyword\_new}$ is the new semantics weight for annotation keywords and $W_{Keyword\_old}$ is the previous semantics weight for annotation keywords. $N_{Kframe\_SK}$ is the number of keyframes with same keywords.

In the meantime, where annotation keywords in the keyframe are difference keywords, new semantics weight is calculated as Formula 2:

$$W_{Keyword\_new} = W_{Keyword\_old} - \frac{1}{N_{Kframe\_SK}}$$

(2)

3.2 Feature Processing Mechanism for Keyframes

An n*m image extracted by a keyframe partitions its pixels into 3*3 regions. Accordingly, a partitioned single image has 9 pixel groups. The average of R,G, and B for the whole keyframe image is calculated as Formula 3:

$$A_{t\_image,red} = \frac{\sum_{i=0}^{w} \sum_{j=0}^{h} R_{[i,j]}}{C_t}$$

(3)

$R_{[i,j]}$ is the value for Red in row i and column j of the whole image. $C_t$ is the number of pixels that has the value for Red(≠ 0) in the whole image.

The value of Red color difference between query images and the whole keyframe image is obtained as follows:

$$D_{t\_red} = |A_{t\_image,red} - A_{t\_frame,red}| \quad \text{(Where, } D_{red} < \alpha)$$

(4)

$A_{t\_image,red}$ is an average value of Red for the whole query image while $A_{t\_frame,red}$ is an average value of Red for the whole keyframe image. $\alpha$ is a threshold.

Formula 5 is used to calculate the similarity between the entered images and the images selected as a primary similar keyframe candidate group.

$$ST_{Q\_image,K\_frame} = D_{t\_red} + D_{t\_green} + D_{t\_blue}$$

(5)
Ultimately, the value obtained where the sum of the difference between R, G, and B is below the threshold value is selected as a primary candidate frame.

Comparison Area Learning Mechanism. We propose algorithm that detect optimized comparison area in order to reduce overhead of system in feature comparison of keyframes in this paper. Mobile indexing agent learns comparison area by user’s last keyframe selection and detect optimized comparison area. Mobile indexing agent calculate color difference weight of comparison area for optimized comparison area detection. Color difference weight of comparison area expresses each area color difference with query image and other keyframe images which become comparison target by weight. Color difference weight calculation method to calculate the color difference rate of comparison area is as following.

\[
W_{\text{diff-color}}[n] = 1 - \frac{R[n] - 1}{N} \quad \text{(6)}
\]

Where, \( R[n] \mid R[n] \in \text{Rank} \)

\( W_{\text{diff-color}}[n] \) is color difference weight of comparison area, is integer between 1 to 9, \( R[n] \) is order value of each area, \( A_{w\_new}[n] \) is accumulation average about present \( W_{\text{diff-color}}[n] \), \( A_{w\_new}[n] \) is accumulation average about previous \( W_{\text{diff-color}}[n] \), \( \beta \) is threshold value for comparison area extraction, N is multi-segmentalized comparison area number.

Whenever user search is achieved, calculate accumulation average about color difference weight continuously. \( A_{w\_new}[n] \) that is new \( W_{\text{diff-color}}[n] \) accumulation average calculates with that add present comparison area color difference weight in previous accumulation average.

\[
A_{w\_new}[n] = \text{Avg}[A_{w\_old}[n] + W_{\text{diff-color}}[n]] \quad \text{(Where, } A_{w\_new}[n] \geq \beta) \quad (7)
\]

After choose area that have value more than threshold value \( \beta \) in accumulation average value of color difference weight that have real value between 0 to 1, decide this area by optimized comparison area.

4 Implementation and Experimental Evaluation

4.1 Implementation

Figure 3 shows the user interfaces for retrieval for video data at a mobile host. The user can find the keyframe of his/her primary target scene by entering several query words. If primary annotation-based retrieval is completed, keyframe lists are displayed on the user’s terminal one by one. Then, the user can select his/her desired image to perform feature-based retrieval. If the user performs feature-based retrieval, he/she can calculate the similarity for query images by using the multiple partition color histogram mechanism and get keyframe lists displayed in order of highest similarity.
4.2 Experimental Evaluation

MPEG formatted movie video files were used as domains for video data in order to evaluate the proposed system. For the purpose of this study, we used some 30 movies that had corresponding videos totaling 38 video clip files, and detected a total of 31,519 keyframes. By default, a single annotation was attached to 5,657 keyframes, except that a keyframe has duplicate portions due to successive cuts or that individual objects of a keyframe are indistinguishable.

In order to evaluate the proposed system for overall retrieval precision and recall, user queries were performed more than 500 times. Figure 4 illustrates the precision and recall of retrieval for this system.

![Fig. 3. Mobile & Web Client Retrieval Interface using Mobile Indexing Agent](image)

![Fig. 4. Retrieval Recall & Precision of the Proposed System](image)
5 Conclusions

This paper presents a video retrieval system that allows the user to perform various semantics retrievals using annotation- and feature-based retrieval techniques for large capacity multimedia data in wireless network.

As with experiment results, the proposed system was able to improve retrieval precision for semantics-based video data retrieval as well as produce high precision at an approximate rate of 95.1 of testing and evaluating user queries. Proposed system showed precision that improve more using optimized comparison area detection and reduced overhead of system and retrieval time in mobile phone.

References

1. Sibel Adali, et al., “The Advanced Video Information System : data structure and query processing,” Multimedia System, pp.172-186, 1996.
2. B. Y. Ricardo and R. N. Berthier, Modern Information Retrieval, ACM press, 1999.
3. T. Kamba, S. A. et al., “Using small screen space more efficiently,” In Proceedings of CHI’96, ACM Press, 1996.
4. O.Buyukkokten, H. et al., “Power Browser: Efficient Web Browsing for PDAs,” In Proceedings of CHI’2000, ACM Press, 2000.
5. Jiang Li. et al., “Scalable portrait video for mobile video communication,” IEEE Trans on CSVT, Vol. 13, No. 5 pp.376-384, 2003.
6. C. Dubuc. et al., “The design and simulated performance of a mobile video telephony application for satellite third-generation wireless systems,” IEEE Trans on Multimedia, Vol. 3, No. 4 pp.424-431, 2001.
7. Rune Hjelsvold, VideoSTAR - A Database for Video Information Sharing, Ph. D. Thesis, Norwegian Institute of Technology, Nov. 1995.
8. H. Rehatsch, and H. Muller, “A Generic Annotation Model for Video Database,” The 3rd International Conference on VISUAL ’99, Amsterdam, Netherlands, pp.383-390, 1999.
9. J. R. Smith and S. F. Chang, “VisualSEEK : a fully automated content-based image query system,” ACM Multimedia, Boston, 1996.
10. Chong-Wah Ngo, Ting-Chuen Pong, Hong-Jiang ZhangOn, “Clustering and retrieval of video shots through temporal slices analysis,” IEEE Trans on Multimedia, Vol.04 No.04 pp.0446-0458, 2002.
11. Jonghee, Lee, Keunwang, Lee, “A Semantics-based Video Retrieval System using Indexing Agent”, Conference on KMMS ’03, Vol. 6, No. 2, pp.281-284, 2003.
Bus Arrival Time Prediction Method for ITS Application

Bongsoo Son\textsuperscript{1}, Hyung Jin Kim\textsuperscript{1}, Chi-Hyun Shin\textsuperscript{2}, and Sang-Keon Lee\textsuperscript{3}

\textsuperscript{1}Dept. of Urban Planning and Engineering, Yonsei University, Seoul, Korea  
\{sbs,hyungkim\}@yonsei.ac.kr
\textsuperscript{2}Dept. of Transportation Engineering, Kyonggi University, Kyunggi-do, Korea  
chshin@kyonggi.ac.kr
\textsuperscript{3}Korea Research Institute For Human Settlements, Kyunggi-do, Korea  
sklee@krihs.re.kr

Abstract. It is known that stoppage times at signalized intersections cause the biggest errors in bus arrival time prediction for real-time Bus Information System (BIS) services and no particular method is proven successful so far. This study developed a prediction method that compares the predicted bus travel times from bus stop to the stop line at signalized intersections by using Kalman filtering technique with the state of green time indications of traffic signals, and then incorporates possible stoppage into a next link travel times. From field surveys and in-lab simulation, the proposed method was found superior to other conventional techniques showing an average of more than 200\% improvement in prediction accuracy.

1 Introduction

In order to combat the ever-increasing congestion in urban area, traffic officials and agencies have embraced Intelligent Transportation Systems (ITS) and technologies in a proactive, systematic way. In recent years interests has grown for granting preferential service to public transport such as buses, since it is desirable to favor public transit over private auto travel due to the basic role of public transport in the city. A number of different public bus preferential treatments have been implemented in many urban areas around the world to offer better service to public buses than private autos. Among bus preferential schemes such as bus gate, bus malls, bus lanes, BIS, bus priority signals, etc., bus information system is one utilized in many cities to balance public and private transport. In fact, the availability of accurate, real-time information is especially useful to operators of vehicle fleets as well as passengers.

However, BIS has not much impacted on the reliability of bus arrival time information. The reason for uncertainty in predicted arrival times of public buses in BIS are waiting times for green phase at the signalized intersections, dwell-times at bus stops, delays resulted from incidents, and so on. Reliability of prediction has increased recently through numerous prediction methods that try to take some of these reasons into considerations [1, 2].

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 88–94, 2004.  
© Springer-Verlag Berlin Heidelberg 2004.
As can be seen from Figure 1, it is well known that waiting times at the signalized intersection cause major errors in forecasting of bus arrival times. In the figure, bus travel trajectories between bus stop $i-2$ and bus stop $i$ are not constant. More specifically, bus type I is associated with the buses arrived at the traffic signal $i-1$ during the red time period. Bus type II represents the buses arrived at the signal during the period between ending of red time and beginning of green time and experienced delay for passing the traffic signal. Bus type III is related to the buses passed the traffic signal without any delay. The three types of bus trajectories indicate that the bus arrival times at bus stop $i$ are widely different and greatly dependent upon whether or not buses await at traffic signal $i-1$, so bus travel times between the two bus stops significantly vary depending upon the state of signal $i-2$.

Figures 2 and 3 show the bus travel times measured between bus stop $i-2$ and traffic signal $i-1$ and between traffic signal $i-1$ and bus stop $i$, respectively, during the same time period. The bus travel times between bus stop $i-2$ and traffic signal $i-1$ are severely fluctuated, while those between traffic signal $i-1$ and bus stop $i$ are relatively somewhat stable. The main cause of travel time fluctuation is the fact how long the bus does wait at traffic signal $i-1$. From the two figures, it is very clear that that waiting times at the signalized intersection cause major errors in forecasting of bus arrival times.

Up to date there is no practical and easy method that accounts for the impact created by traffic signal operation along the bus routes. This paper is an attempt to improve the reliability of predicted bus travel time in urban signalized intersection networks. In the paper, we have come to the result that we can figure out the waiting times at signalized intersections under the Time-Of-Day (TOD) type operation by comparing the state of signals at that very instant if the bus arrival times at the stop line of a signalized intersection are reasonably estimated. The prediction method is described below. The data for this paper came from several sources: specific sources will be given when individual data are discussed in Section 3.
2 The Prediction Method

The method proposed in this paper is based upon two forecasted link travel times of buses: one from a bus stop to a stop line of intersection and another from the stop line to the next downstream bus stop which is supposed to display the remaining time until a next bus of a particular bus line arrives. Conventional techniques have usually considered the distance between two successive bus stops as one link. (See Figure 4)
The information of local bus arrival times between consecutive bus stops needed for the prediction of link travel times can be obtained from ITS devices such as GPS devices with wireless data communication modules atop buses. Then, the GPS based modules have an electronic road map and do map-match for bus stops in order to transmit data. Each link travel time is predicted using Kalman Filtering in which state equations by time are formulated [3]. A particular concern is the link travel time from signalized intersections to immediate downstream bus stop for which two state equations were developed to differentiate signal-awaiting trips at the signalized intersections from non-stop trips.

The computer algorithm used for this method works on a rigid space-time grid. For the analysis, the site interested are broken into sections bounded by grid points \( \{i, i-1, \ldots, i-n\} \) located at each bus stop and at each stop line of signalized intersection as shown in Figure 2. The computational procedure starts at the most upstream point of the site interested, and works downstream (i.e., the bus traveling direction).

Figure 5 is a schematic diagram of the computational procedure. In the figure, amount of waiting times at each signalized intersection can be estimated by comparing the predicted bus arrival times at the subject stop lines with the state of signals (i.e.,
Time-Of-Day signal operation time schedule) at that very instant. For the TOD type signalized intersection, the signal operation time schedules are readily available from the traffic officials and agencies. When stopping at signal is anticipated, the waiting time is then added to the next link travel time that is also predicted. Otherwise, the predicted bus arrival time is determined by the next link travel time, where the next link travel time can be estimated based on the information associated with the proceeding buses traveled on the downstream link at the same time period.

3 The Data Set

The study site is located near Seoul Metropolitan Area. The study site is an 8-lane busy urban arterial highway and approximately 2 km long with four bus stops and three-signalized intersections. The route is therefore segmented into 6 links with consecutive links sharing start and end nodes. The links range from 0.3 to 1.0 km in length. A total of sixteen different bus-lines are in operation along this route. Among them, three bus-lines are currently in BIS service.

| Bus stop I – 2 | Traffic signal i-1 | Travel time from bus stop i-2 to signal i-1 |
|---------------|-------------------|----------------------------------------|
| Number of bus line | Arrival time | Number of bus line | Arrival time | |
| 333 | 8:00:00 | 333 | 8:02:06 | 0:02:06 |
| 51 | 8:00:28 | 51 | 8:02:19 | 0:02:19 |
| 820 | 8:00:37 | 820 | 8:02:20 | 0:02:20 |
| 57 | 8:00:56 | 57 | 8:02:28 | 0:02:28 |
| 736-1 | 8:03:27 | 736-1 | 8:05:27 | 0:02:00 |
| 700 | 8:04:58 | 700 | 8:05:38 | 0:00:40 |
| 1116 | 8:05:40 | 1116 | 8:06:00 | 0:00:20 |
| 333 | 8:06:03 | 333 | 8:08:18 | 0:02:15 |
| 220 | 8:06:22 | 220 | 8:08:27 | 0:02:05 |
| 51 | 8:08:53 | 51 | 8:09:16 | 0:00:23 |
| 17 | 8:11:39 | 17 | 8:14:18 | 0:02:39 |
| 220 | 8:11:54 | 220 | 8:12:09 | 0:00:15 |
| 820 | 8:12:11 | 820 | 8:14:23 | 0:02:12 |
| 116-3 | 8:12:29 | 116-3 | 8:14:30 | 0:02:01 |
| 57 | 8:14:37 | 57 | 8:14:56 | 0:00:19 |

Table 1. Real data set for the bus arrival times collected from the field by manual.
The GPS data for the real time bus locations for the three bus-lines were obtained from SK-Entrac that the private agency provides BIS service in Korea. For the other bus-lines, the bus arrival times at signalized intersections and bus stops were directly collected from the field by manual for the three weekdays from October 14, 2003 to October 16, 2003 during the time period from 08:00 to 12:00. The signal operation time schedules were obtained from the police station. It should be noted that the signalized intersections are operated and/or controlled by the police in Korea.

Table 1 shows real data set for the bus arrival times collected at two consecutive bus stops during the peak time period between 08:00 and 08:15. Average headway of buses during peak period between 08:00 and 10:00 is 44 seconds, and 57 seconds during non-peak period between 10:30 and 12:00.

4 Validation Test

For the model validation test, the accuracy of the results obtained from the proposed method was measured by using three statistical errors such as Mean Relative Error (MRE), Root Squared Error (RSE) and Maximum Relative Error (MAX). Besides, the performances of proposed method were compared with those of four conventional methods such as 1) moving average method, 2) exponential smoothing method, 3) autoregressive moving average method (ARIMA) and 4) Kalman Filtering method. For these purposes, predicted bus arrival times obtained by using the four conventional techniques and the proposed method were compared with observed arrival times obtained from the field measurements.

The results for the statistical analysis of the five methods were summarized in Table 2. As shown in the table, the proposed method is superior to the four conventional methods showing an average of more than 200% improvement in prediction accuracy. Among the conventional methods, Kalman Filtering method has produced the better results than the other three methods. The other methods show the same level of prediction performances in terms of statistical accuracy.

Table 2. The statistical analysis results for the prediction techniques’ performances
Since most of signalized intersections along urban streets operate with cycle lengths of 120 seconds or more, the instance of waiting for the next green phase or non-stop passage of buses at critical intersections would create difficulty in bus arrival time forecasting efforts. It was concluded that the proposed method can offer prompt bus arrival time to bus passengers waiting at bus stops with relatively higher accuracy. Applicability of the proposed model is considered much higher in non-peak time and nighttime BIS service operation.

5 Conclusion

BIS is one utilized in many cities to balance public transport and private transport. The availability of accurate, real-time information is especially useful to operators of vehicle fleets as well as passengers. However, BIS has not much impacted on the reliability of bus arrival time information. It is well known that waiting times at the signalized intersection cause major errors in forecasting of bus arrival times. This paper is an attempt to improve the reliability of predicted bus travel time in urban signalized intersection networks. The emphasis of the proposed method is on estimating waiting times at signalized intersections under the TOD type operation by comparing the state of signals at that very instant.

The performances of the method proposed in this paper were measured by using three statistical errors such as MRE, RSE, and MAX. The performances of proposed method were also compared with those of four conventional methods by using the observed arrival times obtained from the field measurements. The proposed method is superior to the four conventional methods showing an average of more than 200% improvement in prediction accuracy. Among the conventional methods, Kalman Filtering method has produced the better results than the other three methods.

References

1. Wei-Hua Lin and Jian Zeng, “An Experimental Study on Real Time Bus Arrival Time Prediction with GPS Data”, Virginia Polytechnic Institute and State University, 1999.
2. Amer Shalaby and Ali Farhan, “Bus Travel Time Prediction Model for Dynamic Operations Control and Passenger Information System”, Transportation Research Board, 2003.
3. Grewal, M.S. and Andrews, A.P., “Kalman Filtering Theory and Practice”, Prentice Hall, 1993.
RRAM Spare Allocation in Semiconductor Manufacturing for Yield Improvement

Youngshin Han and Chilgee Lee

School of Information and Communication Engineering SungKyunKwan University,
300 Chunchun-dong, Jangan-gu, Suwon,Kyunggi-do 440-746, S. Korea
yshan@ece.skku.ac.kr

Abstract. It takes about four to five weeks to fabricate a semiconductor memory device. As the fabrication process consists of various operations, there is a possibility of fabricating a final product with defects. It would be impossible for us to repair a memory device if it has numerous defects that cannot be dealt with properly. However, in case of a small number of defects, it is desirable to reuse a defective die (standard unit measuring a device on a wafer) after repair rather than to discard it, because reuse is an essential element for memory device manufactures to cut costs effectively. To perform the reuse, laser-repair process and redundancy analysis for setting an accurate target in the laser-repair process is needed. In this paper, cost reduction was attempted by reducing time in carrying out a new type of redundancy analysis, after simulating each defect.

1 Introduction

Recently, it has become possible to mass-product high-capacity memory devices due to technological development in integration such as very large scale integration (VLSI) and wafer scale integration (WSI). These improvements have made it possible to use redundancy extensively in fabricating memory chips and repairing devices with faulty cells. This type of memory device is called redundant random access memory (RRAM). Redundancy is comprised of spare cells made of spare rows and columns, which substitute those for faulty cells. A fabrication method using redundancy is called laser-repair. This method plays a significant role in cutting costs by increasing the yield of faulty devices. Until now, many experiments have been conducted to repair RRAM in the form of wafers. Repair of the algorithm consists of gathering addresses of rows and columns that can be repaired with fault counters on every row and column consisting of a memory device [1]. The fault-driven comprehensive redundancy algorithm [2] depends on user-defined preferences to achieve an optimal repair-solution in every possible collation. In efficient spare allocation in reconfigurable arrays [3], two new algorithm methods for RRAM repair were suggested. The first algorithm uses the branch-and-bound approach which screens subjects in the initial stage of fabrication. The second one uses a heuristic criterion. Redundancy analysis simulation, beyond the idea of the conventional redundancy analysis algorithm, aims at reducing the time spent in the process and strengthening cost.
competitiveness by performing redundancy analysis after simulating each case of defect.

This paper is composed of four parts. Part two gives an explanation on redundancy analysis. Part three shows the performed simulation with analyzed results. Finally part shows the results a future work for research is proposed.

2 Redundancy Analysis Algorithm

In EDS, the wafer test makes it possible to locate defects in a major cell including the location of defect, address of rows and columns. With this information, redundancy analysis allots spare cells of a die (i.e., redundancy analysis is a process of allotting spare cells to repair the defects occurred in a device effectively).

![Fig. 1. Repair process with spare cells](image)

Fig. 1 shows how spare cells (a, b) are used in repairing defects (A, B) in a device with one cell in a spare row and two spare cells in a spare column. If defects (A, B) occur in the direction of the row, it would be more efficient to repair the defects with a spare cell in the row rather than with two cells in the column. Additionally, in the case of only one defect, the result would be the same whether a spare cell in the row or a spare cell in the column is used. In this case, according to priority, a row or column with more spare cells will be used.

There are many redundancy analysis algorithms to repair RRAM. The repair-most algorithm records the number of defects in the counter for each row and column [4]. Fig. 2 shows recorded defect values in the counter of each row and column available redundant columns (ARC) and, available redundant rows (ARR) are counters marking the number of spare cells. As seen in Fig. 3 the repair-most algorithm repairs the row and column with the highest value first, and the process is repeated until all spare cells are used or when all defects are repaired. Generally, most defects can be fixed with the repair-most algorithm. However, Fig. 4 shows some cases in which cells cannot be repaired even with the repair-most algorithm.

In these cases, defects can be repaired with the fault-driven algorithm as seen in Fig. 5. The fault-driven algorithm consists of two phases [5]. The first is forced-repair analysis choosing a certain row or column that will be substituted by spare cells located in the same direction as the defect. The second is the sparse-repair analysis that determines the repair method of remaining defects after the first phase by using
Fig. 2. Example of the occurrence of a defect

Fig. 3. Repairing a row and a column with the highest value

Fig. 4. Cells cannot be repaired even with the repair-most algorithm

Fig. 5. Fault-driven algorithm

Fig. 6. Result of the FLCA algorithm
spare cells that were not used in forced-repair analysis. In the fault-driven algorithm, number of records are generated in repairing the defects. In the fault line covering approach algorithm (FLCA), number of records, less than the number of records in the fault-driven algorithm, are generated [6]. \( R_A \) represents the number of spare cells in a row, \( C_A \) represents the number of spare cells in a column, and \( T_F \) represents the number of single faults.

The FLCA algorithm shows a way to obtain a repair solution through only two rounds of tests as seen in Fig. 6. We should note that parents (B) and (C) have only one descendant, \( (M_F > R_A \text{ and } C_A = 0) \). As parent (D) is \( S_F > R_A + C_A \), it cannot be used. Regarding this, the value of the repair solution will be the record of (E) in Fig. 6 [7].

### 3 Redundancy Analysis Simulation

#### 3.1 Goal of the RA Simulation and Experiment

The goal of the RA simulation is to minimize the time spent by applying a correlation technique to the repair solution analyzed by conventional RA simulation. Using the correlate limit generated when repairing defects of devices takes less time than using the conventional RA algorithm, which in turn decreases a great deal of cost in fabricating memory devices. Visual C++ / Windows 98 is used for the experimental environment. In order to achieve optimal effectiveness, two spare cells only with columns are selected. The size of a device is 1K rows (1024 bit), and 1K columns (1024 bit), i.e., 1M bits. The input element of the simulation is the fail bit map and random correlate limit. The first correlate limit is 0.0. By using this, we will produce the most optimal correlate limit that not only can be repaired through RA simulation but also is the standard for similarity.

#### 3.2 Fail Bit Map Information

A 16 digit figure in a file represents a fail bit of 32bit, and 1001 on the binary scale. This information shows that there are 8bit of defects in most significant bit (MSB) and least significant bit (LSB), and no defects in the middle 2×8 bit. As this paper assumed that a device has a size of 1024 × 1024, the file of fail bit map information

---

| 1234567890 | 1234567890 |
|------------|------------|
| 0000000000 | 0000000000 |
| A000000000 | A000000000 |
| 0000000000 | 0000000000 |
| 0000000000 | 0000000000 |
| 0000000000 | 0000000000 |
| 0000000000 | 0000000000 |
| 0000000000 | 0000000000 |

---

Fig. 7. Fail Bit Map information

will be a perfect square comprised of 32, 16 digit numbers with rows and columns. Fig. 7 shows the sample of fail bit map information file. The character of Row 2 and
Column 1 indicates hexa “A”, hexa “A” can be converted to binary “1010”. So, 1 means ‘fail’ and 0 means ‘pass’ for memory device. The color grade of simulation results are originated from these characters.

3.3 Correlation Definition and Process

The mean is one measure of the central tendency of a random variable.

1. \( E(cX) = cE(X) \) (c is constant)
2. \( E(X+Y) = E(X) + E(Y) \)

The variance of the random variable \( X \), denoted by \( \sigma^2 \) or \( \text{Var}(X) \), is given by

3. \( \sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2 \) (\( \mu \) is mean)

The variance is a measure of the dispersion of a random variable about its mean.

4. \( \text{Var}(cX) = c^2 \text{Var}(X) \)
5. \( \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \) if \( X,Y \) are independent

The covariance between the random variables \( X \) and \( Y \), denoted by \( \text{Cov}(X,Y) \), is defined by

6. \( \text{Cov}(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y) \)

The covariance is a measure of the dependence between \( X \) and \( Y \). Note that \( \text{Cov}(X,X) = \text{Var}(X) \).

Independent random variables are also uncorrelated.

7. \( \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y) \)

If \( X \) and \( Y \) are independent, then

8. \( \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) \)

The correlation between the random variables \( X \) and \( Y \), denoted by \( \text{Cor}(X,Y) \), is defined by

9. \( \text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \)

It can be shown that \(-1 \leq \text{Cor}(X,Y) \leq 1[8]\).

We will store all the possible cases of repairing two spare cells with columns in the database. Fig. 9 represents all the possibilities for repairing defects with two spare cells. We extract the correlation value after obtaining average, covariance, dispersion by comparing each fail type stored in the database as shown in Fig. 9 with the fail bit map's fail type obtained through a test on a controlled device. If there are more than three spare cells in the comparison above, it will be impossible to repair the defects. Therefore, we falsely classify the repairable flag not to go on with RA analysis and save time.

Fig. 8 show how the new RA analysis translates object into the conventional RA analysis when the values are less than the correlation limit set during the correlation process and therefore cannot be repaired. From a different point of view, if the values exceed the correlation limit, or if they have high similarity, the conventional
RA analysis will translate into the new analysis simulation and extract the repairable minimum correlate limit. At first, the initial value of the correlate limit is 0.0. However, after calculating the average, covariance, dispersion, and correlation value, and confirming whether these values can be repaired, we can obtain the optimal correlate limit, which is repairable and has the smallest value in the correlate limit. The reason why we search for the smallest value in the correlate limit is that in this way we can set up a similarity standard able to repair as many fail types as possible. If we set up a standard with high similarity, or a standard with a correlation value near 1, many fail types that are less similar but repairable will be discarded.

3.4 Process of RA Simulation

- Store information of all the possibilities for repair with two spare cells in the database.
- Load files containing fail bit map information obtained from the previous test.
- Begin the simulation by starting the correlate process for each fail type stored in the database and the fail type of files containing fail bit map information.

Fig. 8. Correlate's flow in details
3.5 Result of the Simulation

To analyze the result of the simulation, use buttons on the top right corner of the screen, x2, One1, Two2, Three3, and Four4, and when you choose each map, the program shows the correlate result of the map on the bottom right. If the values are repairable, the result will be green, and if unrepairable, the result will be yellow. In addition, the color grade on the top right side represents the number of fail bits and fail types in each location. Red (level one) means 8bits failed in LSB, and green (level seven) means 8bits failed only in MSB. Black (level 15) means 4x8 bits failed, i.e., 32 bits failed.

The simulation results in Fig. 10 show that the map on the bottom left side is not only repairable but also has the minimum correlation value (0.636) among the four fail bit maps. This means that the simulation has set the optimal correlate limit for repairing values.

According to the simulation results, repairable values similar to the fail type stored in the database as in Fig. 9 are 1.0 (first value), 0.651 (fifth), 0.594 (sixth), 0.581 (seventh), and 0.609 (eighth). Among these values, the minimum correlate limit is 0.581 (seventh). The reason why the first correlate value became 1.0 is that the value coincided with one of the 495 fail types stored in the database as in Fig. 9.

4 Conclusion

After the simulation, if we apply the minimal correlate value that can be repaired to EDS redundancy analysis, we can reduce RA time to almost 0. The conventional RA algorithm analyzes every fail type to calculate the optimal RA result, but the RA simulation stores those fail types in the database and uses them in calculating the optimal RA result with the highest similarity.
Fig. 10. Results of the simulation

As for the conventional RA process, it was impossible to know whether a defect could be repaired before the result of the main cell's type analysis was obtained. However, in the RA simulation, a database of each fail type analysis is already in place. Therefore by correlating the data with the fail type we can reduce the whole process of the analysis. In EDS redundancy analysis, time spent in tests and in the RA process is directly connected to cost. Due to the technological developments in the semiconductor industry, memory volume is increasing and the unit price per volume is decreasing. As bigger volume means more various fail types, The RA simulation will be an effective alternative to save time in the RA process.

References

[1] Memories and redundancy techniques in IEEE ISSCC Dig. Tech. Papers, pp. 80-87, 1981
[2] J. R. Day, A fault-driven comprehensive redundancy algorithm for repair of dynamic RAMs IEEE Design & Test, vol. 2, no. 3, pp.33-44, 1985
[3] S-Y. Kuo and W. K. Fuchs, Efficient spare allocation in reconfigurable arrays, IEEE Design & Test, vol. 4, pp. 24-31, 1987.
[4] M. Tarr, D. Boudreau, and R. Murphy, “Defect Analysis System Speeds Test and Repair of Redundant Memories,” Electronics, p.175, Jan. 12, 1984.
[5] John R. Day, “A Fault-Driven Comprehensive Redundancy Algorithm,” IEEE Design & Test, vol. 2, No. 3, pp. 35-44, 1985.
[6] Fabrizio Lombardi and W. K. Huang, “Approaches for the repair of VLSI/WSI RRAMs by Row/Column deletion,” Proc. 18th Int. Symp. on Fault-Tolerant Computing, pp. 342-347, July 1988.
[7] Wei Kang Huang, Yi Nan Shen, Fabtrizio Lombardi, “New Approaches for the Repairs of Memories with Redundancy by Row/Column Deletion for Yield Enhancement”, Transactions on Computer-Aided Design, Vol. 9, no. 3, march 1990.
[8] Averill M. Law, W.David Kelton, “Simulation Modeling and Analysis”, Third Edition, Chap4. pp. 235-260, 2000.
A Toolkit for Constructing Virtual Instruments for Augmenting User Interactions and Activities in a Virtual Environment

Kyoung S. Park¹ and Yongjoo Cho²

¹ ICU Digital Media Lab, 517-10 Dogok-dong, Gangnam-gu, Seoul, Korea
park@icu.ac.kr

² Sangmyung University, 7 Hongji-dong, Jongno-gu, Seoul, Korea
ycho@smu.ac.kr

Abstract. This paper presents the design and implementation of Virtual Instrument Scripting Toolkit (VIST), a toolkit for constructing virtual instruments. Virtual instruments are handheld computer applications that help probing environmental properties and augmenting user interactions within a virtual environment. The toolkit, VIST, is intended to simplify the creation of components of virtual instruments and to make it possible to quickly prototype a virtual instrument from those components. VIST also provides a mechanism that allows detecting and displaying attributes of virtual objects and environmental properties in a simulated virtual environment. This paper describes the design rationales and architecture that VIST provides, and demonstrates how VIST has been used to build virtual instruments.

1 Introduction

In the real world, a modern instrument consists of hardware probes or sensors, proprietary analog or digital displays, and specialized embedded processors and software that manage the sub-parts of instruments. Some instruments may have control units such as knobs and buttons that can be manipulated by users to tweak the settings of the instruments. Virtual instruments refer to a kind of instrumentation devices that use standard computers and displays as the replacements of some sub-units of the traditional instruments [2]. In this paper, simulated virtual instruments (hereafter simply virtual instruments) refer to handheld applications that are designed to augment user’s interactions and activities within a virtual world. Many virtual instruments are designed to provide simple 2D graphical user interfaces to help users’ interactions within a virtual environment, such as a data collection tool or an information viewer. Some other instruments are also used as simple measuring devices that report the attributes of a virtual environment, such as a thermometer or a pressure gauge. Moreover, a virtual instrument may be designed to provide the combined features of the above devices.

For the past few years, we have developed a series of virtual learning environments to support science inquiry activities for elementary school students, called
virtual ambients [4]. In virtual ambient environments, students are encouraged to explore and observe the phenomena occurring within a virtual world and investigate the underlying causal simulation models of the observed phenomena. In these environments, young learners are prohibited to control the underlying simulation and are limited to familiar concepts and activities: moving around, seeing things at different scales, and imagining the past and future. While they cannot conduct experiments that require manipulation of independent variables, they can still explore issues of causality by conducting “natural experiments” by finding instances of various preconditions in space or time.

Virtual instruments have been used as a means of enhancing learners’ human senses and investigation skills in virtual ambients. Several studies of using virtual instruments showed that the instruments helped students improve their scientific analysis, specially the coverage rate of virtual worlds and data collection rate [1]. Virtual instruments also enable the construction of data-rich virtual ambients. The designers can create worlds in which objects, and the environment itself, have embedded properties—such as the iridium concentration of Martian rocks or the salinity of soil in a virtual field—which form the grist for student investigations. Virtual instruments can be used to examine the attributes of virtual objects or environmental factors to help learners investigate the underlying simulation models of virtual ambients.

While virtual instruments are the essential parts of virtual ambients, the development of virtual instruments is difficult due to required knowledge of handheld programming, especially for novice programmers. This paper describes the design and implementation of a virtual instrument scripting toolkit (VIST) that employs lightweight component architecture and scripting mechanism to enable rapid prototyping of instrumental handheld applications. The primary design goal was to increase the reusability of existing components while decreasing the complexity of programming in the construction of virtual instruments. This paper describes the design issues of virtual instruments and detail descriptions about the architecture of virtual instrument scripting toolkit. Then, it describes the construction of virtual instruments using VIST and the plans for future research direction.

2 Design and Architecture of Virtual Instrument Scripting Toolkits

Virtual instruments can be used in either active or passive mode. In the Virtual Harlem [5], for example, a virtual instrument was used to show the locations of all virtual annotations—users’ recorded voice and gestures in the virtual world—on a 2D map view. This instrument was used to trigger playing an annotation in the virtual world. Virtual instruments can be used to detect and measure both visible and invisible attributes of virtual objects or environmental properties. For instance, students can retrieve the salinity and moisture values in the soils of a virtual field [3]. Based on our experience of building several virtual instruments, we have identified five design issues as follows.
- **Customization.** A virtual instrument can show different representations specially customized to learner’s needs. Simple nominal values can be represented by illustrations (e.g. smiles or frowns), audio (e.g. Geiger counters) or animation (e.g. navigator’s position/orientation on a map) in lieu of numeric readouts. It is also possible to show the same property value with multiple representations. For example, a virtual global position system may show the same positional information with both numeric values (e.g. latitude and longitude) and animation (e.g. “you are here” map).

- **Flexibility.** A virtual instrument should be flexible to support various users’ needs. For instance, a virtual instrument should show increased or decreased precision, thereby providing any precision suitable for users. Young learners are not familiar with decimal or scientific notation, and hence increased precision may lead them confused and obscure more central learning goals. The same is true with the choice of units - e.g. 12 kilometers is much more accessible than 12000 meters for the second graders.

- **Lightweight.** Due to the limited resources (such as computing power, memory, and storage space) of handheld devices, a virtual instrument should be lightweight to minimize the complexity of constructing new components. That is, each component processes simple tasks, and then components can be combined to create a new instrument for more complex task.

- **Transparency.** A virtual instrument should be transparent from virtual worlds. The instrument built for one virtual world should be operated in the same way in other virtual worlds as long as the world supports such instrumentation. For instance, a virtual thermometer should measure temperature in any virtual environments that had temperature property.

- **Interaction.** A virtual instrument can be used as an interface device. First-person view virtual environments are typically weak on object identification and manipulation. Many virtual environments provided a light “ray” for easy interaction with objects, which still needed considerable dexterity and practice. A virtual instrument can be used to simplify such object selection. For instance, a navigator may use a virtual instrument that provides graphical user interface of listing all selectable virtual objects. Then, he/she may interact with an object by simple selection of an item on the instrument rather than using a three-dimensional input device in the virtual world.

The specific nature and design guidelines of virtual instruments led us to develop the virtual instrument scripting toolkit (VIST). VIST was designed to enable novice programmers to easily prototype virtual instruments for augmenting user interactions and activities in a virtual environment. The primary goal was to increase reusability of existing graphical user interface (GUI) and instrument components while decreasing the complexity of programming for virtual instrument and components.

VIST uses a scripting mechanism to address customization and flexibility. An instrument application is constructed by reading a user specified script-based configuration file that will load instrumental or GUI components dynamically at run-time. The configuration file is used to layout all components displayed on a virtual instrument.
It also includes invisible components that process background computations and control of underlying hardware (e.g. timer), input/output (I/O) operations (e.g. file I/O handler) and network communication. The designers can specify event triggering and message passing in the configuration file to combine multiple components together. A new component can be added by specifying a few lines of script code in the configuration file. A component can be removed from the instrument device by commenting out or deleting the script code specific to the component in the configuration file.

VIST also provides a way to specify certain script codes to be executed when a special event occurs in a component. An event is generated in a component and passed to the script level. If there is a script that handles the component’s event, the event handler written in the script will be executed. The designers can add event handlers that would be executed when a specific event happens in a component. Lines 10-12 in Fig. 1 show a quick example of an event handler that responds to the “OnClick” event, which happens when the button is clicked.

VIST provides a message passing mechanism as a means of executing certain code defined in a component. If the component knows how to handle the given message, it processes the message and may return the result of the execution. For example, line 11 in Fig. 1 shows an example of passing a message to a component. When the button is clicked, the component named “label” gets a message called “update” with a text string as a parameter. Although it is not shown in the picture, “label” is an instance of the “Label” component that supports the “update” message; the “Label” component updates the bounded widget with the given text string when the “update” message is received through the Message interface.

VIST is designed to minimize the requirement of the system platforms to address lightweight component architecture. For instance, it utilizes the features of underlying operating system wherever possible instead of building its own architecture, which often adds another layer and hence requires more processing power. VIST’s dynamic loading mechanism is implemented on top of the modern operating system's dynamic module loading mechanism—dynamic link library on Microsoft Windows and shared library on Unix-like systems. Instead of using well-known but complex component architecture, such as CORBA or Java Beans, VIST employs its own simple

![Fig. 1. CLOVES instrument script showing component, event handler](image)
component architecture to minimize the overhead of implementation and execution of component architecture on a handheld computer. In the current version of VIST, a component is implemented by using four abstract interfaces: CreateComponent, Destroy, Initialize, and Message. Since VIST already provides the default implementations of three methods, designers only need to write an interface method, Message, which processes any message received from other components or instrument scripts.

VIST attempts to decouple virtual instruments from the virtual world to minimize coherence while increasing transparency. VIST makes the virtual instrument work similar to the way a real instrument does. In the real world, users can use an instrument to measure the appropriate environment properties specific to the instrument. If the appropriate properties were available in the environment, the instrument would be able to show or process the properties. Otherwise, it would fail to report any information. As shown in Fig. 2, when a virtual instrument is started, it registers the name of environmental properties that the instrument would process to the simulated virtual world. Then, whenever the registered properties are updated in the simulation, the virtual instrument would receive the updated values. If the virtual world does not have the registered properties, nothing will be happened in the virtual instrument as real instruments do in the real world. This approach would increase transparency of virtual instruments while easing usability of the instrument devices.

3 The Use of VIST for Constructing Virtual Instruments

Over the past few years, we have developed a number of virtual instruments to scaffold children’s scientific investigation and data collection in virtual ambient environments [1]. Initially, the virtual instruments were implemented using C++ and Microsoft Pocket PC application programming interfaces (APIs), but later they have been re-written using VIST for easy customizability, usability, and maintenance. VIST reduced the complexity of constructing virtual instruments for a virtual world. For instance, the instrument designed for the virtual field initially took several months to develop since the application specifications evolved and programmers worked with low-level tools (e.g. C++, native Pocket PC APIs, and socket programming for
network communication). However, by using VIST, the developers could build this instrument within a week.

Fig. 3 shows an example of virtual instrument constructed using VIST, called FieldInstrument, in a virtual field [4]. FieldInstrument consists of several visual components, such as simulated global position system (GPS), button, and list widgets. The simulated GPS shows the navigator’s position and orientation in the virtual world on a two-dimensional instrument device. FieldInstrument also has invisible components that work underneath the instrument’s visual interface to handle the communication between the instrument and the virtual field over the network. This network communication component generates “OnReceived” event whenever it receives data from the virtual world over the network. The generated event is passed to VIST’s message handler, which basically updates the GPS module of the instrument. Button modules of FieldInstrument may generate “OnClick” event whenever a user clicks on the image button. Responding to the event, VIST’s event handler method retrieves current user’s position from the GPS module and updates ListWidget component with the position and the types of data that the user collects.

Another virtual instrument was developed for a virtual ocean environment by a naive programmer who had no previous experience with handheld computer programming. The instrument displayed several environmental attributes of the ocean such as pressure, temperature, salinity, and light penetration based on the navigator’s depth in the ocean. Later, this instrument application was divided into two separate instruments: one device showed the navigator’s depth, pressure, and temperature, whereas the other displayed the depth, salinity, and light penetration values. Separating the instrument into two devices was easily done by separating such components in the configuration file. Simple changes, such as adding or removing instrument properties, would have required knowledge of the underlying APIs of the operating system as well as the interfaces of the virtual instruments and the virtual environments when it was written in C++ and native APIs.
4 Conclusion

This paper describes the design and implementation of Virtual Instrument Scripting Toolkit (VIST), a component-based toolkit for constructing virtual instruments. Virtual instruments are handheld computer applications, which run on a Pocket PC, developed to support user interactions and activities in a virtual environment. For the past few years, we have developed a number of virtual instruments to help young children improve their scientific inquiry skills in virtual ambient environments. While virtual instruments are the essential parts of virtual ambients, the development of such instruments requires an understanding of low-level programming and native APIs of handheld computers. Based on our experience of developing and deploying these applications for elementary school science education, we have identified design issues for virtual instruments and proposed a new lightweight component-based scripting toolkit, which addresses these design requirements.

While previous work for handheld applications are focused on helping user “interaction” in a virtual environment, our toolkit puts an emphasis on five design guidelines: customization, flexibility, lightweight, abstraction, and interaction. VIST is intended to ease the construction of virtual instruments for retrieving and displaying environmental properties from a virtual world. It can also be used to augment users’ interactions in a virtual world. This requires identifying some interaction schemes in virtual environments to implement them as components. We have currently re-written virtual instruments using VIST, and more instrument components are added to VIST. More research and work are needed to evaluate the usability of the toolkit.

References

1. Cho. Y., Moher, T., and Johnson, A.: Scaffolding Children’s Scientific Data Collection in a Virtual Field. In Proceedings of International Conference on Virtual Systems and Multimedia (2003), 558-564.
2. Goldberg, H: What is Virtual Instruments? IEEE Instrumentation and Measurement Magazine, December, 1994, 10-13.
3. Johnson, A., Moher T., Cho. Y., Edelson, D., and Reiser, B.: Sixth Graders doing science - Collecting Data and Looking for Patterns in a Virtual Field, In Proceedings of IEEE Virtual Reality (2002), 281-283.
4. Moher, T., Johnson, A., Cho, Y., and Lin, Y.: Observation-based Inquiry in a Virtual Ambient Environment. In Proceedings of International Conference of the Learning Sciences (2000), 238-245.
5. Park, K. Leigh, J., Johnson, A., Carter, B., Brody, J., and Sosnoski, J.: Distance Learning Classroom using Virtual Harlem. In Proceedings of IEEE International Conference on Virtual Systems and Multimedia (2001) 489-498.
Mobility Grouping Scheme to Reduce HLR Traffic in IMT-2000 Networks

Dong Chun Lee, Gwang-Hyun Kim, and Seung-Jae Yoo

1 Dept. of Computer Science Howon Univ., Korea
1dch@sunny.howon.ac.kr
2 Division of CE & CE Gwangju Univ., Korea
3 Dept. of Information Security, Joongbu Univ., Korea

Abstract. This paper proposes the registration grouping scheme that solves the Home Location Register (HLR) bottleneck due to the terminal's frequent Registration Area (RA) crossings and that distributes the registration traffic to each of the local signaling transfer point (LSTP) area.

1 Introduction

The mobility management schemes are based on Interim Standard-95 (IS-95) and Global System for Mobile Communication (GSM) standard. Those standards use the two-tier database system of HLR and visitor location register (VLR). Whenever a terminal crosses a RA, HLR should be updated. Frequent DB accesses and message transfers may cause the HLR bottleneck and degrade the system performance. In this paper, we define the LSTP area as the group of RAs of which the serving VLRs is connected to LSTP. Those RAs are grouped statically not dynamically. We will explain the reason later. The proposed scheme compares the performance to those in IS-95 scheme, Local Anchor (LA) scheme.

2 Proposed Scheme

We define post VLR, PVLR which keeps the callee's current location information as long as the callee moves within its LSTP area. If a terminal crosses the LSTP area, the VLR which serves the new RA is set to a new PVLR. If the terminal moves within the LSTP area, it is registered at its own PVLR not HLR. If the terminal moves out from the area, it is registered at the HLR. In case that a terminal is switched on, the VLR which serves the terminal's current RA is PVLR and the VLRs which serve the intermediate RAs in terminal's moving route within the LSTP area report the terminal's location information to the PVLR. We note that we don't have to consider where the callee is currently. It is because the PVLR keeps the callee's current location as long as the callee moves within its LSTP area. Therefore, without the terminal movements into a new LSTP area, the registration at HLR does not occur.
We statically group the VLRs in LSTP area in order to localize the HLR traffic. It is also possible to group the RAs dynamically regardless of the LSTP area. Suppose that the PVLR and the VLR which serves the callee’s RA belong to the same dynamic group but are connected to the physically different LSTPs. In this case, we should tolerate the additional signaling traffic even though the caller and callee belong to the same dynamic group. A lot of signaling messages for registering user locations and tracking calls are transmitted via RSTP instead of LSTP. If the cost of transmitting the signaling messages via RSTP is large enough compared to that via LSTP, dynamic grouping method may degrade the performance although it solves the ping-pong effect. Fig. 1 shows the message flow due to the location registration according to the status of PVLR change.

Fig. 1. Location registration in RA grouping scheme

3 Performance Analysis

For numerical analysis, terminal-moving probability should be computed. We adopt hexagon model as geometrical RA model. That model is considered to be common for modeling the RA. Generally, it is assumed that a VLR serves a RA.

As shown in Fig. 2, RAs in LSTP area can be grouped. There are 1, 7, and 19 RAs in circle 0, circle 1, and circle 2 areas, respectively. The terminals in RAs inside circle n area still exist in circle n area after their first RA crossing. While the terminals in RAs which meet the line of the circle in figure can move out from their LSTP area. We can compute the number of terminals which move out from the LSTP area as follows.

\[
\frac{\text{Total No. of outside edges in arrow marked polygons}}{\text{No. of edges of hexagon} \times \text{No. of RAs in LSTP area}} \times \text{No. of terminals in LSTP area}.
\]

For example, there are 2 or 3 outside edges in arrow marked area in hexagon model. So 2/6 or 3/6 of terminals in corresponding RA which meets the line of the
circle move out the LSTP area. In case of circle 2 in Fig. 2, the number of RAs is 19 and the number of terminals which move out from LSTP area is given by the terminal in LSTP area $5/19$. The number of VLRs in LSTP area represented as circle n can be generalized as follows.

$$\text{No. of VLRs in LSTP area} = 1 + 3n(n+1) \quad (\text{where } n = 1, 2\ldots)$$

The rate of terminals which move out from LSTP area can be generalized as follows.

$$R_{\text{move_out, No. of VLRs in LSTP area}} = \frac{2n+1}{1+3n(n+1)} \quad (\text{where } n = 1, 2\ldots)$$

RA is said to be locally related when they belong to the same LSTP area and remotely related when the one of them belong to the different LSTP area. The terminal's RA crossings should be classified according to the local and remote relations in the following schemes.

1. LA scheme
   - Relation of LA serving RA and callee's last visiting RA
   - Relation of callee's last visiting RA and callee's current RA

2. The proposed scheme
   - Relation of PVLR serving RA and callee's last visiting RA
   - Relation of RA where callee's last visiting RA and callee's current RA

We define the probabilities that a terminal moves within the LSTP area and crosses the LSTP area as $P(\text{local}) = R_{\text{move_in, no. of VLRs in LSTP area}}$ and $P(\text{remote}) = R_{\text{move_out, no. of VLRs in LSTP area}}$, respectively. Suppose that the number of RAs in LSTP area is 7. If a terminal moves into a new LSTP area in its $n^{th}$ movement, the terminal is located in one of outside RAs - 6 RAs - in the new LSTP area. If the terminal moves into a new RA in its $(n + 1)^{th}$ movement, $R_{\text{move_in, no. of VLRs in LSTP area}}$ and $R_{\text{move_out, no. of VLRs in LSTP area}}$ are both 3/6. If the terminal's $(n + 1)^{th}$ movement occurs
within the LSTP area, the two probabilities according to its \((n + 2)^{th}\) movement, \(4/7\) and \(3/7\) are respectively. To evaluate the performance, we define the signaling costs (SCs) as follows.

- SC1: Cost of transmitting a message from one VLR to another VLR through HLR
- SC2: Cost of transmitting a message from one VLR to another VLR through RSTP
- SC3: Cost of transmitting a message from one VLR to another VLR through LSTP

We evaluate the performance according to the relative values of SC1, SC2, and SC3 which are needed for location registration. To evaluate the registration traffic, we define registration cost set (RCS). RCS is composed of the SC1, SC2, and SC3. SC2 may not belong to the registration cost according to the applied schemes. In IS-41 scheme, only SC1 is used. We assume the various RCSs, SC3 < SC2 < SC1 (Case 1), SC3, SC2 << SC1 (Case 2), SC3 << SC2, SC1 (Case3), and SC3, SC2 < SC1 (Case 4).

In following figures, we assume that the larger RCS number is, SC2 and SC3 are larger and SC3/SC2 is larger in above 4 cases. It is to investigate how much SC2 and SC3 affect the registration cost. Because SC1 is fixed regardless of the RCS number, the traffic dependency on SC1 is reduced with larger RCS number. Using the RCS, we can compare the registration costs in several cases with that in IS-95 scheme.

![Fig. 3. RCS with 7VLRs in LSTP](image-url)
If a terminal moves within the LSTP area - in case of the local relation -, the PVLR is updated and the entry in old VLR is cancelled. If a terminal moves into a new LSTP area, the HLR is updated. Subsequently, the entries in old PVLR and old VLR are cancelled. The registration cost is computed as follows.

\[ C_{\text{static Reg. Grouping, Loc.Reg}} = \frac{R_{\text{move_in, no. of VLRs in LSTP area}}}{2^{\cdot}2S_{3}} + \frac{R_{\text{move_out, no. of VLRs in LSTP area}}}{2(S_{1} + S_{3})} \]

In Fig.3, we assume that SC3 in case 4 - SC3, SC2 < SC1 - is larger than those in the other cases. The registration traffic dependency on SC1 in proposed scheme is smaller than that in IS-41 scheme. Therefore, with a common SC1, case 1, 2, and 3 shows the better results compared to case 4. Registration cost is reduced with larger number of the VLRs in LSTP area. In case that the average speed of the terminal is fixed and \( R_{\text{move_out, no. of VLRs in LSTP area}} \) is lessen, the HLR access cost is reduced.

In IS-95 scheme, whenever a terminal moves into a new RA, it is registered at the HLR. The registration cost is computed as follows: \( C_{IS-41, \text{Loc.Reg}} = 2 \cdot SC_{1} \).

\[ \text{Fig. 4. Registration cost comparison} \]

In LA scheme, the local and remote relations between the RA into which a terminal moves and LA serving RA should be considered. When a terminal moves into a new RA, the LA is updated or a new LA is selected, and subsequently the entry in old VLR is cancelled. Suppose that a call is requested to a terminal after terminal's n time movements. We define the number of terminal's RA crossings before call request as \( M_{\text{no.}} \). The registration cost can be written as follows.
\[
C_{\text{LA,Loc.Reg}} = (1 - \frac{1}{(M_{\text{no.}})}) \left\{ 2 \cdot 2SC \cdot P(\text{local}) + 2 \cdot 2SC2 \cdot P(\text{remote}) \right\} + \left( \frac{1}{(M_{\text{no.}})^2} \right) \left\{ 2(SC1 + SC3) \cdot P(\text{local}) + 2(SC1+SC2) \cdot P(\text{remote}) \right\} 
\]

In Fig. 4, the proposed scheme shows the better results compared to the LA scheme in case of \( (1 / (M_{\text{no.}}) \geq 0.5 \) where \( M_{\text{no.}} \) is the number of terminal's RA crossings before call request. The smaller \( (1 / (M_{\text{no.}}) \) is in LA scheme, the more VLRs in LSTP area can be taken for the similar performance. This implies that proposed scheme with large number of VLRs shows the similar performance to LA scheme in high mobility environments.

4 Conclusions

In this letter we presented the static registration grouping scheme that is mainly focused on reducing the HLR bottleneck traffic and distributing the HLR traffic to each of the LSTP area. The proposed scheme decreases the additional signaling traffic compared to the dynamic one, and is relatively insensitive to the user's call-to-mobility ratio (CMR), compared to other schemes. As a result of cost evaluation, the more VLRs in LSTP area are, the performance is improved greater than other schemes.

Acknowledgement

This work was supported (in part) by the Ministry of Information & Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

[1] Andrew D. Malyan, Leslie J. Ng, Victor C.M. Leung, and Robert W. Donaldson, “Network Architecture and Signaling for Wireless Personal Communications,” IEEE JSAC., Vol. 11, No. 6, pp. 830-840, August 1998.
[2] EIA/TIA, “Cellular Radio telecommunications Intersystem Operations: Automatic Roaming,” Technical Report IS-95 (Revision A), EIA/TIA, July 1997.
[3] G.P. Pollini, “Signaling Traffic Volume Generated by Mobile and Personal Communications,” IEEE Comm. Mag., Vol.33 No. 6, pp. 60-65, June 1997.
[4] J.S.M. Ho and I.F. Akyildiz, “Local Anchor scheme for Reducing Location Tracking Cost in PCNs,” ACM MOBICOM’99, 1999.
Security Requirements for Software Development

Tai-hoon Kim¹, Myong-chul Shin², Sang-ho Kim¹, and Jae Sang Cha³

¹KISA, 78, Garak-Dong, Songpa-Gu, Seoul, Korea
{taihoon, shkim}@kisa.or.kr
http://www.kisa.or.kr

²School of Electrical and Computer Engineering, Sungkyunkwan University, Korea

³Dept. of Information and Communication Eng. SeoKyeong University, Korea

Abstract. This paper proposes a security engineering based approach considering security when developing software. If the cases security requirements are not considered, the use of security products (as an example, firewall or intrusion detection system) may be not the secure solution. Therefore, when making some kinds of software products, security-related requirements must be considered.

1 Introduction

The focus of security engineering has expanded from one primarily concerned with safeguarding classified government data to broader applications including financial transactions, personal information, and the Internet. These trends have elevated the importance of security engineering [1].

When we are making some kinds of software products, ISO/IEC TR 15504 may provide a framework for the assessment of software processes, and this framework can be used by organizations involved in planning, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software. But, in the ISO/IEC TR 15504, considerations for security are relatively poor to other security-related criteria such as ISO/IEC 21827, the Systems Security Engineering Capability Maturity Model (SSE-CMM), or ISO/IEC 15408, Common Criteria (CC) [2-6]. Security-related software development is concerned with many kinds of measures that may be applied to the development environment or developer to protect the confidentiality and integrity of the IT product or system developed. The customer’s requirements must be implemented to software perfectly, but this is not sufficient. The secure software may be implemented by not only applying Firewall or IDS but also considering security requirement appended to customer’s requirement.

This paper proposes a concept of security requirement appended to customer’s requirement.

2 Scope of Security Engineering

In fact, the scope of security engineering is very wide and encompasses:

- the security engineering activities for a secure software or a trusted system addressing the complete lifecycle of: concept definition, analysis of customer’s re-
requirements, high level design and low level design, development, integration, installation and generation, operation, maintenance and decommissioning;
• requirements for product developers, secure systems developers and integrators, organizations that develop software and provide computer security services and computer security engineering;
• applies to all types and sizes of security engineering organizations from commercial to government and the academe.

The security engineering should not be practiced in isolation from other engineering disciplines. Maybe the security engineering promotes such integration, taking the view that security is pervasive across all engineering disciplines (e.g., systems, software and hardware) and defining components of the model to address such concerns.

The field of security engineering has several generally accepted principles, but it currently lacks a comprehensive framework for evaluating security engineering practices. It must be stressed that security engineering is a unique discipline, requiring unique knowledge, skills, and processes which warrants the development of a distinct CMM for security engineering. This does not conflict with the premise that security engineering is done in context with systems engineering. In fact, having well-defined and accepted systems engineering activities will allow security engineering to be practiced effectively in all contexts.

3 Application of Security Engineering

A wide variety of organizations can apply security engineering to their work such as the development of computer programs, software and middleware of applications programs or the security policy of organizations. Appropriate approaches or methods and practices are therefore required by product developers, service providers, system integrators, system administrators, and even security specialists. Some of these organizations deal with high-level issues (e.g., ones dealing with operational use or system architecture), others focus on low-level issues, and some do both.

Software or product developers may use security engineering to gain an understanding of the customer’s security needs and append security requirements to the customer’s requirements. Interaction with the customer is required to ascertain them. In the case of a product, the customer is generic as the product is developed a priori independent of a specific customer. When this is the case, the product marketing group or another group can be used as the hypothetical customer, if one is required.

The main objective of application of security engineering is to provide assurance about the software or system to customer, and the assurance level of software or system may be the critical factor has influence on deciding purchase. Therefore, the meaning of the application of security engineering to the software is the application of some assurance methods to the software development lifecycle phases.
4 Application of Security Engineering to Software Development

There are many methodologies for software development, and security engineering does not mandate any specific development methodology or life cycle model. Fig. 1 depicts underlying assumptions about the relationship between the customer’s requirements and the implementation. The figure is used to provide a context for discussion and should not be construed as advocating a preference for one methodology (e.g. waterfall) over another (e.g. prototyping).

![Diagram](image)

**Fig. 1.** The relationship between the customer’s requirements and the implementation

It is essential that the requirements imposed on the software be effective in contributing to the objectives of consumers. Unless suitable requirements are established at the start of the development process, the resulting end product, however well engineered, may not meet the objectives of its anticipated consumers.

The process is based on the refinement of the customer’s requirements into a software implementation. Each lower level of refinement represents design decomposition with additional design detail. The least abstract representation is the software implementation itself.

4.1 Security-Related Requirements

For the development of software, the first objective is the perfect implementation of customer’s requirements. And this work may be done by very simple processes. However, if the software developed has some critical security holes, the whole network or systems that software installed and generated are very vulnerable.

Therefore, developers or analyzers must consider some security-related factors and append a few security-related requirements to the customer’s requirements. Fig. 2 depicts the idea about this concept.

The processes based on the refinement of the security-related requirements are considered with the processes of software implementation.
4.2 Implementation of Appended Security-Related Requirements

Developers can reference the ISO/IEC 15408, Common Criteria (CC), to implement security-related requirements appended.

The multipart standard ISO/IEC 15408 defines criteria, which for historical and continuity purposes are referred to herein as the CC, to be used as the basis for evaluation of security properties of IT products and systems. By establishing such a common criteria base, the results of an IT security evaluation will be meaningful to a wider audience.

The CC will permit comparability between the results of independent security evaluations. It does so by providing a common set of requirements for the security functions of IT products and systems and for assurance measures applied to them during a security evaluation. The evaluation process establishes a level of confidence that the security functions of such products and systems and the assurance measures applied to them meet these requirements. The evaluation results may help consumers to determine whether the IT product or system is secure enough for their intended application and whether the security risks implicit in its use are tolerable.

In support of the three parts of the CC listed above, it is anticipated that other types of documents will be published, including technical rationale material and guidance documents.

4.3 Implementation Example of Appended Security-Related Requirements

For example, if the customer’s requirement ‘management by web’ is specified, we can append some security-related requirements:

1. Customer’s requirement: We want to manage the program by web
2. Appended requirements: Next lists are not perfect but general to all cases
– Requirements for identify mechanisms.
– Requirements for authentication mechanisms.
– Requirements for Cryptographic operations (Key generation and so on)
– Requirements for the SF to provide the capability for SF-initiated and user initiated locking and unlocking of interactive sessions.
– Requirements for the creation of a trusted channel between the SF and other trusted IT products for the performance of security critical operations.
– Requirements for recording the occurrence of security relevant events that take place under SF control.
– Etc.

(3) Grouping

| Grouping                  | Requirements                                                                 |
|---------------------------|------------------------------------------------------------------------------|
| IA (Identification and   | – Requirements for identify mechanisms.                                      |
| Authentication)           | – Requirements for authentication mechanisms.                                 |
| TA (TOE Access)           | – Requirements for the SF to provide the capability for SF-initiated and user |
|                           | initiated locking and unlocking of interactive sessions.                     |
| CS (Cryptography Support) | – Requirements for Cryptographic operations (Key generation and so on)       |
| TP (Trusted path/channels)| – Requirements for the creation of a trusted channel between the SF and other |
|                           | trusted IT products for the performance of security critical operations.     |
| AU (Audit Review)         | – Requirements for recording the occurrence of security relevant events that |
|                           | take place under SF control.                                                |

(4) Selection of security family in the class

| Security class | Mapping to security family                                                                |
|----------------|------------------------------------------------------------------------------------------|
| FIA            | User authentication (FIA_UAU),                                                          |
|                | User identification (FIA_UID),                                                           |
| FAU            | Security audit data generation (FAU_GEN),                                               |
|                | Security audit review (FAU_SAR),                                                         |
| FCS            | Cryptographic key management (FCS_CKM),                                                 |
|                | Cryptographic operation (FCS_COP),                                                       |
| FTA            | Session locking (FTA_SSL),                                                               |
| FTP            | Inter-TSF trusted channel (FTP_ITC),                                                     |
(5) Selection of security component in the family

| Security family | Mapping to security component |
|-----------------|-------------------------------|
| FIA_UAU, FIA_UID | FIA_UAU.2 User authentication before any action, FIA_UID.2 User identification before any action |
| FAU_GEN, FAU_SAR | FAU_GEN.2 User identity association, FAU_SAR.2 Restricted audit review, |
| FCS_CKM, FCS_COP | FCS_CKM.1 Cryptographic key generation, FCS_CKM.2 Cryptographic key distribution, FCS_CKM.4 Cryptographic key destruction, FCS_COP.1 Cryptographic operation |
| FTA_SSL | FTA_MCS.2 Per user attribute limitation on multiple concurrent sessions, FTA_SSL.1 TSF-initiated session locking, FTA_SSL.2 User-initiated locking, FTA_TSE.1 TOE session establishment |
| FTP_ITC | FTP_ITC.1 Inter-TSF trusted channel |

(6) Implementation of requirements to software, hardware or firmware.

5 Conclusions

This paper proposed significance of security-related requirements appended to the customer’s requirements. For the development of software, the first objective is the perfect implementation of customer’s requirements. However, if the software developed has some critical security holes, the whole network or systems that software installed and generated may be very vulnerable.

Therefore, developers or analyzers must consider some security-related factors and append a few security-related requirements to the customer’s requirements. The processes based on the refinement of the security-related requirements are considered with the processes of software implementation.

References

1. ISO. ISO/IEC 21827 Information technology – Systems Security Engineering Capability Maturity Model (SSE-CMM)
2. ISO. ISO/IEC TR 15504-2:1998 Information technology – Software process assessment – Part 2: A reference model for processes and process capability
3. ISO. ISO/IEC TR 15504-5:1998 Information technology – Software process assessment – Part 5: An assessment model and indicator guidance
4. ISO. ISO/IEC 15408-1:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model
5. ISO. ISO/IEC 15408-2:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 2: Security functional requirements
6. ISO. ISO/IEC 15408-3:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 3: Security assurance requirements
7. Tai-Hoon Kim, Byung-Gyu No, Dong-chun Lee: Threat Description for the PP by Using the Concept of the Assets Protected by TOE, ICCS 2003, LNCS 2660, Part 4, pp. 605-613
8. Tai-Hoon Kim, Haeng-kon Kim: A Relationship between Security Engineering and Security Evaluation, ICCSA 2004, LNCS 3046, Part 4, pp. 717-724
9. Tai-Hoon Kim, Dong Chun Lee: Reduction Method of Threat Phrases by Classifying Assets, ICCSA 2004, LNCS 3043, Part 1, pp. 1052-1059
Intelligent Control Model of Information Appliances

Huey-Ming Lee¹, Ching-Hao Mao¹, and Shu-Yen Lee²

¹Department of Information Management, Chinese Culture University, 55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114) Taiwan
²Dep. of Private Participation in Infrastructures, China Engineering Consultants, Inc., 2nd Fl., No. 1, Ln. 240, Kiang Fu S. Road, Taipei, Taiwan

Abstract. In this study, we proposed an intelligent control model of information appliances (ICMIA). This model not only can collect the related users’ information appliances preference messages automatically, but also generate the IA control rules by the fuzzy neural network learning. Via this model, it can make home network environment more convenient and comfortable.

1 Introduction

During the past decades, information appliances (IAs) grows so fast that the need of control mechanism becomes much more important. The control mechanism of IAs can’t be standardized. It needs to have the capability of self-learning and inference; that way, they can adapt to the changing of home networks and suit users’ needs.

Lee et al. [1] came up with the idea of IAs intelligent agent model (IAIA), making home environments safer and convenient. Lee et al. [5] proposed fuzzy neural network model of information appliances with the functions of self-learning and fuzzy inference, it enables IAIA to maximize efficiency of IAs in a more humane way. Lee and Mao [3] proposed a clustering model of information appliance, it processes user’s recognitions of IAs to cluster IA devices, and it facilitates the management of control. Lee et al. [4] proposed a fuzzy aggregative clustering model of information appliances (FACIA) which is capable to cluster the IAs, filter and extract the IAs’ messages automatically. If there is a mechanism to interaction with environment automatically, then we can simplify the control procedure of IAs.

In this study, we proposed an intelligent control model of information appliances (ICMIA). This model not only can collect the related users’ IAs preference messages automatically but also generate the IA control rules by the fuzzy neural network learning. Via this model, we can have the more convenient and comfortable home network environment.

2 Intelligent Control Model of Information Appliances

In this section, we presented an intelligent control model of information appliances (ICMIA) under the supervision of IAIA [1] as shown in Figure 1. Via the IAs control
sub-model IAIA, ICMIA can exchange the messages of IAs. We not only modify the interaction components of IAIA in FCIA but also enhance the capability of messages interaction in FNNIS. Therefore, there are two sub-models in ICMIA, saying IA clustering and filtering sub-model (IACF) and IA Learning and Inference sub-model (IALI), as shown in Figure 2.

The functions of these sub-models are as the followings:

- **IACF**: It can simplify the chaos messages of IAs and filter the contradictive messages.
- **IALI**: It learns from the messages of IAs, generate fuzzy control rule, and produce a suitable inference to control IAs.

### 2.1 IACF

IACF sub-model can filter, extract the messages of IA, and exchange the message between IA Message Sub-model with IALI automatically. There are three components, saying IA clustering engine (IACE), IA message clustering engine (IAMCE) and interaction management component (IMC) in IACF, as shown in Figure 3.
The functions of these components are as the followings:

- **IACE**: IACE is the core of this model. The users’ recognitions of IA device are clustered through fuzzy cluster.
- **IAMCE**: IAMCE can not only simplify the complicated IA message, but also filter the contradiction values.
- **IMC**: IMC plays the interface among IAIA, IACF and IALI. It can redirect the messages to appropriate components according to functions.

### 2.2 IALI

IALI learns from the messages of IAs, generate fuzzy control rule, and produce a suitable inference to control IAs. Also it can interact with other sub-models via the IAs’ messages. There are two components in IACI, saying IA Learning mechanism, and Fuzzy Inference Engine, as shown in Figure 4.

The functions of these components are as the followings:

- **FNN Learning Mechanism**: This is a fuzzy neural network (FNN) which can learn from the status values produced from IAs and form a fuzzy control rule.
- **Fuzzy Inference Engine**: This sub-model can find the apt rule base based on the IA status value and conduct fuzzy-inference. Its result can be used as a reference in IA control.
- **Rule Base**: It is used to store all the rules produced by fuzzy logic and provide the rules to the fuzzy inference engine for its use.
- **Data Base**: It is used to store all kinds of colloquial variances and their fuzzy membership functions. Membership functions can be modified through fuzzy neural network and provided for model calculations.
3 Implementation

In this section, we illustrate the implementation environment of the model and take the home temperature control as an example to implement.

3.1 Practical Environment

We follow the specification of UPnP and implement six kinds of IA devices. The home temperature controlling environment framework is shown as Figure 5. For the purpose of ease manipulation, cross-platform, and remote-control capability, we have applied Java Server Page (JSP) and Java Servlet written Web Server structure, and Java 2 Platform Standard Edition v 1.4.2 API Specification is utilized for constructing ICMIA.

3.2 Model Implementation

For illustrating the processes of implementation, we take the 499 messages of IAs as examples.
(1) IACF

After the operating of IACF, there are 121 IAs’ message as output, as shown in Table 1.

| Before Clustering | After Clustering |
|-------------------|------------------|
| 499               | 121              |

(2) IALI

- Rule-based generated

When IALI received the 121 IAs’ messages of the output of IACF, then it generated 9 control rules of IAs as rule-base, as shown in Table 2.

| No | Control Rules                                      |
|----|----------------------------------------------------|
| 1  | if tem is low and smk is low then air is high     |
| 2  | if tem is low and smk is high then air is high    |
| 3  | if tem is high and smk is low then air is lowest  |
| 4  | if tem is high and smk is high then air is low    |
| 5  | if tem is high and vol is highest then air is lowest |
| 6  | if tem is low and hum is high then air is highest |
| 7  | if tem is high and hum is high then air is low    |
| 8  | if tem is high and vol is low then air is low     |
| 9  | if tem is high and ray is high then air is lowest |

- Case implementation

(1) If we input the value of T as 17, H as 0.5, S as 0.8, R as 0, V as 10, then we can have the temperature of air_condition increase 4.062498255 degrees.

(2) If we input the value of T as 15, H as 0.54, S as 0.8, R as 0, V as 10, then we can have the temperature of air_condition increase 4.287032967 degrees.

(3) If we input value of T as 35, H as 0.58, S as 0.5, R as 5, V as 68, then we can have the temperature of air_condition decrease 4.949370629 degrees.

As the above statements, ICMIA not only can filter 75.5% of the abnormal or unmeaning messages of IAs without user’s operation but also generate 9 control rules of IAs automatically. Furthermore, this model can maintain home temperature even if the home environment status change constantly.
4 Conclusion

In this study, we proposed an intelligent control model of information appliances (ICMIA). This model not only can filter the contradiction or chaos messages of IAs but also can generate the control rules automatically. Via this model, it can make home network environment more convenient and comfortable.

References

[1] Lee Huey-Ming, Chen Yen-Chih, Chen Jan-Jo, “The Intelligent Agent Design of Information Appliance,”, JCIS, 2003, Proceeding of the 7th Join Conference on Information Sciences, Cary, NC, USA, pp.1681-1684 September 26-30, 2003
[2] Lee Huey-Ming & Huang Jun-Hong, “The study of IA devices monitoring model”, The sixth seminar of the research and practices of information management, pp.430-437, May 2002.
[3] Lee Huey-Ming, Mao Ching-Hao, “A Fuzzy Clustering Model of Information Appliance”, Third International Conference on Electronic Business (ICEB2003) Singapore, pp. 241-243, Dec 10-12, 2003.
[4] Lee Huey-Ming, Mao Ching-Hao, Lee Shu-Yen, “A Fuzzy Aggregative Clustering Control model of Information Appliance”, WSEAS Transaction of System, May 2004.
[5] Lee Huey-Ming, Mao Ching-Hao, Lee Shu-Yen, “A Fuzzy Neural Network of Information Appliance”, Fuzzy System & Innovation Computing (FIC) 2004, Kitakyushu, Japan, June 2-3, 2004
[6] Wu Chi-Hsiang, Jan Rong-Hong, System integration of WAP and SMS for home network system, COMPUTER NETWORKS.
Effective Solution of a Portfolio Selection Based on a Block of Shares by a Meta-controlled Boltzmann Machine

Teruyuki Watanabe\textsuperscript{1} and Junzo Watada\textsuperscript{2}

\textsuperscript{1} Faculty of Business Administration, Osaka Sangyo University, 3-1-1 Nakagaito, Daito-shi, Osaka 574-8530, Japan  
   watanabe\textasciitilde adm.osaka-sandai.ac.jp

\textsuperscript{2} Graduate School of Information, Production and Systems, Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135, Japan  
   junzow\textasciitilde osb.att.ne.jp

Abstract. In a real investment, stocks are dealt with based on a block of shares. A block of shares is a minimum unit for trading stocks. However, a conventional portfolio selection problem does not consider about a block of shares. If we deal with stocks according to a block of shares, real allocations of funds to each stock should differ among the cases of different amounts of money. Furthermore, a decision maker should be unable to buy less than one block even if the investing ratio for some stock is much smaller.

The objective of this paper is to build a portfolio selection model in consideration of the amount of investing funds and a block of shares. Our model is formulated as an integer quadratic programming problem. In general, an integer nonlinear programming problem is difficult to solve for all but the smallest cases. So we also propose the efficiently approximate model employing a Meta-controlled Boltzmann machine.

1 Introduction

In 1952, H. Markowitz\textsuperscript{11} proposed a method to allocate an amount of funds to plural stocks for investment. The method was named a portfolio selection problem. Based on time-series data of return rate, it theoretically decides the best investing rate to each of stocks, which minimizes the risk, i.e. the variance of the profits in keeping the least expected return rate that a decision maker desires. It is characteristic that the model can reduce the risk by means of allocating the amount of funds to many stocks. The model is excellently concise for real problems.

Since then, researches have been pursued on various aspects of the model, such as realizing efficient calculation\textsuperscript{245,6}.

In a real investment, stocks are dealt with based on a block of shares\textsuperscript{8}. A block of shares is a minimum unit for trading stocks. For example, in a Japan market, except for some cases, dealings are basically conducted according to a block of shares, as shown in Table\textsuperscript{11}.
Table 1. Basic block of shares in Japan markets

| Face value (yen) | The block of shares |
|------------------|---------------------|
| 50               | 1000                |
| 500              | 100                 |
| 50,000           | 1                   |

By the way, the portfolio selection problem proposed by H. Markowitz is formulated based on the premise that the amount of funds can be divided into infinitely small amount without limit. If a decision maker invests the amount of funds according to investing ratio, it is obvious that a block of shares can not be maintained in dealing. If we deal with a stock in a block of shares, real allocations to each stock should differ among the cases where the funds are 1 billion yen and 500 million yen. Furthermore, it is possible that a decision maker may be unable to buy less than one block even if the investing ratio for an objective stock is much smaller.

In consideration of the above point, the objective of this paper is to build a portfolio selection model in consideration of the amount of investing funds and a block of shares. This model defines the investing ratio in terms of a stock unit price, a block of shares, the number of dealing block and the amount of investing funds. Therefore, our model can not obtain a solution with an investment ratio but with the number of dealing blocks. It is possible to obtain a solution more realistically by this formulation than the conventional model.

In general, an integer nonlinear programming problem will be very difficult to solve for all but small cases. This paper proposes an efficiently approximate approach by a Meta-controlled Boltzmann machine. The Meta-controlled Boltzmann machine is a neural network model which is proposed by T. Watanabe and J. Watada et al [5]. This model deletes the units of lower layer which are not selected in the Meta-controlling layer in its execution. Then the lower layer is restructured by using the selected units. Because of this feature, a Meta-controlled Boltzmann machine converges more efficiently than a conventional Boltzmann machine.

In this paper, as a numerical example, we will compare a solution and computing time between the Meta-controlled Boltzmann machine and LINGO 8.0. LINGO 8.0 is a commercial software package in order to solve mathematical programming problems. We will discuss the results in three cases of funds 1 billion yen, 500 million yen and 100 million yen.

2 Portfolio Selection Based on a Block of Shares

In this section we build a portfolio model in consideration of the amount of funds and a block of shares. First, we explain a portfolio selection problem which is proposed by H. Markowitz. Next, we build our model.
2.1 Markowitz Model

A mean-variance approach to a portfolio selection problem was originally proposed by H. Markowitz [1]. H. Markowitz assumed that almost all the decision maker should be a risk aversion person. According to this assumption, H. Markowitz has formulated a portfolio selection problem as the following quadratic programming problem under the restriction that the expected return rate is made sure to be more than amount $R$.

Formulation 1

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij}x_i x_j \\
\text{subject to} & \quad \sum_{i=1}^{m} \mu_i x_i \geq R \\
& \quad \sum_{i=1}^{m} x_i = 1 \\
& \quad x_i \geq 0 \quad (i = 1, 2, \ldots, m)
\end{align*}
\]

where $\sigma_{ij}$ denotes a covariance between stocks $i$ and $j$, $\mu_i$ an expected return rate of stock $i$, $R$ an acceptable least rate of the expected return, $x_i$ an investing rate to stock $i$, respectively.

The objective of this model is to minimize its risk in allocating the amount of funds to many stocks. In a real investment problem, there is a limitation in the amount of funds which a decision maker can manage. Since each of stocks is dealt in a block of shares, it is ordinarily impossible to spend 100% of the provided fund in buying or selling. However, the portfolio selection problem proposed by H. Markowitz is formulated based on assumption that each of stocks can be dealt in an infinitely small unit. The objective of this paper is to propose the portfolio selection model in consideration of a block of shares.

2.2 Formulation of Our Model

Our model is formulated on the basis of the Markowitz model in Formulation 2. In our model, we express an investment ratio of a Markowitz model using a unit price of stock, a block of shares, the number of dealing blocks and the amount of funds which a decision maker can manage. Our model can be rewritten as the following integer quadratic programming model.

Formulation 2

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} \frac{n_i p_i q_i}{B} \frac{n_j p_j q_j}{B} \\
\text{subject to} & \quad \sum_{i=1}^{m} \mu_i \frac{n_i p_i q_i}{B} \geq R
\end{align*}
\]
\[ \sum_{i=1}^{m} n_i p_i q_i \leq B \quad (7) \]
\[ \frac{n_i p_i q_i}{B} \geq 0 \quad (i = 1, 2, \ldots, m) \quad (8) \]

where, \( n_i \) denotes the number of dealing blocks of stock \( i \), \( p_i \) a unit price of stock \( i \), \( q_i \) a block of shares of stock \( i \) and \( B \) an amount of funds, respectively. Other notations are given previously.

The characteristic of our model can obtain a solution not with an investment ratio but with the number of dealing blocks.

### 3 Meta-controlled Boltzmann Machine for Portfolio Selection Based on a Block of Shares

A Meta-controlled Boltzmann machine [5] employs a Hopfield network [8] as a Meta-controlling layer and a Boltzmann machine [7] as a lower layer. The Meta-controlling layer supervises the subordinate lower layer to obtain the best portfolio within the optimal combination of invested stocks and the lower layer decides the optimal investment rate over the limited number of stocks supervised by the Meta-controlling layer. This model deletes units of the lower layer which are not selected in the Meta-controlling layer in its execution. Then the lower layer is restructured by using the selected units. Executing the Meta-controlled Boltzmann machine according to the above mentioned algorithm, the Meta-controlled Boltzmann machine converges more efficiently than a conventional Boltzmann machine.

In this section, we illustrate the energy functions of a Meta-controlled Boltzmann machine to solve a portfolio selection problem formulated as the Formulation 2. The energy functions are written as follows:

**Meta-controlling Layer**

\[ E_u = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} s_i s_j - \frac{K_u}{2} \sum_{i=1}^{m} \mu_i s_i, \quad (9) \]

**Lower Layer**

\[ E_l = \frac{1}{2} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} \frac{n_i p_i q_i}{B} \frac{n_j p_j q_j}{B} \right\} + \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{n_i p_i q_i}{B} \frac{n_i p_i q_i}{B} \right\} \]
\[ - \sum_{i=1}^{m} \frac{n_i p_i q_i}{B} - \frac{K_l}{2} \sum_{i=1}^{m} \mu_i \frac{n_i p_i q_i}{B}, \quad (10) \]

where \( K_u, K_l \) are weight of the expected return rate for each layer and \( s_i \) is output value of unit \( i \) of the Meta-controlling layer. Other notations are defined previously.
Output value, \( s_i \), of the Meta-controlling layer takes 0 or 1. If corresponding stock \( i \) is selected, the output value of unit \( i \) is 1.

In the Meta-controlling layer if \( K_u \) is set to a larger value, the selected number of stocks will increase. In the lower layer if \( K_l \) is set to a smaller value, we can obtain a solution sufficiently approximate to the minimum risk solution.

Therefore, as we change each parameter \( K \), we will obtain various investing pattern according to the aspiration level of a decision maker’s in the same as in a fuzzy portfolio selection model which is proposed by J. Watada et al [6].

4 Numerical Example

In this simulation, we mainly compare computing time employing 225 Japanese stocks. The employed data is daily stock price data which is chosen from Nikkei 225 on the 60th from June 2 to August 26, 1997.

The simulation conditions are shown in Table 2. An acceptable least rate of the expected return \( R \) for LINGO 8.0 is set to 0.001. Moreover, 1 billion yen, 500 million yen, and 100 million yen are an amount of funds which a decision maker can manage as.

Results of Meta-controlled Boltzmann machine are shown in Tables 3 to 5 and solutions of Formulation 2 by LINGO 8.0 are shown in Table 6. We compare the result between the Meta-controlled Boltzmann machine in the case of \( K_l = 0.0 \) and LINGO 8.0. The case where the amount of funds is 1 billion yen, we can obtain the solution which the risk is almost not different and the expected return rate is higher by employing the Meta-controlled Boltzmann machine. The case where an amount of funds is 500 million yen and 100 million yen, we can obtain

| Table 2. Simulation conditions |
|--------------------------------|
| (1) CPU & OS                  |
| CPU Pentium 4 2.4GHz          |
| OS Windows XP                |
| (2) Parameters for Meta-controlled Boltzmann machine |
| \( K_u \) 10.0                |
| \( K_l \) 0.0, 0.3, 0.5      |

| Table 3. Results of Meta-controlled Boltzmann machine(\( K_l = 0.0 \)) |
|------------------------------------------------------------------------|
| Amount of investing funds                                              |
| 1 billion yen 500 million yen 100 million yen                          |
| Expected Return Rate 0.00147 0.00169 0.00138 |
| Risk 0.000050 0.000085 0.000087 |
| Investing funds (yen) 990,775,000 487,778,000 83,679,000 |
| Computing Time (second) 10.92 10.47 10.58 |
Table 4. Results of Meta-controlled Boltzmann machine ($K_l = 0.3$)

| Amount of investing funds | 1 billion yen | 500 million yen | 100 million yen |
|---------------------------|---------------|-----------------|-----------------|
| Expected Return Rate      | 0.00260       | 0.00223         | 0.00146         |
| Risk                      | 0.000122      | 0.000103        | 0.000083        |
| Investing funds (yen)     | 982,501,000   | 484,836,000     | 84,531,000      |
| Computing Time (second)   | 11.72         | 11.10           | 10.29           |

Table 5. Results of Meta-controlled Boltzmann machine ($K_l = 0.5$)

| Amount of investing funds | 1 billion yen | 500 million yen | 100 million yen |
|---------------------------|---------------|-----------------|-----------------|
| Expected Return Rate      | 0.00273       | 0.00251         | 0.00145         |
| Risk                      | 0.000158      | 0.000166        | 0.000083        |
| Investing funds (yen)     | 983,740,000   | 484,113,000     | 82,288,000      |
| Computing Time (second)   | 11.97         | 11.19           | 10.53           |

Table 6. Results of LINGO 8.0

| Amount of investing funds | 1 billion yen | 500 million yen | 100 million yen |
|---------------------------|---------------|-----------------|-----------------|
| Expected Return Rate      | 0.00105       | 0.00101         | 0.00108         |
| Risk                      | 0.000045      | 0.000035        | 0.000028        |
| Investing funds (yen)     | 981,540,500   | 499,977,500     | 88,388,000      |
| Computing Time (second)   | 79.0          | 339.0           | 238.0           |

the sufficiently large expected return rate, although it is a little inferior in the risk.

Let us compare computing time. Table 6 shows that LINGO 8.0 has taken 79.0 seconds by the shortest. Moreover, the case where the investing funds is smaller, LINGO 8.0 has taken hundreds seconds to obtain the solution. Therefore, its solution is not efficient. On the other hand, the Meta-controlled Boltzmann machine can obtain the solution in 10 to 12 seconds.

Moreover, as the value of $K_l$ changes, Tables 3 to 5 shows that the various expected return rate and risk are obtained to satisfy the decision maker’s aspiration level. The above result shows that our Meta-controlled Boltzmann machine is an effective and efficient method in order to solve an integer quadratic programming problem.

5 Concluding Remarks

In this paper, we proposed the portfolio selection model in consideration of the amount of funds and a block of shares and an effective solution by a Meta-controlled Boltzmann machine.
The following points should be emphasized from the result of numerical examples:

- Our model which is shown in Formulation 2 can obtain a solution not with an investment ratio but with the number of dealing blocks. So the solution of our model provides more realistic solution or a real problem.
- Comparing results of numerical example, it is shown that our approximate solution employing a Meta-controlled Boltzmann machine is an effective model in order to solve an integer quadratic programming problem.

References

1. H. Markowitz, “Portfolio Selection,” *Journal of Finance*, Vol. 7, No. 1, pp. 77–91, 1952.
2. H. Konno and K. Suzuki, “A fast algorithm for solving large scale mean-variance model by compact factorization of covariance matrices,” *Journal of the Operations Research Society of Japan*, Vol. 35, pp. 93–104, 1992.
3. H. Konno, Financial Engineering I, Nikkagiren Shuppan-Sha, 1995, in Japanese.
4. T. Watanabe, K. Oda and J. Watada, “Hierarchical Decision Making of Strategic Investment”, *International Journal on Fuzziness, Uncertainty, and Knowledge-Based Reasoning (IJUFKS)*, Vol.7, No. 4, pp.429–438, 1999.
5. T. Watanabe and J. Watada, “A Meta-Controlled Boltzmann Machine for Rebalancing Portfolio Selection,” *Central European Journal of Operations Research*, to appear.
6. J. Watada, “Recent Development of Soft-computing Approach to Portfolio Selection Problem,” Tutorial at ECIT2002: The 2nd European Conference on Intelligent Technology, Iasi, Romania, on July 17–20, pp.1–22, 2002.
7. D. H. Ackley, G. E. Hinton and T. J. Sejnowski, “A Learning Algorithm for Boltzmann Machines,” *Cognitive Science*, Vol. 9, pp.147–169, 1985.
8. J. J. Hopfield: “Neural Networks and Physical Systems with Emergent Collective Computational Abilities,” *Proceedings of National Science*, pp. 2554–2558, 1982
Abstract. A library provides us with books, journals, newspapers and so on according to the demand of its users. The selection of books to buy is also done based on the mission of each library. Usually, they purchase books according to their object as a city, an institute, a school, a university, etc. are founded. It is more important to install books which users are requiring. For example, a city has its life cycle. The population of a young city consists of more young generations, so its library should provide books required by young people but an aged city has elder generations. So the library in the aged city has high rate of books required by elder people. Therefore, a library should adjust their selection of books to the demand of users.

In this paper we provide a method how to adjust selection of books to the demand of users. The objective of the paper is to propose a method for strategic decision of selecting books. A method is also to deal with latitude in an aspiration level of the decision maker according to their mission. In this method, a librarian and a manager define, for each of expected utility rate and its variance, a necessity level and a sufficient level of users. And then, they can obtain a solution that satisfies an aspiration level of the users.

We employ this method to analyze and decide which kind of books and how much rate of the total budget should be spent to buy, where the method of fuzzy mean-variance analysis is employed to solve the problem.

Keywords: Library Strategy, Budget Allocation to Book Category, Fuzzy Mean-Variance Analysis, Portfolio Selection

1 Introduction

A library provides us with books, journals, newspapers and so on according to the demand of its users. The selection of books to buy is also done based on their own mission of each library. Usually, they purchase books according to the mission as a city, an institute, a school, a university, etc. are founded. It is more important to install books which users are requiring. For example, a city has its life cycle. The population of a young city consists of more young generations, so its library should provide books required by young people but an aged city has elder generations. So the library in the aged city has high rate of books required by elder people. Therefore, a library should adjust their selection of books to the demand of users. In this paper we provide a method how to adjust selection of books to the demand of users.
Mean-Variance Analysis proposed by H. Markowitz is widely employed in stock investment. The fuzzy mean-variance analysis enables us to obtain a solution which realizes the best within a vague aspiration level and a fuzzy goal given as a fuzzy number, which is obtained from the expertise of decision makers. It should be important to take both variance and expected return under the consideration in decision making on budget allocation, because sales volume is not stable and widely changed. It is also hard to express an aspiration beheld of a decision maker using a constant rigid value.

The objective of the paper is to propose a method for strategic decision of selecting books in order to purchase based on a fuzzy portfolio theory. A method is also to deal with latitude in an aspiration level of the decision maker according to their mission. In this method, a librarian and a manager defines, for each of expected renting or using rate and its variance, a necessity level and a sufficient level of users. And then, they can obtain a solution that satisfies an aspiration level of the users.

We employ this method to analyze and decide which kind of books and how much rate of the total budget should be spent to buy, where the method of fuzzy mean-variance analysis is employed to solve the problem.

Since the demand of users in future is uncertain, it is difficult to decide the proper demand of books. In this paper we propose a method of decision making under consideration of the past utility variance of some book category (classification of books) and utility frequency in order to realize some amount of utility. Generally, it is not easy to express an aspiration level of users employing a constant value. In this paper, we employ a fuzzy number to define it with latitude, and propose the fuzzy mean-variance analysis which satisfies an aspiration level of users. Fuzzy mean-variance analysis is flexible about interpreting of a decision maker.

In the conventional portfolio selection, its problem is formulated as a quadratic programming which minimizes variance under the realization of the expected return rate which a decision maker expects. The expected return and risk hold the trade off relation. Therefore, we formulate two objective mean-variance analysis, which can obtain the best solution under the consideration of uncertain aspiration levels that a decision maker might have for both indices of expected return rate and risk. The fuzzy mean-variance analysis enables us to obtain a solution that realizes the best within a vague aspiration level and a fuzzy goal given as a fuzzy number.

This analysis provides us with a satisfactory solution which clears the objective level. In this paper, we propose the strategic method of selecting book category and allocating the total budget which can be solved by means of fuzzy mean-variance analysis both to maximize expected utility of books and to minimize its variance.

2 Mean-Variance Analysis

A mean-variance analysis, that is, portfolio selection analysis is widely used as investment theory proposed by H. Markowitz in the early 1950s. In the formulation of the mean-variance analysis, H. Markowitz started his discussion with the assumption that almost all decision makers have aversion to risk even if they lose its return more.
However, it should be difficult to identify a utility function because they have a different utility function of their own. So, H. Markowitz has formulated the mean-variance analysis problem as the following quadratic programming problem under the constraint that the expected return rate is made sure to be more than some amount.

Let us discuss the budget allocation to book categories according to the mean-variance analysis. Its optimal solution with the least variance is searched under the constraint that the expected access rate to each book category should be more than value $R$ that decision-maker arbitrarily gives. The allocation rate of budget to each book category is decided for the solution with the least variance to the given expected access rate to the category. Since the variance is estimated under the condition of fixing the expected access rate to the category, the decision-maker cannot be fully satisfied of the solution. Therefore, it is much proper to employ the following Formulation 1:

Formulation 1.

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{n} \mu_i \chi_i \\
\text{minimize} & \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} \chi_i \chi_j \\
\text{Subject to} & \quad \sum_{i=1}^{n} \chi_i = 1 \\
& \quad \chi_i \geq 0 (i = 1, 2, \ldots, n) 
\end{align*}
\]

where $R$ denotes an acceptable least rate of the expected access to each book category, $\sigma_{ij}$ a covariance between categories $i$ and $j$, $\mu_i$ an expected access rate of book category $i$, and $\chi_i$ a budget allocation rate to book category $i$, respectively.

Formulation 1 is a quadratic programming problem with two objective functions of an expected access rate to each book category and its variance.

### 3 Fuzzy Mean-Variance Analysis

In order to overcome the issue of degeneration mentioned above, we should employ a non-linear membership function such as a tangent one defined by H. Leberling, which has asymptotic lives $\lambda_1 = 1$ and $\lambda_2 = 0$.

In this paper we employ the logistic function for a non-linear membership function as follows:

\[
f(x) = \frac{1}{\left(1 + \exp(-\alpha(x))\right)}
\]
The logistic function has a similar shape as the tangent hyperbolic function employed by H. Leberling, but it is more easily handled than the tangent hyperbola. And also a trapezoidal membership function is an approximation of the logistic function. Therefore, the logistic function is considered much more appropriate to denote a vague goal level which a decision maker considers.

The goal rate for an expected return can be described using the logistic membership function in the following:

$$
\mu_E(E(x)) = \frac{1}{1 + \exp(-\alpha_E(E(x) - E_M))}
$$

where $E_M$ is the mid point where membership grade $\lambda$ is 0.5. Figure 1 shows the logistic membership function of the goal rate of an expected return.

The logistic membership function of the goal for variance. The goal for variance can be described using the logistic membership function in the following:

$$
\mu_V(V(x)) = \frac{1}{1 + \exp(\alpha_V(V(x) - V_M))}
$$

where $V_M$ is the mid point where membership value $\lambda$ is 0.5. Notations $\alpha_E$ and $\alpha_V$ influence respectively on the shapes of membership functions $\mu_E$ and $\mu_V$, respectively, where $\alpha_E > 0$ and $\alpha_V > 0$. The larger parameters $\alpha_E$ and $\alpha_V$ get, the less their vagueness becomes. Since the logistic function is monotonously increasing, maximizing $\lambda$ makes $\lambda$, maximized, that is, maximizes $\log\{\lambda/(1-\lambda)\}$.

$$
\lambda = \frac{1}{1 + \exp(-\hat{\lambda})}
$$

Accordingly, Formulation 1 is equivalent to Formulation 2 as follows:

Formulation 2.

maximize $\lambda$

Subject to

$$
\alpha_V V(x) + \lambda_i \leq \alpha_V V_M
$$

$$
\alpha_E E(x) + \lambda_i \geq \alpha_E E_M
$$

$$
\sum_{i=1}^{n} \chi_i = 1
$$

$$
\lambda_i, \chi_i \geq 0 (i = 1, 2, \ldots, n)
$$
4 Applications to Books Allocation Strategy for Library

4.1 Initial Setting of Parameters

In this analysis parameters are initially set as follows:

1. Necessity and saficiency levels of the expected return rate and variance are set to as $V_L = 0.14$, $V_U = 0.04$, $E_L =0.02$ and $E_U =0.14$.

2. The relation Equation (15) or (16) between $\alpha_V$ and $\alpha_E$ decides $\alpha_V = 1.2 \alpha_E$.

3. Set values $\alpha_V = 60$, $\alpha_E = 50$ will be standard ones.

4.2 Result

In the case $\alpha_V = 60$ and $\alpha_E = 50$ which emphasizes the policy of maximization of its $\lambda$ as Table 2 shows, book category A should be allocated as 9.8%, book category B as 7.2%, book category C as 20.1%, book category D as 25.6%, book category E as 8.7%, book category G as 55.3%, book category H as 15.8%, book category L as 18.7%, book category M as 11.6%, and book category F, book category I, book category J and book category K has no budget allocation.

Table 1. Membership grade lambda, variance and expected return of fuzzy mean-variance analysis with the logistic membership functions

| $\alpha_V$ | $\alpha_E$ | $\lambda$ | Variance | Expected return |
|-------|-------|--------|---------|----------------|
| 60    | 50    | 0.64846| 0.07980 | 0.09225        |

Table 2. Budget allocation rate of fuzzy mean-variance analysis with the logistic membership functions

| $\alpha_V$ | $\alpha_E$ | A   | B   | C   | D   | E   | F   | G   |
|-------|-------|-----|-----|-----|-----|-----|-----|-----|
| 60    | 50    | 0.0982 | 0.0718 | 0.2010 | 0.0256 | 0.0872 | 0.0000 | 0.0553 |
| $\alpha_V$ | $\alpha_E$ | H   | I   | J   | K   | L   | M   |
|-------|-------|-----|-----|-----|-----|-----|-----|
| 60    | 50    | 0.15782 | 0.00000 | 0.00000 | 0.00000 | 0.18696 | 0.11612 |

As Table 2 shows, it will be decided the budget allocation to each book category. The purchasing budget of each book category is obtained from multiplying the total budget volume by the selection rate of the book categories. The purchasing budget of the book category is calculated by dividing the total budget of the
purchasing books. The following is the result which is obtained by the mean-variance analysis.

As a result, book category A should be purchased as 117 pieces, book category B as 86 units, book category C as 168 units, book category D as 31 units, book category E as 74 units, book category G as 110 units, book category H as 354 units, book category L as 418 units and book category M as 263 units.

Table 3. Decided budget allocation of each book category

| $\alpha_V$ | $\alpha_E$ | A  | B  | C  | D  | E  | F  | G  |
|------------|------------|----|----|----|----|----|----|----|
| 60         | 50         | 117.5 | 86.6 | 168.9 | 31.2 | 74.2 | 0.0 | 110.8 |

| $\alpha_V$ | $\alpha_E$ | H  | I  | J  | K  | L  | M  |
|------------|------------|----|----|----|----|----|----|
| 60         | 50         | 354.2 | 0.0 | 0.0 | 0.0 | 418.2 | 263.0 |

5 Concluding Remarks

It is hard to decide the amount of budget allocation because the utility amount and volume of each book category are not known previously. In this paper, we employed a fuzzy number to define it with latitude, and propose the fuzzy budget allocation portfolio management which satisfies an aspiration level of a decision maker.

As expected return rate and variance is in a trade-off relation, it is much more proper to formulate two objective mathematical programming which maximizes the expected return rate and minimizes its variance than over objective mathematical minimizes the variance under the fixed expected return rate. In this paper we formulated two objective mean-variance analysis, then we employed the aspiration level of a decision maker's in the two objective mathematical programming. Especially, a sigmoid function is employed as a membership function which enables to differentiate among feasible solutions a trapezoidal one has membership grade 1.

In this paper, we proposed a fuzzy mean-variance analysis in the strategic decision of sales management which satisfies a decision maker, that is, maximize the expected return rate and minimize its variance.

References

[1] Wtada J: “Fuzzy portfolio selection and its applications to decision making”, Tatra Mountains Math. Publ. 13, pp.219-248 (1997)

[2] Konno Hiroshi: Financial Engineering 1, Nikkagiren Publishing, pp1-14 (1995) in Japanese.

[3] Kawaura Takayuki and Watada Junzo: “Mean-Variance Analysis of Agricultural Management based on a Boltzmann Machine,” Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'99), -1196-1201 (1999)
[4] Kawaura Takayuki, Watada Junzo and Watanabe Teruyuki: “Neural Network Approach to Sales Portfolio Management,” Proceedings of The International Symposium on Medical Informatics and Fuzzy Technology(MIF’99), pp120-123 (1999)

[5] Mizunuma Hiroto and Watada Junzo: “Fuzzy Portfolio Selection -Realization of an Aspiration Level Given by a Decision Maker,” Trans.of The Institute of Systems, Control and Information Engineers, Vol.8, No.12, pp.667-684 (1995) in Japanese.

[6] Markowitz H: Mean-Variance Analysis in Portfolio Choice and Capital Ma
Analysis of Human Feelings to Colors

Taki Kanda

Bunri University of Hospitality, Department of Service Management, 311-1 Kashiwabarashinden Shinogawara, Sayama, Saitama 350-1336, Japan

Abstract. In this paper human feelings to colors are studied investigating the relation between colors of food and feelings of tastiness of food. It is important for meal marketing to take consumer’s feelings into consideration since consumer’s choice of foodstuff is much influenced by human feelings. What we feel tasty to food is not related with only the sense of eating but also the sense of smell, sight, touch, hearing, etc. because it does not concern only taste but also savor, flavor, dishing up or coloring on dining tables, the temperature or texture of food, the sound of chewing food, etc. The sense of sight is the most delicate sense organ compared with other sense organs of human body. As for the sense of sight it is considered that the appearance of food such as the colors of the surface, the shape or gloss has a great influence on the human feelings of tastiness. Here the colors of the surface of food are perceived and it is studied how we feel tasty to food depending on the colors of the surface of food.

1 Introduction

This study is concerned with the relation between colors and human feelings. Human feelings to colors are considered to be different among individuals and also very with object. Taking food for example, what we feel tasty to food is not related with only the sense of eating but also the sense of smell, sight, touch or hearing. Especially the sense of sight is the most delicate sense organ among many human organs. It is therefore considered that colors of food influence human feelings of tastiness. Here human preferable colors for candies are investigated based upon the data obtained from the experiment on evaluation of preference of colors for candies conducted for this study letting subjects with different properties-Japanese kindergarten pupils, Japanese University students, Foreign University students in Japan from Asia. In the investigation whether human feelings to colors agree among subjects is statistically found out and human preference of colors for candies is evaluated using order statistics defined based upon the standard normal distribution. Here how to conduct experiments on evaluation of preference to colors for candies and statistically analyze the obtained data is described and what to find or conclude from the results is considered.
2 Experiment on Evaluation of Preference

Experiments on evaluation of preference concerning the feelings of tastiness to the colors of the surface of food have been conducted for this study. Subjects are 20 Japanese kindergarten pupils of ages from 3 to 6, 25 Japanese University students of ages from 19 to 26 and 31 foreign University students in Japan from Asia of ages from 21 to 30. In the experiments subjects are asked to range 7 candies in different colors (red, blue, yellow, white, black, green and orange) according to how the candies look tasty. The results of the experiments for 20 Japanese kindergarten pupils, 25 Japanese University students and 31 foreign University students from Asia are shown in Tables 1, 2 and 3 respectively. Based upon the results of the experiments the concordance of preference among subjects of each group has been investigated by Kendall’s coefficient of concordance and Freedman’s test, and preference of Japanese kindergarten pupils, Japanese University students and foreign University students from Asia concerning the feelings of tastiness to the colors of the surface of candies has been numerically evaluated by ranking method.

3 Test on Concordance of Preference

For the investigation of the concordance of preference to be evaluated by ranking method Kendall’s coefficient of concordance and Freedman’s test are used. For Kendall’s coefficient of concordance let \( n \) and \( k \) be the number of subjects and stimuli respectively and \( V_i (i = 1, 2, \cdots, k) \) be the total of the ranking which subjects give to each stimulus then the mean of \( V_i \) is

\[
\overline{V} = \frac{n(k + 1)}{2}
\]

and the sum of the squares of the deviation from the mean \( \overline{V} \)

\[
S = \sum_{i=1}^{k} (V_i - \overline{V})^2
\]

divided by the maximum of \( S \) becomes Kendall’s coefficient of concordance as

\[
W = \frac{S}{S_{\text{max}}}, 0 \leq W \leq 1
\]

where \( S_{\text{max}} \) is the maximum of \( S \). For Freedman’s test to confirm the concordance of the ranking by \( n \) subjects the test statistic
\[ F_0 = \frac{(n - 1)W}{1 - W} \]  

is calculated and if the calculated value of \( F_0 \) is not less than the critical value of the \( F \) distribution for the degrees of freedom

\[ f_1 = k - 1 - \frac{2k}{n}, f_2 = (n - 1) f_1, \]

**Table 1.** Ranking of 20 Japanese kindergarten pupils

| Subjects | Red | Blue | Yellow | White | Black | Green | Orange |
|----------|-----|------|--------|-------|-------|-------|--------|
| No.1     | 1   | 6    | 7      | 2     | 3     | 5     | 4      |
| No.2     | 2   | 6    | 1      | 5     | 7     | 4     | 3      |
| No.3     | 1   | 3    | 2      | 5     | 4     | 6     | 7      |
| No.4     | 2   | 5    | 4      | 6     | 7     | 3     | 1      |
| No.5     | 4   | 6    | 2      | 3     | 7     | 5     | 1      |
| No.6     | 2   | 6    | 4      | 3     | 7     | 5     | 1      |
| No.7     | 6   | 5    | 3      | 1     | 7     | 4     | 2      |
| No.8     | 1   | 3    | 5      | 6     | 7     | 4     | 2      |
| No.9     | 6   | 5    | 3      | 1     | 7     | 4     | 2      |
| No.10    | 2   | 7    | 4      | 6     | 1     | 5     | 3      |
| No.11    | 5   | 1    | 4      | 2     | 6     | 3     | 7      |
| No.12    | 3   | 6    | 1      | 4     | 7     | 5     | 2      |
| No.13    | 4   | 3    | 6      | 5     | 7     | 2     | 1      |
| No.14    | 2   | 6    | 4      | 3     | 7     | 5     | 1      |
| No.15    | 4   | 6    | 2      | 1     | 7     | 5     | 3      |
| No.16    | 1   | 6    | 3      | 5     | 7     | 4     | 2      |
| No.17    | 4   | 1    | 3      | 7     | 6     | 2     | 5      |
| No.18    | 6   | 1    | 4      | 7     | 2     | 3     | 5      |
| No.19    | 2   | 6    | 3      | 5     | 7     | 4     | 1      |
| No.20    | 2   | 6    | 3      | 4     | 7     | 5     | 1      |
| Total    | 60  | 94   | 68     | 81    | 120   | 83    | 54     |
it is concluded that preference is concordant among \( n \) subjects, that is, if 
\[ F_0 \geq F(f_1, f_2, \alpha), \]
the concordance of preference among \( n \) subjects is significant.

### Table 2. Ranking of 25 Japanese University Students

| Subjects | Red | Blue | Yellow | White | Black | Green | Orange |
|----------|-----|------|--------|-------|-------|-------|--------|
| No.1     | 5   | 6    | 4      | 1     | 7     | 2     | 3      |
| No.2     | 5   | 4    | 1      | 2     | 7     | 3     | 6      |
| No.3     | 5   | 6    | 2      | 1     | 7     | 3     | 4      |
| No.4     | 2   | 3    | 5      | 7     | 6     | 4     | 1      |
| No.5     | 2   | 1    | 6      | 5     | 7     | 4     | 3      |
| No.6     | 1   | 6    | 3      | 4     | 7     | 5     | 2      |
| No.7     | 6   | 1    | 3      | 2     | 7     | 4     | 5      |
| No.8     | 3   | 2    | 6      | 7     | 1     | 5     | 4      |
| No.9     | 5   | 2    | 6      | 7     | 1     | 4     | 3      |
| No.10    | 3   | 5    | 2      | 1     | 6     | 7     | 4      |
| No.11    | 5   | 6    | 3      | 4     | 7     | 2     | 1      |
| No.12    | 3   | 4    | 6      | 7     | 5     | 1     | 2      |
| No.13    | 5   | 3    | 2      | 1     | 4     | 7     | 6      |
| No.14    | 5   | 4    | 1      | 2     | 7     | 3     | 6      |
| No.15    | 3   | 6    | 2      | 1     | 5     | 7     | 4      |
| No.16    | 1   | 6    | 7      | 2     | 4     | 3     | 5      |
| No.17    | 2   | 5    | 4      | 7     | 6     | 1     | 3      |
| No.18    | 3   | 6    | 4      | 1     | 7     | 5     | 2      |
| No.19    | 3   | 7    | 1      | 5     | 6     | 2     | 4      |
| No.20    | 1   | 4    | 5      | 7     | 6     | 2     | 3      |
| No.21    | 5   | 2    | 3      | 6     | 4     | 7     | 1      |
| No.22    | 4   | 7    | 3      | 1     | 6     | 5     | 2      |
| No.23    | 1   | 5    | 3      | 7     | 6     | 4     | 2      |
| No.24    | 4   | 2    | 3      | 6     | 7     | 5     | 1      |
| No.25    | 2   | 5    | 4      | 7     | 6     | 1     | 3      |
| Average  | 84  | 108  | 89     | 101   | 142   | 96    | 80     |

### 4 Result of Test on Concordance of Preference

Statistical test described in 3 has been conducted to find out whether preference among 20 Japanese Kindergarten pupils, 25 Japanese University students and 31 Foreign University students from Asia is concordant respectively and the results are obtained as shown in Table 4.
For significance test since we have

$$F_0 = 7.03 \geq F(5, 60, 0.05) = 2.37 > F(f_1, f_2, 0.05) \quad (6)$$

for 20 Japanese kindergarten pupils,

$$F_0 = 4.23 \geq F(5, 120, 0.05) = 2.29 > F(f_1, f_2, 0.05) \quad (7)$$

**Table 3.** Ranking of 31 foreign University Students

| Subjects | Red | Blue | Yellow | White | Black | Green | Orange |
|----------|-----|------|--------|-------|-------|-------|--------|
| No.1     | 3   | 5    | 2      | 4     | 7     | 6     | 1      |
| No.2     | 2   | 4    | 6      | 1     | 7     | 5     | 3      |
| No.3     | 5   | 1    | 2      | 7     | 6     | 4     | 3      |
| No.4     | 4   | 3    | 6      | 1     | 7     | 5     | 2      |
| No.5     | 5   | 6    | 3      | 2     | 7     | 4     | 1      |
| No.6     | 4   | 1    | 3      | 7     | 6     | 5     | 2      |
| No.7     | 1   | 5    | 4      | 6     | 7     | 2     | 3      |
| No.8     | 6   | 5    | 4      | 2     | 7     | 3     | 1      |
| No.9     | 6   | 5    | 3      | 1     | 7     | 2     | 4      |
| No.10    | 7   | 6    | 1      | 3     | 5     | 2     | 4      |
| No.11    | 1   | 2    | 3      | 5     | 6     | 4     | 7      |
| No.12    | 2   | 3    | 5      | 7     | 1     | 6     | 4      |
| No.13    | 6   | 5    | 4      | 3     | 7     | 2     | 1      |
| No.14    | 6   | 5    | 7      | 3     | 4     | 2     | 1      |
| No.15    | 5   | 7    | 3      | 1     | 6     | 2     | 4      |
| No.16    | 5   | 6    | 2      | 3     | 7     | 1     | 4      |
| No.17    | 1   | 6    | 3      | 5     | 7     | 2     | 4      |
| No.18    | 1   | 4    | 7      | 5     | 6     | 3     | 2      |
| No.19    | 6   | 5    | 7      | 3     | 4     | 2     | 1      |
| No.20    | 2   | 4    | 7      | 7     | 5     | 3     | 1      |
| No.21    | 2   | 1    | 6      | 7     | 4     | 5     | 3      |
| No.22    | 1   | 4    | 5      | 6     | 7     | 3     | 2      |
| No.23    | 5   | 6    | 2      | 1     | 7     | 3     | 4      |
| No.24    | 3   | 1    | 7      | 5     | 2     | 6     | 2      |
| No.25    | 7   | 6    | 4      | 2     | 5     | 1     | 3      |
| No.26    | 2   | 1    | 6      | 7     | 4     | 5     | 3      |
| No.27    | 5   | 6    | 3      | 2     | 7     | 1     | 4      |
| No.28    | 6   | 3    | 2      | 5     | 7     | 4     | 1      |
| No.29    | 7   | 5    | 2      | 4     | 1     | 3     | 6      |
| No.30    | 1   | 2    | 5      | 6     | 7     | 4     | 3      |
| No.31    | 5   | 7    | 2      | 1     | 6     | 3     | 4      |
| Average  | 122 | 130  | 126    | 122   | 176   | 103   | 88     |
for 25 Japanese University students and

\[ F_0 = 3.24 \geq F(5,120,0.05) = 2.29 > F(f_1, f_2, 0.05) \]  \hspace{1cm} (8)

for 31 foreign University students from Asia respectively, the calculated values of the test statistic \( F_0 \) are not less than the critical values \( F(f_1, f_2, 0.05) \) for 20 Japanese kindergarten pupils, for 25 Japanese University students and for 31 foreign University students in Japan from Asia and it is concluded that the concordance of preference concerning colors and good tasting is significant as far as our conducted experiments on evaluation of preference go.

5 Evaluation of Preference

In order to evaluate preference on tastiness to the colors of candies based upon the ranking given by subjects, the weight of each order is determined by using order statistics which is defined based upon the normal distribution as shown in Table 5. Tables 6 and 7 show the scales and the ranking of colors for the feelings of tastiness to the colors of candies respectively. These tables state that orange is the most favorable colors for candies and black is the most unfavorable colors among Japanese kindergarten pupils, among Japanese University students and among foreign University students from Asia respectively. It is seen from this table that preference of Japanese kindergarten pupils and Japanese University students are similar but differs from that of foreign University students from Asia on the whole.

6 Comparison of Preference

Now it is found out whether preference concerning the colors of candies and the feelings of tastiness varies depending on properties of subjects. To do this Pearson’s moment coefficient of correlation and Spearman’s rank coefficient of correlation are obtained for Japanese kindergarten pupils and Japanese University students, Japanese University students and foreign University students from Asia, and foreign University students from Asia and Japanese kindergarten pupils respectively as shown in Table 8. Pearson’s moment coefficient of correlation is calculated based upon the scales on preference for 7 colors of candies given by Table 6 and Spearman’s rank coefficient of correlation is calculated based upon the ranking of the scales on preference for 7 colors of candies given by Table 7. In terms of Table 8 this states that preference is

| Subjects                          | Kendall's coefficient of concordance | Degree of freedom | Test statistic |
|----------------------------------|--------------------------------------|-------------------|---------------|
| 20 Japanese kindergarten pupils | 0.27                                 | 5.90              | 112.10        | 7.03          |
| 25 Japanese University students | 0.15                                 | 5.92              | 142.08        | 4.23          |
| 31 foreign University students  | 0.10                                 | 5.94              | 178.06        | 3.24          |
similar between Japanese kindergarten pupils and Japanese University students but relatively different between Japanese University students and foreign University students from Asia and also different between foreign University students from Asia and Japanese University students.

Table 5. Weights to the ranking given by subjects

| Order | 1st  | 2nd  | 3rd  | 4th  | 5th  | 6th  | 7th  |
|-------|------|------|------|------|------|------|------|
| Weight| 1.352| 0.757| 0.353| 0.000| -0.353| -0.757| -1.352|

Table 6. Scales of colors with respect to the feelings of tastiness

| Subjects | Red  | Blue | Yellow | White | Black | Green | Orange |
|----------|------|------|--------|-------|-------|-------|--------|
| S1       | 0.422| -0.243| 0.232  | -0.006| -0.899| -0.050| 0.545  |
| S2       | 0.279| -0.119| 0.179  | -0.022| -0.703| 0.053 | 0.333  |
| S3       | 0.049| -0.048| -0.056 | 0.024 | -0.723| 0.276 | 0.483  |

S1: 20 Japanese kindergarten pupils
S2: 25 Japanese University Students
S3: 31 foreign University students from Asia

Table 7. Ranking of colors with respect to the feelings of tastiness

| Subjects | 1st  | 2nd  | 3rd  | 4th  | 5th  | 6th  | 7th  |
|----------|------|------|------|------|------|------|------|
| S1       | Orange | Red  | Yellow | White | Green | Blue | Black |
| S2       | Orange | Red  | Yellow | Green | White | Blue | Black |
| S3       | Orange | Green | Red  | White | Blue  | Yellow | Black |

Table 8. Comparison of preference depending on subjects

| Subjects       | Pearson’s moment coefficient of correlation | Spearman’s rank coefficient of correlation |
|----------------|---------------------------------------------|-------------------------------------------|
| S1 and S2      | 0.987                                       | 0.964                                     |
| S2 and S3      | 0.883                                       | 0.714                                     |
| S3 and S1      | 0.847                                       | 0.643                                     |

7 Concluding Remarks

In this paper the relation between the colors of candies and good tasting was studied conducting the experiments on evaluation of preference to colors of candies and statistically analyzing the data from the conducted experiments. It has been seen from the experiments and analysis that the concordance of preference with respect to the relation
between the colors of candies and good tasting has been confirmed among 20 Japanese kindergarten pupils, 25 Japanese University students and 31 foreign University students in Japan from Asia respectively. Preference for colors and good tasting has been also evaluated by using order statistics and it has been found that preference between Japanese kindergarten pupils and Japanese University students are quite similar but that between Japanese kindergarten pupils and foreign University students from Asia, Japanese University students and foreign University students from Asia are slightly different.
Possibilistic Forecasting Model and Its Application to Analyze the Economy in Japan

Yoshiyuki Yabuuchi¹ and Junzo Watada²

¹ Shimonoseki City University, 2-1-1 Daigaku, Shimonoseki, Yamaguchi 751-8510, Japan 
yabuuchi@shimonoseki-cu.ac.jp
² Waseda University, 2-7 Hibikino, Wakamatsuku, Kitakyushu, Fukuoka 808-0135, Japan 
junzow@osb.att.ne.jp

Abstract. It is hard to separate samples according to each latent system in the case of multivariate data. Hitherto, there are many researches to investigate the structure under obtained data and analyze such data. J. C. Bezdek proposes Switching Regression Model based on Fuzzy Clustering Model to formulate a forecasting model. The model proposed by Bezdek is to separate mixed samples coming from plural latent systems and apply each regression model to the group of samples coming from each system. That is a Fuzzy c-Regression Model.

In this paper, in order to deal with the possibility of a system, we employ a fuzzy forecasting model such as a Fuzzy Switching Regression Model and a Fuzzy Switching Time Series Model. The fuzzy forecasting regression model is explained to analyze the economy in Japan.

1 Introduction

There is a possibilistic forecasting model to analyze a system and interpret the possibility of the system by forecasting mode. These forecasting model are possibilistic regression model, possibilistic time series model [3,5,8], and so on. The objective of these possibilistic forecasting model is to interpret the possibility of the system has.

Possibilistic forecasting model interpret the possibility of the system by including data realized the possibility. If outlier ware included in data, a possibility that a possibilistic forecasting model interpreted might be widely distorted, some times. In addition, if data ware observed from some systems, forecasting model should be hard to correctly interpret system. Many work are done to correspond to these problems.

Models to correspond to outlier are proposed by [2,6]. And a mode to correspond to data that mixed possibility of some systems is proposed by [1,4,7,9].

In this paper, we formulate possibilistic forecasting model and these model was explained to analyze the economy in Japan.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 151–158, 2004. © Springer-Verlag Berlin Heidelberg 2004
2 Possibilistic Regression Model

As an observed sample embodies one of possibilities that the considered system has, the measured data should be interpreted as possibilities of the system. Therefore, a fuzzy regression model is built in terms of the possibility and evaluates all observed values as possibilities which the system should contain. Then, the fuzzy regression model is named a possibilistic regression model. In other words, the fuzzy regression model should be built so that it could contain all observed data in the estimated fuzzy numbers resulted from the model.

The fuzzy regression equation is written as in the following:

\[ Y_j = A_1 x_{1j} + A_2 x_{2j} + \cdots + A_p x_{pj} = A x_j \]

\[ x_{1j} = 1; j = 1, 2, \cdots, n \]  (1)

Where a regression coefficient \( A_i \) is a triangular-shaped fuzzy number \( A_i = (a_i, c_i) \) with center \( a_i \) and width \( c_i \), and \( x_j \) is an input vector. \( A \) is an \( p \) fuzzy coefficient row vector, \( X \) is an \( n \times p \) input matrix, \( Y \) is an \( n \) fuzzy output row vector.

According fuzzy arithmetic \( A x_j = (a, c) x_j = (a x_j, c |x_j|) \), the output of the fuzzy regression equation (1), whose coefficients are fuzzy numbers, results in a fuzzy number. Where, \( |x_j| = [|x_{ij}|] \). The regression model with fuzzy coefficients can be expressed with lower boundary \( a x_j - c |x_j| \), center \( a x_j \) and upper boundary \( a x_j + c |x_j| \). The fuzzy regression model is built to contain all data in the model. As a result, since the model contains all given data in itself, the width of the model gets larger in the case where data are widely scattered. On the other hand the width gets narrower in the case where data gather near the regression line. When the width of the model is large, the expression of its regression equation is vague. It is better and more convenient to obtain a clear and rigid expression. Therefore, the width of the regression should be minimized as removing the vagueness of the model as possible. The fuzzy regression model is formulated to minimize its width the including relations between data and the fuzzy regression model. This problem results in a linear programming.

Using the notations of observed data \( D_j = (y_j, d_j) \), fuzzy coefficients \( (a, c) \), and the \( j \) th column component, \( x_j \), of input values \( X \), the regression model can be mathematically written in the following LP problem:

\[
\begin{align*}
\text{minimize} & \quad \sum_{j=1}^{n} c |x_j| \\
\text{subject to} & \quad y_j + d_j \leq a x_j + c |x_j| \\
& \quad y_j - d_j \geq a x_j - c |x_j| \\
& \quad c \geq 0 \quad (j = 1, 2, \cdots, n)
\end{align*}
\]  (2)

This fuzzy regression contains all data in its width and results in expressing all possibilities which data embody, that is, which the considered system should have. It is possible in the formulation of the fuzzy regression model to treat non-fuzzy data with no width by setting the width \( d \) to 0 in the above equations.
3 Switching Regression Model Based on Possibilistic Regression Analysis

The switching regression model proposed by J. C. Bezdek is to divide samples into clusters by means of a fuzzy cluster analysis. In the Bezdek’s method, fuzzy c-regression models is to clusterize whole data using a fuzzy cluster analysis and to analyze the samples of each cluster using a regression analysis. In terms of statistics, a regression model analyzes data based on the minimum of mean squares errors to build a model. Therefore, a cluster analysis has consistence with a regression model from the statistical point of view. On the other hand, as a fuzzy regression model is to express the possibility included in the observed system, the clusters should be parted in terms of possibility.

When a fuzzy cluster analysis is directly applied to clusterize data, the obtained model should have torsted. The reason is because the model based on statistics is hard to be influenced by the data distribution as the model is built by the minimum of least squire errors.

If the distribution of data has the shape of the ellipse. The 2nd axis and 3rd axis which have less variance than the 1st axis are better because it is avoided to make the width of the model large.

Applying switching regression model based on fuzzy regression model to data, one cluster should be assigned to data which have the same possibilistic distribution. Out of such distribution of data, samples are parted to minimize the width of data distribution.

In this paper, we employ a principal component analysis to separate each possibilistic distribution. The purpose which we employ a fuzzy covariance is to figurize a distribution of data. This is a mean that we employ eigenvalues which are a variance of the separated data to evaluate a distribution of data. In order to analyze data by a fuzzy regression model using a fuzzy covariance, we indicate the following two key words as follows:

1. The 1st eigenvalue should have a large value to have a linearity
2. The other eigenvalues should have a small value to minimize the width of the model or to minimize a fuzziness in other words.

Let us denote an eigenvalue of the cluster \( i (i = 1, \ldots, m) \) as \( \lambda_i (\lambda_{i1} > \lambda_{i2} > \cdots > \lambda_{ip}) \).

The evaluating conditions have the same meaning as the objective of a fuzzy regression model. The model has two objects which are to express the possibility that the system has and to build a fuzzy regression model to minimize a fuzziness of the model. These intend to separate data which have a linearity on the 1st axis and to minimize the width of the possibility of the system along the 2nd axis.

And we employ a distance between data and the 1st principal component in order to make a high density of possibilistic distribution.

The combination of \( n \) samples which are separated to \( m \) clusters is approximately rewritten as \( m^n / m! \). For example, when \( m = 2 \) and \( n = 40 \), a combination size which separates samples into \( m \) clusters is \( 2^{40} / 2! = 2^{39} \).
We should effectively search an optimum separation from very large combina-
tion size. Therefore, in this paper, we employ a genetic algorithm to separate
data into each of systems.

An individual has a gene with a grade to which extent it is included in each
of systems; these fitnesses are calculated by an eigenvalue of each system and a
distance between the data and the 1st principal component of the same system.

On distribution \( i \) and individual \( j \), using eigenvalue \( \lambda_{i}^{(j)} \), data size \( n_{i}^{(j)} \) and
distance \( d_{il}^{(j)} \) between data \( l \) and 1st principal component, fitness \( J_{j} \) of individual
\( j \) is as follows:

\[
J_{j} = (1 - K) \sum_{i=1}^{m} \frac{\lambda_{i2}^{(j)}}{\lambda_{i1}^{(j)}} + K \sum_{i=1}^{m} \sum_{l=1}^{n_{i}^{(j)}} d_{il}^{(j)}
\]  

(3)

Using the above condition to separate data, \( \lambda_{i1} \) is larger then \( \lambda_{i2} \). Using the
above facts, the fitness of individual \( j \) should be highly scored when \( \lambda_{i2}^{(j)}/\lambda_{i1}^{(j)} \) is
a small value. And when the distance between data and 1st principal component
is a small value, the distribution of data has high density.

Parameter \( K \) of the function of fitness(3) is weigh in optimizing the shape
of the distribution of data or separating each of systems to have high density
distributions. The value of this parameter \( K \) depends on a problem.

Otherwise, parameter \( K \) is set to a small value because \( \lambda_{i2}^{(j)}/\lambda_{i1}^{(j)} \), which shows
the shape of the distribution of data, is less than 1.

We employ a fuzzy regression model to build a switching regression model as
we mentioned above. Using the evaluation function(3), we separate data to each
of systems by a genetic algorithm. Each system is analyzed by a fuzzy regression
model.

4 Possibilistic Autocorrelation Model

In the fuzzy autocorrelation model, time-series data \( z_{t} \) are transformed into a
fuzzy number \( Y_{t} \) to express the possibility of data[5].

Next, we employ a calculus of finite differences to filter out the time-series
data of trend. It enables us first-order difference-equation to write the following:

\[
T_{t} = (T_{t}^{L}, T_{t}^{C}, T_{t}^{U}) = (\min(Y_{t} - Y_{t-1}, Y_{t}^{C} - Y_{t-1}^{C}, \max(Y_{t} - Y_{t-1}))
\]  

(4)

Generally, if we take finite differences then we reduce the trend variation, and
only an irregular pattern is included in the difference series. However, when we
calculate the fuzzy operation, the ambiguity may become large and the value of
an autocorrelation coefficient may also take 1 or more and -1 or less value. In or-
der to solve this problem in the case of the fuzzy operation, we adjust the width
of a fuzzy number using \( \alpha \)-cut when we calculate the difference series. An \( \alpha \)-cut
level \( h \) is decided from the value of the autocorrelation. When we calculate the
fuzzy autocorrelation, we employ usual fuzzy operation under condition that the
fuzzy autocorrelation of lag 0 is set \( \rho_{0} = (1, 1, 1) \). It results in the following linear
programming to decide the value at the $\alpha$-cut level. When we set $\alpha$-cut level to 1, the ambiguity of fuzzy autocorrelation is made the smallest, but we can not obtain the fuzzy autocorrelation which reflects the possibility of the system. So, we maximize the width of autocorrelation. However, the size of width is decided automatically as the value of autocorrelation should be include in $[-1,1]$.

$$\maximize \sum_{i} (\rho_{i}^U - \rho_{i}^L)$$
subject to $-1 \leq \rho_{i}^L, \rho_{i}^U \leq 1$
$$\rho_{i}^L \leq \rho_{i}^C \leq \rho_{i}^U$$
$(i = 1, 2, \cdots, p)$

(5)

We can define the fuzzy covariance and the fuzzy autocorrelation as follows:

$$
A_k \equiv Cov[T_t T_{t-k}] = [\lambda_k^L, \lambda_k^C, \lambda_k^U], \quad r_k = A_k/A_0 = [\rho_k^L, \rho_k^C, \rho_k^U]
$$

We adjust the ambiguity of the difference series employing the $\alpha$-cut level $h$ which is obtained by solving the above linear programming. Using fuzzy autocorrelation coefficient which is calculated by employing $\alpha$-cut level $h$, we redefine Yule-Walker equations as in linear programming, and calculate the partial autocorrelation.

We calculate the following autoregressive process.

$$T_t = \Phi_1 T_{t-1} + \Phi_2 T_{t-2} + \cdots + \Phi_p T_{t-p}$$

where $\Phi = [\phi^L, \phi^C, \phi^U]$ is a fuzzy partial autoregressive coefficient.

As mentioned above, the next value of observation value exceeds observed value at present by the size of the value of autocorrelation, or it is less. For this reason, autocorrelation is important for the time-series analysis. So, we build the model which illustrates ambiguity of the system shown by fuzzy autocorrelation. The reason for the autocorrelation is also fuzzy autocorrelation, Yule-Walker equations can be also calculated by the fuzzy equation in the same way.

$$R_t = \Phi_1 r_{t-1} + \Phi_2 r_{t-2} + \cdots + \Phi_p r_{t-p}$$

(6)

$\Phi$ in Equation (6) is an unknown coefficient. We are building the model in terms of fuzzy autocorrelation which can describe the ambiguity of the system. However, when ambiguity of a model is large, the relation between a model and a system becomes ambiguous. Therefore, the possibility of the system can not be described properly. So, in order to obtain the fuzzy partial autocorrelation coefficient whose ambiguity of a time-series model should be minimized, we come down to the following linear programming:

$$\minimize \Phi \sum (\rho_{t}^U - \rho_{t}^L)$$
subject to $R_t^L \leq \rho_{t}^L, R_t^C = \rho_{t}^C, \rho_{t}^U \leq R_t^U$
$$\rho_{t}^L \leq \rho_{t}^C \leq \rho_{t}^U$$
$(t = 1, 2, \cdots, p)$

(7)
As mentioned above, $R$ is obtained by the fuzzy operation employing fuzzy autocorrelation $r$ and fuzzy partial autocorrelation $\Phi$. $R^L$, $C_R^C$, and $R^U$ represent the lower limit, the center, and the upper limit of $R$, respectively.

A fuzzy autocorrelation model expresses the possibility that the change of the system is realized in data, different from the conventional statistical method. We are building the model which can show an ambiguous portion called a possibility that it has not expressed clearly by the conventional statistics technique.

5 Possibilistic Switching AR Model

Sometimes, a time series system changes these state. In this Case, it is hard to interpret time series data as a single system by single model. In case of data ware observed from some systems or a state of system was changed, we shall employ some model for every system. In this paper, we propose possibilistic switching AR model that is recognize each system included in time series data and each system was interpreted by possibilistic AR model.

In this paper, we employ possibilistic autocorrelation model mentioned above. That is, possibility of system has is separate to each possibility, possibilistic autocorrelation model interpret possibility of time series system.

In this model, we employ genetic algorithm to separate each systems. Where, $E = \sum_{i=1}^{m} w_i$ is evaluate function, $w_i$ is width of $i$th data's possibility and $m$ is number of systems. A genetic algorithm separate each system to minimize the value of evaluates function that is fuzziness. And each system is interpreted by possibilistic autocorrelation model.

6 Numerical Example

6.1 A Consideration of Fluctuations in Prices of Japan

Let us discuss total price indexes as an application. The application, relative price indexes of 1934 through 1936 is set to 1 and a calendar year average.

As a result of analyzing, we obtained two fuzzy regression models as follows:

$$y_1 = -77.941 + (0.041, 0.000)x$$
$$y_2 = -94470.810 + (48.572, 0.026)x$$

Models shows the result of analyzing that data include two systems that is from 1925 to about 1945 and after about 1946. The fact that the price indexes keep stable until 1946, and seems not to have a large change. The fact that the price indexes precipitously rise after 1946 seems to express the effect of the 2nd world war.

The fact which an economical system is changed comes from a change of the before war to the after war separated two system into two. Our method could abstract two systems from observed data.
6.2 A Consideration of a Nikkei Stock Average

In this section, we employ Nikkei stock average that indicates the trend of the whole stock market as an index of Japanese stock market. We use the monthly data from 1970 to 1998. We obtained following possibilistic switching AR model by analyzing finite difference series.

Model 1: \( \nabla \nabla_6 X_t = [-0.3200, -0.0645, 0.3165] \nabla \nabla_6 X_{t-6} \)
Model 2: \( \nabla \nabla_6 X_t = [-0.0837, -0.0020, 0.1235] \nabla \nabla_6 X_{t-6} \)

In these model, switching point is June 1997 and value of a evaluate function is 0.844.

7 Conclusion

In this paper, We discussed a character of possibilistic forecasting model. We discussed possibilistic regression model that reduce an influences of outlier is included in data. And we illustrated possibilistic switching regression model that interpret each system of observed data from some systems.

Possibilistic time series model is discussed as a model to interpret the possibility of system has.

And possibilistic switching time series model was illustrated. Possibilistic switching time series model can interpret each possibility that are states changed or each data observed from some systems.

At last, we illustrate characters of possibilistic switching regression model and possibilistic switching AR model by analyzing numerical examples.

References

1. R. J. Hathaway and J. C. Bezdek, Switching Regression Models and Fuzzy Clustering, *IEEE Transactions on Fuzzy Systems*, Vol.1, No.3, pp.195–204, 1993
2. H. Ishibuchi and H. Tanaka, Interval Regression Analysis based on Mixed 0-1 Integer Programming problem, *Journal of Japan Industrial Management Association*, Vol.40, No.5, pp.312-319, 1988, in Japanese.
3. K. Ozawa, T. Watanabe and M. Kanke, Forecasting Fuzzy Times Series with Fuzzy AR Model, Proceedings of 20th International Conference on Computers & Industrial Engineering, in Kyongju, Korea, pp.105–108, 1996
4. M. Ryoke and Y. Nakamori, Simultaneous Analysis of Classification and Regression by Adaptive Fuzzy Clustering, *Journal of Japan Society for Fuzzy Theory and Systems*, Vol.8, No.1, pp.136–146, 1996, in Japanese.
5. J. Watada, Fuzzy Time-series Analysis and Its Forecasting of Sales Volume, *Fuzzy Regression Analysis*, edited by J. Kacprzyk & M. Fedrizzi, pp.211–227, 1992
6. Y. Yabuuchi and J. Watada, Fuzzy Robust Regression Analysis based on A Hyperelliptic Function, Proceedings of the 4th IEEE International Conference on Fuzzy Systems, pp.1841–1848, 1995
7. Y. Yabuuchi and J. Watada, Fuzzy Switching Regression Model based on Genetic Algorithm, *Proceedings of the 7th International Fuzzy Systems Association World Congress*, in Prague, Czech Republic, pp.113–118, 1997

8. Y. Yabuuchi, Y. Toyoura and J. Watada, Fuzzy AR Model of Stock Price, *Proceedings of 5th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*, Mt. Koya, Japan, pp.127–132, 2002

9. Y. Yabuuchi, J. Watada and Y. Toyoura, Fuzzy Switching AR Model, *Proceedings of the 19th Fuzzy System Symposium*, pp.697-698, 2003, in Japanese.
A Proposal of Chaotic Forecasting Method Based on Wavelet Transform

Yoshiyuki Matsumoto¹ and Junzo Watada²

¹ Shimonoseki City University, 2-1-1 Daigakumachi, Shimonoseki, Yamaguchi 751-8510, Japan
² Waseda University, 2-2, Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0135, Japan

Abstract. Recently, the chaotic method is employed to forecast a short-term future using uncertain data. This method makes it possible to restructure the attractor of given time-series data in the multi-dimensional space through Takens’ embedding theory.

However, some time-series data have less chaotic characteristic. In this paper, Time-series data are divided using Wavelet Transform. It will be shown that the divided orthogonal elements of time-series data are employed to forecast more precisely than original time-series data. The divided orthogonal time-series data are forecasted using Chaos method. Forecasted data are restored to the original data by inverse wavelet transform.

1 Introduction

The chaotic short-term forecasting method [1][2] based on time-series data enables us to know a future value, which we could not predict before. Nevertheless, it is still difficult to definitely forecast a value even in near future because many kinds of data have less chaotic nature. Even though such data are less chaotic, it is possible to abstract and pull out the partially chaotic portion out of the data [3][4][5].

In this research, time-series data are divided into parts using wavelet transform[6]. It will be shown that the divided orthogonal elements of time-series data are employed to forecast more precisely than the original time-series data. Forecasted data are rebuilt into the original form by inverse wavelet transform. A forecasted result is compared among various kinds of mother wavelet functions.

2 Forecasting by Chaotic Method

The general objective to employ a chaotic method in forecasting is 1) to find a deterministic structure in given time-series data and 2) to predict a value in such a near future from a certain point using this structure on that the present state can sufficiently influence.

The chaotic method enables us to forecast its near future with high precision using time-series data that show very unpredictable and un-periodical changes.
This forecasting bases on the Takens’ embedding theory [7] which tells us that it is possible to restructure the trajectory of a dynamic system in a high dimensional space by using only the information (that is, time-series data) of partial component dimensions or variables.

Using time-series data \( x(t) \), let us define vector \( Z(t) \) as follows:

\[
Z(t) = (x(t), x(t-\tau), x(t-2\tau), \ldots, x(t-(n-1)\tau))
\]  

(1)

Where \( \tau \) denotes an arbitrary constant time interval. The vector \( Z(t) \) shows one point in \( n \) dimensional space (Data Space). Therefore, changing \( t \) generates a trajectory in the \( n \) dimensional data space. When \( n \) is sufficiently large, this trajectory shows a smoothly changed one of a multi-dimensional dynamic system. That is, if the dynamic system has some attractor, the attractor obtained from the original one should come out on the data space. In other words, the original attractor of the dynamic system can be embedded in the \( n \) dimensional phase space. Number \( n \) is named an embedded dimension. Denoting the dimension of the original dynamic system by \( m \), it can be proved that this dimension \( n \) should be sufficiently large such as \( n \) holds the following:

\[
n \geq 2m + 1
\]

(2)

Equation (2) is a sufficient condition on the embedded dimension. Therefore, it is required to employ data with more than \( 3m+1 \) to \( 4m+1 \) samples within a certain time length in short-term forecasting.

3 Wavelet Transformation

Fast Fourier Transform is a widely employed method to transform signal into the portions of each frequencies. A sin function is employed as a base function. The sin function is an infinitive smooth function. Therefore, the information obtained by the Fast Fourier Transform does not include the local information such as the place and the frequency where and which frequencies the original signals have.

On the other hand the wavelet transform employs a compact portion of a wavelet as a base function. Therefore, it is a time and frequency analysis such as it enables us to determine the signal using time and frequency.

The mother wavelet transform \( \psi(x) \) of a function \( f(x) \) can be defined as follows:

\[
(W_{\psi f})(b, a) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{|a|}} \overline{\psi\left(\frac{x-b}{a}\right)} f(x) dx
\]

(3)

where “\( a \)” is a scale of the wavelet, “\( b \)” is a translate. \( \overline{\psi(x)} \) is a conjugate of a complex number. It is also possible to recover the original signal \( f(x) \) using inverse wavelet transform. That is, the inverse wavelet transform can be written as follows:
The wavelet transform is a useful method to know the characteristics of the signal but not an efficient one. It is partly because the signal has a minimum unit and the wavelet method expresses many-duplicated information’s. This point can be resolved by discrediting a dimensional axis. Let us denote a dimension as \((b,1/a)=(2^{-j}k,2^j)\), then the discrete wavelet transform can be rewritten as

\[
d_k^{(j)} = 2^j \int_{-\infty}^{\infty} \psi(2^j x - k) f(x) dx
\]  

(5)

4 Moving Wavelet Transform

When time-series data are analyzed on the basis of wavelet transform, the number of employable samples should decrease such as \(1/2\) of the number of the original samples for \(j=-1\) and \(1/4\) for \(j=-2\), respectively. Accordingly the time interval should be expanded to the 2 times and 4 times wider, respectively. This is because the wavelet transform produces lower frequency components according that \(j\) becomes smaller. This situation makes it difficult to inverse-transform the forecasted components of wavelet transform into the original time-series form.

![Fig. 1. Moving Wavelet Transform](image)

This paper proposes a method to move the starting and terminated points and wavelet-transform them so as to obtain a value at the same focal future time-point.
This wavelet transform is named a moving wavelet transform in this paper. This method makes us easily inverse-wavelet transform forecasted component values for each \( j \) at the next future time-point into the original time-series. For example, let us wavelet transform original time-series data until \( j=-3 \) (Fig. 1). In this case, the number of employable samples for \( j=-1 \) should be 1/2 of the total number of samples and one for \( j=-2 \) is 1/4 and one for \( j=-3 \) is 1/8. If we move the starting point one by one, we can interpolate the lacked samples or components that should be inverse-wavelet-transformed and calculate the inversed forecasted value.

5 Forecast Based on Inverse-Wavelet-Transform

This section will be spent to explain the short-term forecast based on inverse wavelet transform. The wavelet transform decomposes time-series data into each frequency band. Decomposed time-series data is more convenient to forecast the short-term future comparing the original time-series data. Therefore, it is much reasonable to forecast the short-term future using components decomposed by wavelet transform and to inverse wavelet transform these forecasted components values into the original time-series value which means the short-time forecasted value of the original time-series data. This method enables us to forecast the short-term future based on wavelet transform with more precision than on the original time-series data.

Basically, the chaotic forecast is to evaluate a value in the one term future from the present. Nevertheless, the decomposing by wavelet transform divides the original data into each of frequency bunds. This brings out that the length \( \tau \) of one term for each component should be different. Therefore, it is impossible to inverse-wavelet-transform the forecasted component values into the original value. We should employ moving wavelet transform to bring the forecasted component values at one term future and using these forecasted component values at the one term future can be inverse-wavelet-transformed into the original time-series.

6 A Kind of Mother Wavelet Function to Employ

In this research, several mother wavelet functions are employed.

1) Haar Wavelet
   It is the easiest wavelet function made by Haar. The Haar wavelet has a dis-continuous part, and looks like a step function.

2) Daubechies Wavelet
   It is an orthonormal wavelet function responded by Daubechies. It appeared as wavelet which was continuous and compact support for the first time. It is numbering by natural number \( N \). It becomes smooth as much as to increase \( N \).

3) Symlet Wavelet
   Daubechies Wavelet is not symmetrical. The Symlet wavelet is only nearly symmetrical. This function was built based on Daubechies.
4) Coiflet Wavelet

Coiflet Wavelet have the same number of vanishing moments for both the scaling functions and the mother wavelets.

5) Spline Wavelet

Cardinal B-Spline Wavelet is satisfying a two-scale relation. But this function is not orthogonal to itself. So orthogonal dual basis is used in necessity. This system is called dual orthogonal. Compact support B-Spline Wavelet was proposed by Chui and Wang.

We examined the effect of difference among various wavelet functions in chaotic forecasting.

7 Forecast Results

In this section, the movement of TOPIX was forecasted by using a proposed method. The number of the data employed is 1125 samples from January 1991. 1025 were used as the past data, a forecast was done the values for the rest 100 samples. We forecasted stock prices on the following day. As data a ratio par its previous day is used. The B-Spline function is employed as a mother wavelet function. In this case,

![Chart showing forecast results](image)

**Fig. 2.** Forecast Results

**Table 1.** Forecast Results

|       | Proposed model | Conventional model |
|-------|----------------|--------------------|
| Minimum | 0.988535       | 0.980187           |
| Average | 1.020612       | 1.050553           |
| Maximum | 1.042376       | 1.136912           |
the conventional method and this method. Forecast results are shown in the Fig. 2 and Table 1. The error of proposed method is better than the conventional method on the average.

This result was obtained divided time series data in the completely same dimension. The optimum forecasting dimension is different from the divided data. So, it should be forecasted in search of the optimum dimension for the divided time series data again. Forecasted results are shown in the Fig. 3 and Table 2. The forecasted error of the proposal model is better in all the dimensions.

![Diagram showing forecast errors](image)

**Fig. 3.** Forecast Results (Optimization)

| Embedding Dimension | Proposed model | Conventional model |
|---------------------|----------------|--------------------|
| 3                   | 0.96083        | 0.980187           |
| 4                   | 0.96970        | 1.050553           |
| 5                   | 0.96970        | 1.050553           |
| 6                   | 0.97921        | 1.136912           |
| 7                   | 0.97921        | 1.136912           |
| 8                   |                |                    |
| 9                   |                |                    |
| 10                  |                |                    |

**Table 2.** Forecast Results (Optimization)

8 Concluding Remarks

In this paper, we proposed chaotic forecasting on moving wavelet transform. We simulated our method about the forecasting of TOPIX. The error of proposal method is smaller than usual method.

Acknowledgements

We express our appreciation to Mr. Shinju Tomozuka, Director, Fuji Investment Management Co. Ltd for his valuable comments and advices on this research.
A Proposal of Chaotic Forecasting Method Based on Wavelet Transform

References

1. Matsuba: “Chaos and forecast”, Mathematical Sciences, No.348, pp.64-69, 1992, in Japanese.
2. T. Nagashima, Y. Nagai, T. Ogiwara, T. Tsuchiya: “Time series data analysis and Chaos”, Sensing Instrument Control Engineering, Vol.29, No.9, 1990, in Japanese.
3. Y. Matsumoto, J. Watada: “Short-term Prediction by Chaos Method of Embedding Related Data at Same Time”, Journal of Japan Industrial Management Association, Vol.49, No.4, pp.209-217, 1998, in Japanese.
4. Y. Matsumoto, J. Watada: “Application of Chaotic Short-term Forecast to Economics and Business Problem”, Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications, pp.219-225, 1998, Vietnam
5. Y. Matsumoto, J. Watada: “Chaotic Short-term Forecasting on Up & Down Movements of Stock Prices”, 2nd International Symposium on Advanced Intelligent Systems, pp287-291, 2001, Korea
6. C.K. Chui: “Introduction to wavelets”, Academic Press, New York, 1992.
7. F. Takens: “Detecting Strange Attractors in Turbulence,” in Dynamical Systems and Turbulence, ed. by D.A.Rand and L.S.Young, Lecture Notes in Mathematics, vol.898 (Springer-Verlag, Berlin), pp.366-381, 1981
8. Junzo Watada: “Fuzzy System Methods in Financial Engineering,” in Special Issue on Advanced Methods for Fuzzy Systems, Official Journal: Systems, Control And Information, Vol. 48, No. 1, pp. 28-33, 2004
Fuzzy Multivariant Analysis*,**

Junzo Watada¹, Masato Takagi¹, and Jaeseok Choi²

¹ Waseda University, Graduate School of Information, Production and Systems
² Gyeonsang National University, School of Electric Engineering

Abstract. In this paper we will analyze data obtained from a fuzzy set, where samples are defined using a membership grade of a fuzzy set. Generally, multivariant model is to analyze samples characterized using plural variants. In this paper such samples are also characterized by a fuzzy set. Our aim is to build fuzzy discriminant model and fuzzy pattern classification model.

1 Introduction

In this paper we will analyze data obtained from a fuzzy set, where samples are defined using a membership grade of a fuzzy set. Generally, multivariant model is to analyze samples characterized using plural variants. In this paper such samples are also characterized by a fuzzy set. Our aim is to build fuzzy discriminant model and fuzzy pattern classification model.

2 Fuzzy Group

We will deal with fuzzy data and fuzzy events that provides the basis for Fuzzy Multivariant Analysis. Since the sample sets are commonly called groups in multivariate analysis, we will call the fuzzy sets of the samples here fuzzy groups.

In order to review that definition, we will rewrite probability \( P(A) \) of the fuzzy event determined by fuzzy set \( A \) over \( n \)th dimensional interval \( R_n \), which is defined by degree of probability \( P \) according L.A. Zadeh [5]:

\[
P(A) = \int_{R_n} \mu_A(x) dP = E(\mu_A).
\]

(1)

Here, \( E(\mu_A) \) is the expected value of membership function \( \mu_A \).

* This article is presented at KES2004 on September 20-25, 2004, in New Zealand.
** All inquiries can be sent to Junzo Watada, Professor / Waseda University / Graduate School of Information, Production and Systems / 2-2, Hibikino Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan / e-mail: junzow@osb.att.ne.jp
Using equation 1, the fuzzy mean and fuzzy variance for variable $x$ can be calculated as follows:

\[
m_A = \frac{1}{P(A)} \left\{ \int_{R_A} x \mu_A(x) dP \right\}
\]

\[
\sigma^2_A = \frac{1}{P(A)} \left\{ \int_{R_A} (x - m_A)^2 \mu_A(x) dP \right\}.
\]

3 Fuzzy Discriminant Analysis

The object of Fuzzy Discriminant Analysis as proposed in Watada et al. [2] is to express several fuzzy groups in terms of variables. These variables take the form of values (membership values) on [0,1].

Table 1. Data Handled by Fuzzy Discriminant Analysis

| No. | Fuzzy Group | Category |
|-----|-------------|----------|
| $\omega$ | $B_1, \ldots, B_M$ | $A_1, \ldots, A_i, \ldots, A_K$ |
| 1 | $\mu_{B_1}(1) \cdots \mu_{B_M}(1)$ | $\mu_1(1) \cdots \mu_i(1) \cdots \mu_K(1)$ |
| 2 | $\mu_{B_1}(2) \cdots \mu_{B_M}(2)$ | $\mu_1(2) \cdots \mu_i(2) \cdots \mu_K(1)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $\omega$ | $\mu_{B_1}(\omega) \cdots \mu_{B_M}(\omega)$ | $\mu_1(\omega) \cdots \mu_i(\omega) \cdots \mu_K(\omega)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $\mu_{B_1}(n) \cdots \mu_{B_M}(n)$ | $\mu_1(n) \cdots \mu_i(n) \cdots \mu_K(n)$ |

Table 1 shows the data handled by Fuzzy Discriminant Analysis. The object of Fuzzy Discriminant Analysis is to express, as well as possible using the linear equation of weight $a_i$, of $A_1$, the discrimination of fuzzy groups on the real value axis:

\[
y(\omega) = \sum_{i=1}^{K} a_i \mu_i(\omega); \quad \omega = 1, \ldots, n.
\]

In other words, it is determining the $a_i$ that gives the best separation of the fuzzy groups on the real value axis.

The degree of separation of the fuzzy groups is defined as the fuzzy variance ratio $\eta^2$, which is the ratio of the total variation $T$ and variation between fuzzy groups $B$:

\[
\eta^2 = \frac{B}{T}.
\]

In the following we will determine $a_i$ for linear equation 4, which maximizes fuzzy variance ratio $\eta^2$. 

The fuzzy mean \( \bar{y}_{Br} \) within fuzzy group \( Br \) for value \( y(\omega) \) for the linear equation and total fuzzy mean \( \bar{y} \) come out as follows:

\[
\bar{y}_{Br} = \frac{1}{N(B_r)} \left\{ \sum_{\omega=1}^{n} y(\omega) \mu_{B_r}(\omega) \right\}; \quad r = 1, \ldots, M.
\]

\[
\bar{y} = \frac{1}{N} \left\{ \sum_{r=1}^{M} \bar{y}_{Br} N(B_r) \right\}
\]

Fuzzy mean \( \bar{\mu}_i^r \) within each fuzzy group \( B_r \) for the membership value of \( A_i \) and total fuzzy mean \( \bar{\mu}_i \) are expressed as follows:

\[
\bar{\mu}_i^r = \frac{1}{N(B_r)} \left\{ \sum_{\omega=1}^{n} \mu_i(\omega) \mu_{B_r}(\omega) \right\}; \quad i = 1, \ldots, K, \quad r = 1, \ldots, M
\]

\[
\bar{\mu}_i = \frac{1}{N} \left\{ \sum_{r=1}^{M} \bar{\mu}_i^r N(B_r) \right\}; \quad i = 1, \ldots, K.
\]

In order to simplify the writing, the \((Mn, K)\) matrices \( A, \tilde{A}_G, \) and \( \tilde{A} \) for \( \mu_i(\omega), \bar{\mu}_i^r, \) and \( \bar{\mu}_i \) are defined as follows:

\[
A = \left[ \mu_i(\omega) \right];
\]

\[
\tilde{A}_G = \begin{bmatrix}
\bar{\mu}_1^1 & \ldots & \bar{\mu}_i^1 & \ldots & \bar{\mu}_K^1 \\
\vdots & \ddots & \vdots \\
\bar{\mu}_1^i & \ldots & \bar{\mu}_i^i & \ldots & \bar{\mu}_K^i \\
\vdots & \ddots & \vdots \\
\bar{\mu}_1^K & \ldots & \bar{\mu}_i^K & \ldots & \bar{\mu}_K^K
\end{bmatrix};
\]

\[
\tilde{A} = \begin{bmatrix}
\bar{\mu}_1 & \ldots & \bar{\mu}_i & \ldots & \bar{\mu}_K \\
\vdots & \ddots & \vdots \\
\bar{\mu}_1 & \ldots & \bar{\mu}_i & \ldots & \bar{\mu}_K \\
\vdots & \ddots & \vdots \\
\bar{\mu}_1 & \ldots & \bar{\mu}_i & \ldots & \bar{\mu}_K
\end{bmatrix}
\]

In addition, the \( K \) dimensional row vector \( a \) for weight \( a_i \) and the \((M_n, M_n)\) diagonal matrix \( G \) formed from membership value \( \mu_B \) are defined as follows:
\[ A = [A_1, \cdots, A_i, \cdots, A_K]; \]

\[
\bar{G} = \begin{bmatrix}
\mu_{B_1}(1) & \cdots & 0 \\
\cdots & \ddots & \cdots \\
0 & \cdots & \mu_{B_M}(n)
\end{bmatrix}.
\]

Using these, the total variation \( T \) and variation between fuzzy groups \( B \) from We have the following relations:

\[ T = a'(A - \bar{A})'G(A - \bar{A})a \quad (6) \]
\[ B = a'(\bar{A}_G - \bar{A})'G(\bar{A}_G - \bar{A})a. \quad (7) \]

We obtain the following relationship if we substitute Equations 6 and 7 for Equation 5 and partially differentiate by \( a \):

\[ G^{1/2}(\bar{A}_G - \bar{A})'G^{1/2}(\bar{A}_G - \bar{A})a = \eta^2 G^{1/2}(\bar{A}_G - \bar{A})'G^{1/2}(\bar{A}_G - \bar{A})a. \quad (8) \]

If we now define \( S_G \) and \( S \) by the \((K, K)\) matrix as

\[ S_G = \{G^{1/2}(\bar{A}_G - \bar{A})\}'\{G^{1/2}(\bar{A}_G - \bar{A})\} \]
\[ S = \{G^{1/2}(\bar{A}_G - \bar{A})\}'\{G^{1/2}(\bar{A}_G - \bar{A})\}. \]

We can decompose \( S \) using triangular matrix \( \Delta \) as \( S = \Delta'^{-1}\Delta \), and we get the following:

\[ [(\Delta')^{-1}S_G\Delta'^{-1}]\Delta a = \eta^2 \Delta a. \]

Because of this, \( C a \) for Equation 4, which maximizes fuzzy variance ratio \( \eta^2 \), can be obtained from eigenvector \( \Delta a \), which maximizes eigenvalue \( \eta^2 \) of the matrix \([(\Delta')^{-1}S_G\Delta'^{-1}] \).

### 4 Fuzzy Dual Scaling

Fuzzy Dual Scaling is a method in which pattern classification is done; similar methods were developed independently in various countries. They are known by different names such as dual scaling, correspondence analysis, pattern classification, etc., and are described in detail in a paper by Tenenhaus and Young and in a book by Lebart, et al.

Fuzzy Dual Scaling is a method in which, if samples are taken from young people, we think of these samples as being elements of fuzzy set \( B, \) “young people,” and attempt quantitatively to classify each sample \( \omega \) and category by considering the membership values of this fuzzy set. In this case, the response to
each category is given a degree of attribution, not on \( \{0, 1\} \) but on \([0,1]\), and we determine whether the response pattern for each sample differs. Especially when the response to the categories are given by \( \{0, 1\} \), we can think of the frequency of the patterns as the degree of attribution.

Table 2. Data Handled by Fuzzy Dual Scaling

| No. | Fuzzy Group \( B \) | Category \( 1 \cdots i \cdots u_K \) | Total |
|-----|---------------------|---------------------------------|-------|
| \( v_1 \) | \( \mu_B(1) \) | \( \mu_1(1) \cdots \mu_i(1) \cdots \mu_K(1) \) | \( \mu_1 \) |
| \( v_2 \) | \( \mu_B(2) \) | \( \mu_1(2) \cdots \mu_i(2) \cdots \mu_K(1) \) | \( \mu_2 \) |
| ... | ... | ... | ... |
| \( v_\omega \) | \( \mu_B(\omega) \) | \( \mu_1(\omega) \cdots \mu_i(\omega) \cdots \mu_K(\omega) \) | \( \mu_\omega \) |
| ... | ... | ... | ... |
| \( v_n \) | \( \mu_B(n) \) | \( \mu_1(n) \cdots \mu_i(n) \cdots \mu_K(n) \) | \( \mu_n \) |

Table 2 shows a generalization of the data handled by Fuzzy Dual Scaling. The object of Fuzzy Dual Scaling is to give a mutually close numerical value to the samples resembling reactions, when real numerical values \( \nu_\omega \) and \( u_i \) are assigned to samples \( \omega \) and to the various categories \( i \), and at the same time to give a close numerical value to resembling categories. In order to do this, the correlation coefficient of the category is used as an indicator.

The size of reactions for all categories for sample \( \omega \) is defined as

\[
m_\omega = \sum_{j=1}^{K} \mu_i(\omega)\mu_B(\omega).\]

Samples with a large membership value to fuzzy set \( B \) are evaluated highly in this analysis, but those with low ones are not given much consideration. As a result, the reaction for all of the data is defined as the product of the membership values to fuzzy set \( B \) and the above equation:

\[
T = \sum_{\omega=1}^{n} m_\omega \mu_B(\omega).
\]

Means \( u \) and \( v \), variances \( \sigma_u^2 \) and \( \sigma_v^2 \), and covariance \( \sigma_{uv} \) for numerical values \( u_i \) and \( v_\omega \), based on fuzzy event \( B \), come out as follows:

\[
\bar{u} = \frac{1}{T} \left\{ \sum_{\omega=1}^{n} \sum_{i=1}^{K} \mu_i(\omega)\mu_B(\omega)u_i \right\}
\]

\[
\bar{v} = \frac{1}{T} \left\{ \sum_{\omega=1}^{n} m_\omega \mu_B(\omega)v_\omega \right\}
\]
\[ \sigma^2_u = \frac{1}{T} \left\{ \sum_{\omega=1}^{n} \sum_{i=1}^{K} \mu_i(\omega) \mu_B(\omega) u_i^2 \right\} - \bar{u}^2 \]

\[ \sigma^2_v = \frac{1}{T} \left\{ \sum_{\omega=1}^{n} m_\omega \mu_B(\omega) v_\omega^2 \right\} - \bar{v}^2 \]

\[ \sigma_{uv} = \frac{1}{T} \left\{ \sum_{\omega=1}^{n} \sum_{i=1}^{K} \mu_i(\omega) \mu_B(\omega) u_i v_\omega \right\} - \bar{u} \bar{v}. \]

Based on the conditions that \( u = 0 \), and \( v = 0 \), the problem here is to determine the \( u_i \) and \( v_\omega \) that maximize correlation coefficient

\[ \rho = \frac{\sigma_{uv}}{\sqrt{\sigma^2_u \sigma^2_v}}. \]

A partial differentiation of correlation coefficients \( u_k \) and \( v_\tau \) gives us

\[ \frac{\partial \rho}{\partial u_k} = 0; \quad k = 1, \ldots, K \]

\[ \frac{\partial \rho}{\partial v_\tau} = 0; \quad \tau = 1, \ldots, n. \]

If we do these calculations, we get the following:

\[ \sum_{\omega=1}^{n} \mu_k(\omega) \mu_B(\omega) v_\omega = \rho \frac{\sigma_v}{\sigma_u} \sum_{\omega=1}^{n} \sum_{i=1}^{K} \mu_i(\omega) \mu_B(\omega) u_k \]

\[ \sum_{i=1}^{K} \mu_i(\tau) \mu_B(\tau) u_i = \rho \frac{\sigma_u}{\sigma_v} m_\tau \mu_B(\tau) v_\tau \]

Eliminating \( v_\tau \), we get

\[ \sum_{\omega=1}^{n} \sum_{i=1}^{K} \frac{\mu_B(\omega)}{m_\omega} \mu_k(\omega) \mu_i(\omega) u_i = \rho^2 \cdot \sum_{\tau=1}^{n} \mu_k(\tau) \mu_B(\tau) u_k; \quad k = 1, \ldots, K \quad (9) \]

In order to consolidate the above, we introduce the following notation:

\[ b_k = \sum_{\omega=1}^{n} \mu_k(\omega) \mu_B(\omega); \quad k = 1, \ldots, K \]

\[ z_k = \sqrt{b_k u_k}; \quad k = 1, \ldots, K \]

\[ c_{ki} = \frac{1}{\sqrt{b_k b_i}} \sum_{\omega=1}^{n} \frac{\mu_B(\omega)}{m_\omega} \mu_k(\omega) \mu_i(\omega); \quad k, i = 1, \ldots, K \]

\[ c = [c_{ki}] \]

\[ z' = [z_1, \ldots, z_K] \]
Using this, the matrix expression of equation 9 results in the following:

\[ b_k = \sum_{\omega=1}^{n} \mu_k(\omega)\mu_B(\omega); \quad k = 1, \ldots, K \]

\[ z_k = \sqrt{b_k u_k}; \quad k = 1, \ldots, K \]

\[ c_{ki} = \frac{1}{\sqrt{b_k b_i}} \sum_{\omega=1}^{n} \frac{\mu_B(\omega)}{m_\omega} \mu_k(\omega)\mu_i(\omega); \quad k, i = 1, \ldots, K \]

\[ c = [c_{ki}] \]

\[ z' = [z_1, \ldots, z_K] \]

\[ cz = p^2 z \]

In other words, under the conditions \( u = 0 \) and \( v = 0 \), all we have to do is to find the maximum values for the eigenvalue equation. In addition, if we solve it based on the condition that \( v_\tau \) is \((1/\rho \cdot \mu_u/\mu_v)=1\), we get

\[ v_\tau = \frac{1}{m_\tau} \left\{ \sum_{i=1}^{K} \mu_i(\tau)\mu_i \right\}. \]

References

[1] Watada, J., Tanaka, E., and Asai, K., “Fuzzy Quantification Type II,” *The Japanese Journal of Behaviormetrics (Behaviormetric Society of Japan)*, 9, 2, pp.24-32 (in Japanese).

[2] Watada, J., Tanaka, E., and Asai, K., “Analysis of Purchasing Factors Using Fuzzy Quantification Theory Type II,” *Journal of the Japan Industrial Management Association*, 32, 5, pp.51-65 (1981)(in Japanese).

[3] Watada, J., Tanaka, E., and Asai, K., “Fuzzy Quantification Theory Type T,” *The Japanese Journal of Behaviormetrics (Behaviormetric Society of Japan)*, 11, 1, pp.66-73 (1984).

[4] Zadeh, L. A., “Fuzzy Sets,” *Information and Control*, 8, pp.338-353 (1965).

[5] Zadeh, L. A., ”Probability Measures of Fuzzy Events,” *Journal of Mathematical Analyses and Applications*, 23, pp.421-427 (1968).
Using Coherent Semantic Subpaths to Derive Emergent Semantics

D.V. Sreenath¹, W.I. Grosky², and F. Fotouhi¹

¹ Department of Computer Science, Wayne State University, Detroit, MI 48202
² Department of Computer and Information Science, University of Michigan-Dearborn, MI 48128

Abstract. We contend that the author of a Web page cannot completely define that document’s semantics and that semantics emerge through use. Semantics is context-sensitive. User browsing paths over a multimedia collection provide, we believe, the necessary context to derive semantics. We attempt to use information from several modalities to improve retrieval using a single modality. We have developed and tested several algorithms to detect breakpoints and cluster Web pages into groups that exhibit uniform semantics. We have also used visual-keywords to improve textual-keyword based searches.

1 Introduction

The current Web was designed for human interaction and not for computer-computer interaction. The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [1]. The semantic Web will be an extended web of machine understandable information and automated services. This requires the explicit representation of the semantics of every Web resource. The first step towards this goal is to determine the semantics of Web pages. We contend that a Web page does not have a fixed semantics, but multiple semantics that vary over time. As users browse through this page, they contribute to its semantics. The path traversed by the user contributes to the semantics of the Web pages along the path. Our goal is to derive the emergent semantics of Web pages from the browsing paths of the users.

We contend that short sub-paths of a user’s browsing path through the web exhibit uniform semantics, and that these semantics can be captured, easily represented, and used to our advantage. The browsing path of a user contributes to the semantics of the pages traversed. For example, a professor of chemistry could author a Web page on ammonium nitrate, describing its chemical properties, its composition and its uses. One may assume that there is only one fixed semantics - chemicals. If there are users who have been browsing through a series of pages related to farming, crops, fertilizers, organic fertilizers, chemical fertilizers and then to the page on Ammonium Nitrate, then that user’s browsing path contributes the semantics: fertilizers. If there is another set of users who reached the
page after visiting Web pages on violence, making bombs, explosives, etc., then that browsing path contributes *chemical weapons* to the semantics. Yet another browsing path could be that of a student who reached the page looking for a graduate research assistantship under the professor. The goal of this research is to derive the emergent multiple semantics of the Web pages using the browsing paths of the users.

In [2], Web usage mining techniques are used to provide a scalable architecture to automate better personalization. In [3], path clustering techniques are used to cluster users with similar interests. Sclaroff [4] uses the singular-value decomposition as a dimension reduction technique for a text-image matrix. The terms closer to an image are weighted more than the terms farther away. The experiments in [4] combine textual and image features for image retrieval using latent semantic indexing. In [5], it is argued that that the images by themselves do not have any semantics, but they do when placing them in the context of other images and by user interaction. The images do have a meaning, but with the best object recognition technologies, it would not be possible to extract the semantics of the image to satisfy the user’s requirements. Image semantics are context sensitive. From an image database perspective, the semantics of an image can be extracted by interpreting the sequence of queries posed by various users.

In this research, user browsing paths will be used to derive emergent semantics of Web pages. The basic idea is to include textual features and image features of the Web pages along the path to build a feature-path (also called term-path) matrix. Several variations of normalizing techniques and rank-k approximations of the matrices (dimension reduction) were explored as part of the latent semantic analysis to discover the semantics.

This paper is organized as follows. In Section 2, we describe our approach to deriving emergent semantics. The various algorithms for detecting break points are discussed in Section 3. The implementation details, the architecture, design criteria and experimental results are in 4 and the conclusions in 5.

### 2 Multimedia Semantics

Recent research in the field of content-based image retrieval has focused largely on the syntactic structural composition of low-level image features. Images are also represented using some form of metadata to enable searching and indexing at the metadata level. From an image semantics perspective, the results from research that maps low-level features to high-level semantics are promising, but have their limitations. Our belief is that image semantics is subjective and cannot be derived from pure low-level image analysis. Image semantics is context-sensitive and evolves depending on its use. We propose an approach to derive image semantics using both the context in which it appears and image browsing paths. For a user, an image of a snow-covered mountain could be as peaceful and serene as a sun-soaked beach. We attempt to derive the semantics as an emergent property of the interaction of the user with the multimedia col-
lection, as he browses through a sequence of Web pages. We believe that just as
textual features help derive image semantics, visual features will help document
semantics.

We use the technique of latent semantic analysis to determine the semantics
of web pages, derived from their textual and image features. Zhao and Grosky[7]
have previously shown that latent semantic analysis can discover that certain sets
of different image features co-occur with the same textual annotation keywords.
In this technique, each document is represented in a reduced dimensional space,
where each dimension corresponds to a concept, each concept representing a
set of co-occurring keywords and image features. Resulting keyword searches for
Web pages become much more efficient after this transformation; more efficient
than using latent semantic analysis on just the keywords. The keywords (visual
and textual) are represented as rows and the paths as columns. Each path is rep-
resented by a multi-dimensional column vector, a point in a \( n \)-dimension space.
Each dimension represents a particular keyword: visual or textual. The value or
weight associated with each keyword represents the importance of that keyword
for that path. The importance of a keyword in a path is directly proportional
to the frequency of its occurrence in that path. In order to effectively represent
the visual and textual keywords for a document, the keyword-path matrix is
subjected to a process of weighting and normalization.

We focus on the JPEG images that are predominantly used in the Web and
we process these images in the compressed domain. The luminance component
\((Y)\) contains the grayscale and the chrominance components, while \((C_{red}\) and
\(C_{blue}\) contain the color information. The DC \((0,0)^{th}\) coefficient of the cosine
transforms of the \(8 \times 8\) block represents the average intensity value. We use the
DC coefficients of the all three components, \(Y\), \(C_b\) and \(C_r\), as the basis for our
visual keywords.

The motivation for choosing the visual keywords came from Taylor’s analysis
of Pollock’s paintings [8] where the fractal dimension is used to analyze these
paintings. This article discusses the fractal nature of all Pollock’s paintings. The
image space (2-dimensions) is divided into grid cells of size \(r\). For an image, the
fractal dimension \(D_0\) is computed by counting the number of features in each
cell. If \(N(r)\) is the number of cells containing a feature, then

\[
D_0 = \lim_{r \to 0} \frac{\log N(r)}{\log \left(\frac{1}{r}\right)}
\]  

(1)

For a finite number of cells, the \(\lim_{r \to 0}\) can be ignored. The number of neigh-
borors considered can be expanded to simulate the variable-size sliding window as
in the WALRUS system[9]. The benefit of our approach is that we are able to ex-
tract the features directly by reading the compressed domain image file without
any additional wavelet computational overhead. Also, the use of the deviation
helps in identifying a uniform region or an edge.

Each image is represented by 859 visual keywords, composed of three his-
tograms of the absolute DC coefficient values, three histograms of the deviations
of these values from their neighbors, the maximum absolute values, and the max-
imum deviation. There are a total of 6 histograms, 3 for the actual values and 3
for the deviation of these values from those of its 8 neighbors. For the deviation
features, the Y component is represented by 130 bins, $C_r$ and $C_b$ by 40 bins
each.

Every element in the keyword-path matrix is a number representing the raw
frequency of the occurrence of the keyword in that path. We compute the log
of the raw frequency $f_{i,j}$ to dampen the extreme values. The local weight $l_{i,j}$
applied to each raw frequency is given by Equation 2.

$$l_{i,j} = \log_2(1 + f_{i,j})$$  \hspace{1cm} (2)

From the earlier work by Zhao and Grosky [7], we continue to use the entropy
method for global weight. If $p_{i,j}$ is the probability of the occurrence of the
keyword located in row i, then global weight $g_i$ is given by Equation 3

$$g_i = 1 + \sum_j p_{i,j} \log_2(p_{i,j}) \log_2(Total \ Number \ of \ Paths)$$  \hspace{1cm} (3)

where $p_{i,j}$ is computed by

$$p_{i,j} = \frac{f_{i,j}}{\sum_j (f_{i,j})}$$  \hspace{1cm} (4)

The weighted frequency $wf_{i,j}$ is the product of local and global weights. The
elements of the matrix are scaled to ensure that the Euclidean norm of each
column is 1. Normalization ensures that each path in the collection has equal
importance regardless of its length. The normalized weighted frequency $nwf_{i,j}$
is given by

$$nwf_{i,j} = \frac{wf_{i,j}}{\sqrt{\sum_i (wf_{i,j})^2}}$$  \hspace{1cm} (5)

Using the reduced-rank vector space model in [10], we compare a query vector
q to the columns of the approximation $A_k$ to the term-by-path matrix $KP$.
Identifying the appropriate value of $k$ is not a trivial task. We have experimented
with fixed value for $k$ (100-150), 80% rule ( $k = r \times 0.8$ ), and slope ( $\sigma_k - \sigma_{k+1}$).

3 Coherent Semantic Subpaths

The semantic break points are identified along the user’s browsing path where
the semantics change appreciably. The sequence of pages that exhibit similar
semantics is clustered as a sub-path. We contend that the semantics of a web page
is derived from its context. The location of a web page in any path contributes to
its emergent semantics. Due to this dependency, the identification of a semantic
breakpoint in a long browsing path (with multiple semantics) is crucial. We have
used vantage points as reference in the vector space to compute the distance
between URLs in a path. Shapiro [11] had shown that significant improvements
can be achieved by the proper choice and location of multiple reference points.

To summarize, each URL in each path is placed as a query in the vector space
and the distance from each path is computed. Each path is treated as a vantage
point. The set of such distances is treated as a vector. Two such distance sets are used to compute the distance between two adjacent URLs to determine if there should be a breakpoint between the two. As a pessimistic approach, this algorithm is efficient when we expect few breakpoints in the paths. But when the number of breakpoints is expected to be large, it is efficient to identify all the breakpoints for all the paths before another iteration. A variation of this algorithm is our Greedy Vantage Point algorithm.

In the merging-concepts algorithm, every possible sub-path of the each path is predetermined. If there is a path \((abc)\), the sub-paths \((a)\), \((ab)\) and \((abc)\) are added. Only unique sub-paths are added. Also, in this algorithm, if there is no breakpoint, the two URLs are intersected. The intersection operation is simulated by having a temporary path in the database that is made of \((U_j, U_j - 1)\). The Link-distance algorithm takes all possible tuples of the relation between the two sets and computes the average of the minimum distance of each point in a set with points in the other set.

When a typical browsing path was analyzed, it was found that the semantics of Web pages sometimes deviate gradually from the starting topic. A typical threshold value is not acceptable in such situations. The idea is to use the Link Distance algorithm, but instead of identifying the breakpoint using a threshold, analyze all distances for a path and if the slope exceeds the cluster-threshold, then there is a breakpoint. Finally, the cluster-merge algorithm merges the URLs together when the distance between the URLs are within the threshold. This approach takes care of the semantic decay along the browsing path. The merging of two adjacent URLs simulates an intersect operation.

4 Experimental Results

The data for detecting breakpoints in long browsing paths consisted of 26 long paths with about 100 Web pages in each path. The initial keyword-path matrix consisted of 30,000 keywords. The plots represent the cosine distances of each URL along a path. The URLs are along the X-axis and the cosine distances along the Y-axis. We experimented with 32 long browsing paths with approximately 100 web pages each. Each one of these paths consisted of Web pages related to a high-level topic. Some of the topics concerned Arnold Schwarzenegger, Beckham, Bush, Carter, explosives, ferry crash, fertilizers, Gaza, hubble telescope, Soccer, Iraq, UN draft, Kobe, NATO, nitrate, Nobel, Saddam, war crimes, Mother Teresa and the World Cup. These topics were randomly chosen by several participating individuals. The 32000 keywords were reduced to 16250 keywords using WordNet. This was a one-time filtering process. All the remaining experiments used these 16250 keywords.

The first set of experiments was to identify the semantics using a query. A web page about fertilizers was used as a query. The idea was to identify all the paths in the database that are similar to this query. The X-axis in the plot represents the various topics in alphabetical order - arnold to world-cup. The concepts explosives(10), fertilizers(12) and nitrates(20) are relevant with higher
cosine distance values than the rest of the topics in the collection. Graphically, it is easier to identify the topics, but, this can be missed easily when a threshold value is used. Instead of using a single URL as a query, the entire path (5 URLs) leading to the previous URL on fertilizers was used to derive emergent semantics. It can be clearly seen that a browsing path captures the context and is able to discriminate between the explosives and fertilizers. The fertilizers and nitrates match the query better than explosives. The static and emergent semantics derived using the long browsing paths is compared in Figure 1. The improvement of a path query over a single URL can be clearly seen for the two emergent concepts.

The 32 long browsing paths (100 URLs each) were then processed using our breakpoint algorithms. This resulted in 278 uniform coherent subpaths. This database of 278 subpaths was used to derive the static and emergent semantics using the fertilizer query. In Figure 2 we can clearly see the improved matching cosine values for the concepts represented as coherent semantic subpaths: explosives(81), fertilizers(136) and nitrate(220). The cosine values 0.72, 0.62 and 0.81 respectively, are high enough to be used in conjunction with threshold values. It is to be noted that the cosine value for fertilizers is actually lower than that of explosives or nitrates. The browsing path of fertilizers was used as a query to derive the emergent semantics using the coherent semantic subpaths. The results are in Figure 3. It can be seen that by using the path, one is able to use the context to clearly identify the emergent semantics of the browsing path query. The distance values 0.51, 0.70 and 0.55 respectively, clearly show that the emergent semantics is that of fertilizers. The necessity to use semantic coherent subpaths (breakpoint analysis) is highlighted by comparing Figures 2 and 3.

![Fig. 1. URL vs Path](image1)

![Fig. 2. Static semantics](image2)

![Fig. 3. Emergent semantics](image3)

5 Conclusions

We have successfully implemented a platform-independent, modular Java-based system for analyzing Web browsing paths. From the various experiments, we can infer the following: The technique for deriving emerging semantics from Web browsing paths works. Multiple semantics of a Web page or a path can be derived. Using the multiple semantics of a Web page, it is possible to detect the intention of a browsing path. Using a path as a query instead of a single Web page improves the retrieval precision of relevant documents. It is necessary
to identify the semantic breakpoints and break a long path into sub-paths that exhibit coherent uniform semantics. The breakpoint algorithms can also be used to eliminate irrelevant pages in the middle of a path. The breakpoint algorithms can be used as a generic clustering technique, to cluster Web pages from different domains into topics. The use of visual keywords does aid textual keywords. The use of visual keywords from images in a path helps the textual keywords only if the images themselves exhibit uniform semantics. It is well known that image features are not image semantics - it depends on the context. We would need some other metadata for the images, that could be used to identify the concept. Our research can be used to provide that metadata for images that co-occur with pages related to the same topic. The choice of vantage points and number of vantage points is critical in deriving emergent semantics. The number of vantage points directly affects the scalability of our approach.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” Scientific American, 2001.
2. B. Mobasher, R. Cooley, and J. Srivastava, “Automatic personalization based on web usage mining - web usage mining can help improve the scalability, accuracy, and flexibility of recommender systems,” Communications of the ACM, vol. 43, no. 8, pp. 142–151, 2000.
3. C. Shahabi, A. M. Zarkesh, J. Adibit, and V. Shah, “Knowledge discovery from users web-page navigation,” Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97), pp. 20–30, 1997.
4. S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher, “Unifying textual and visual cues for content-based image retrieval on the world wide web,” Computer Vision and Image Understanding, vol. 75, no. 1-2, pp. 86–98, 1999.
5. Z. Pecenovic, M. N. Do, and M. Vetterli, “Integrated browsing and searching of large image collections,” Advances in Visual Information Systems, Proceedings, vol. 1929, pp. 279–289, 2000.
6. S. Santini, A. Gupta, and R. Jain, “Emergent semantics through interaction in image databases,” IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 3, pp. 337–351, 2001.
7. R. Zhao and W. I. Grosky, “Narrowing the semantic gap - improved text-based web document retrieval using visual features,” IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 189–200, 2002.
8. Taylor, “Order in pollock’s chaos,” Scientific American, 2002.
9. A. Natsev, R. Rastogi, and K. Shim, “Walrus: A similarity retrieval algorithm for image databases,” Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 395–406, 1999.
10. M. W. Berry, Z. Drmac, and E. R. Jessup, “Matrices, vector spaces, and information retrieval,” Siam Review, vol. 41, no. 2, pp. 335–362, 1999.
11. M. Shapiro, “The choice of reference points in best file searching,” Communications of the ACM, 20(5), pp. 339–343, 1977.
Retrieval of Product Reputations from the WWW

Takahiro Hayashi, Yosuke Kinosita, and Rikio Onai

University of Electro-Communications,
Department of Computer Science, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan
{haya, kinoshita}@seman.cs.uec.ac.jp, onai@cs.uec.ac.jp

Abstract. We present a system for retrieving web pages which contain product reputations. We assume that product reputations are contained in personal pages and message board pages, and propose the system to retrieve those web pages. We illustrate that the system can distinguish those web pages automatically from other pages using six kinds of features extracted from a web page. We have evaluated the performance of the proposed system by computing precision and recall measures.

1 Introduction

It is well known that product reputations influence people who buy the product [1][2]. The WWW is a huge resource from which a user can obtain product reputations. A merit of searching the WWW is that a user can obtain various product reputations written by many people. Especially a user can obtain reputations about new products or rare products. It is difficult for a user to obtain such reputations by word-of-mouth because few close people (his family, friends, coworkers, etc.) usually have these products. A search engine is utilized for retrieving product reputation from the WWW. When a user gives a product name as a query to the search engine, the user is provided many web pages which contain various kinds of information about the product. Some of the web pages contain product reputations, however, many other web pages do not contain product reputations. Therefore, it is very costly for a user to find product reputations among many web pages which a search engine has provided. The purpose of this paper is to propose and evaluate a system for retrieving product reputations effectively from the WWW.

In the next section, related works are described. Section 3 introduces some definitions, describes the components of the proposed system, and explains the retrieval process used in the system. Section 4 describes experiments which have conducted to evaluate the performance of the proposed system. Finally, Section 5 concludes the paper.

2 Related Work

Morinaga et.al.[3] have proposed a system for retrieving product reputations from the WWW. They have assumed that product reputations are contained in web pages which include characteristic expressions. The characteristic expressions depend on the...
product domains. In personal computer domain, for example, the characteristic expressions are “fast”, “lightweight”, “unstable” and so on. However, it is impossible for their system to expect beforehand what kind of product a user is interested in. On the other hand, our system proposed in this paper adopts domain-independent approach. This is a big difference between the two systems.

Various methods for retrieval of product reputation using collaborative filtering have been proposed. An example is the method used in Amazon\(^1\)\[^4\]. Using the method, many product reputations can be collected from many users if these products are familiar to many people, such as personal computers, books, cell phones, cosmetics, etc. However, reputations regarding rare products are not collected so much. Reputations of rare products are often contained in personal pages or message board pages managed by volunteers. Therefore our proposed system aims to classify various kinds of web pages into personal pages, message board pages, etc.

Our system adopts a traditional classification model which define mutually exclusive classes a priori, find boundaries of classes in a feature space by learning training data, and choose one best matching class for each testing data. Our method does not require linear separation in a feature space which many traditional classification models such as nearest neighbor assume as a pre-condition. Though neural networks (NNs) are often used as classification methods for non-linear separation\[^5\]\[^6\], it is difficult to provide a complete explanation for results obtained by NNs. In contrast, our method is simple and can give the clear explanation for classification results.

3 Proposed System

In this section, we propose a system for retrieving product reputations from the WWW.

3.1 Page Categories

Web pages which contain information about products can be classified into some types. We have defined five page categories for various kinds of web pages. The five page categories are “personal pages”, “message board pages”, “advertisements or news pages”, “shopping pages” and “other pages”. We assume that product reputations are contained in “personal pages” and “message board pages”. Our proposed system automatically classifies web pages into these five page categories and informs a user of “personal pages” and “message board pages”.

3.2 Components of the Proposed System

A high-level system diagram is shown in Figure 1. First, a product name is given as a query to the system by a user, the search module forwards it to a search engine, and the search module receives web pages from the search engine. Second, six features are extracted for every web page by the feature extraction module. Finally, the classification module classifies these web pages into the five page categories using the six features

\(^1\) http://www.amazon.com.
and category data from category database, and the user obtains only “personal pages” and “message board pages” as a retrieval result. Next two subsections explain the six features and category data.

3.3 Features

In this subsection, six kinds of features, which are required for classifying web pages, are introduced. These features are extracted from a web page. The six kinds of features $f_1, f_2, ..., f_6$ are defined as follows.

$$f_1 = \frac{n_k}{N}$$

where $n_k$ is the number of occurrences in a page of a product name, which is given as a query to the system. $N$ is the total number of words in a page. A product name is repeatedly used as a keyword in “personal pages”, “message board pages” and “advertisements or news pages”, therefore $f_1$ becomes high in these pages. In contrast, $f_1$ becomes low in “shopping pages”.

$$f_2 = \begin{cases} 1 & \text{(if } n_a > 50) \\ 0.02n_a & \text{(otherwise)} \end{cases}$$

$n_a$ is the number of occurrences of anchor tags in a page. Since anchor tags are used more frequently in “advertisements or news pages” than in other page categories, $f_2$ is one of criteria to distinguish “advertisements or news pages” from others.

$$f_3 = \begin{cases} 1 & \text{(if URL of a web page belongs to “.com” or “.co.jp” domain)} \\ 0 & \text{(otherwise)} \end{cases}$$
URLs of “advertisements or news pages” and “shopping pages” often belong to “.com” or “.co.jp”. In contrast, there are few “personal pages” which belong to “.com” or “.co.jp”. $f_3$ gives information which is needed to distinguish “advertisements or news pages” and “shopping pages” from “personal pages”.

$$f_4 = \begin{cases} 1 & \text{if } n_d > 5 \\ 0.2n_d & \text{otherwise} \end{cases}$$ (4)

$n_d$ is the number of occurrences of strings which match to the regular expression of “[0123456789]+/[.:.] [0123456789]+” (brackets metacharacter matches one occurrence of any character inside the brackets and the “+” modifier matches one or more times). Strings which express dates or times such as “9/22” and “10:20” match to the regular expressions. Since these expressions are often used in “message board pages”, $f_4$ is an important feature for identifying “message board pages”.

$$f_5 = \begin{cases} 1 & \text{if } n_s > 4 \\ 0.25n_s & \text{otherwise} \end{cases}$$ (5)

$n_s$ is the number of occurrences of the expressions of “cart”, “order”, “tax”, “shipping fee” or “privacy policy”. Since these words are often used in “shopping pages”, $f_5$ is an important feature for identifying “shopping pages”.

$$f_6 = \begin{cases} 1 & \text{if the word of “copyright” is included in a page} \\ 0 & \text{otherwise} \end{cases}$$ (6)

The word of “copyright” is included in most of “advertisements or news pages” and “shopping pages” because these pages are created by companies. The companies insist their copyright for the contents. Therefore, $f_6$ is an important feature to distinguish “advertisements or news pages” and “shopping pages” from other page categories.

Since each of the above six features is reflected only one aspect of a web page, each individual feature cannot classify web pages perfectly. However, accuracy of classification can be improved by combining the six features. Subsection 3.5 describes the details of the classification method used in the proposed system.

### 3.4 Category Data

Category data is also required in classification module. The category data is created for each of four kinds of page categories except “other pages” and defined as a set of twelve values $\mu_1, \mu_2, \ldots, \mu_6$ and $\sigma_1, \sigma_2, \ldots, \sigma_6$. Where, $\mu_i, \sigma_i (1 \leq i \leq 6)$ are the mean and standard deviation of $f_i$ computed for every sample web page. The samples for each page category are collected from the WWW beforehand by hand, and the category data for each page category is computed and stored into category database.

---

2 “co.jp” means “company in Japan”.
3.5 Classification Method

In the classification module, web pages are classified using the features and the category data. In this module, $n_k$ (the number of occurrences in a page of a product name given as a query) is checked at first. When $n_k \leq 2$, the module determines that the web page belongs to “other pages” category. When $n_k > 2$, the module classifies the web page through the following steps.

To each of four kinds of page categories, i.e., “personal pages”, “message board pages”, “advertisements or news pages” and “shopping pages”, the closeness $S$ of every web page is computed using six features $f_1, f_2, \ldots, f_6$ and category data $\mu_1, \mu_2, \ldots, \mu_6, \sigma, \sigma_2, \ldots, \sigma_6$. The classification module determine a web pages belongs to the page category to which the closeness $S$ of the web page becomes highest. The closeness $S$ is defined by the following equations.

$$ S = \sum_{i=1}^{6} t_i $$

$$ t_i = \begin{cases} 
1 & \text{(if } a_i \leq 1) \\
\frac{a_i - 2}{a_i} & \text{(if } a_i > 1) 
\end{cases} $$

$$ a_i = \frac{|f_i - \mu_i|}{\sigma_i} $$

Figure 2 visually illustrates classification process. The left radar chart shows the six features extracted from a web page. Each of gray regions on the right radar charts
shows the range between $\mu_i - \sigma_i$ and $\mu_i + \sigma_i$ ($1 \leq i \leq 6$) in a page category. When a web page belongs to a certain page category, the features of the web page are involved in the gray region with high probability. When all of the six features are involved in the gray region, the closeness to the page category becomes 6, which is the best value, else becomes worse than 6. For the web page in figure 2, the closeness to the category of “personal pages” is 5.8, which is the highest in all four categories. Therefore, the system determines that the web page belongs to “personal pages” category.

4 Experiments

This section describes the experiments for evaluating the proposed system. In this experiment, the system has attempted to retrieve “personal pages” and “message board pages” for four kinds of products (the names and domains of these four products are shown in table 1). To retrieve the two kinds of pages, the system has classified 50 pages for each product into the five page categories. These 50 pages for each product have been obtained from google by giving each product name as a query to google. We have evaluated the performance of the system by computing precision and recall measures [7]. Precision $P$ and recall $R$ are defined as $P = \frac{|A \cap B|}{|A|}$ and $R = \frac{|A \cap B|}{|B|}$. Where $A$ is the set of pages the system has identified as “personal pages” or “message board pages” and $B$ is the set of pages we have manually identified as “personal pages” or “message board pages”.

Table 1. Experimental Results

| Product Names | Product Domains | Precisions($|A \cap B|/|A|$) | Recalls($|A \cap B|/|B|$) |
|---------------|-----------------|------------------|-----------------|
| Prius         | Car             | 75.0%(12/16)     | 100%(12/12)     |
| ThinkPad330Cs | Note PC         | 63.2%(12/19)     | 80.0%(12/15)    |
| Yukichigo     | Custard Cake    | 85.7%(12/14)     | 41.4%(12/29)    |
| The Lord of the Rings | Movie | 62.5%(5/8)   | 27.8%(5/18)    |

The experimental results are shown in Table 1. The left two columns show product names and domains respectively and the right two columns show precisions and recalls respectively. The precisions are no lower than the precisions Morinaga et.al. have reported [3] though the classification method used in the proposed system does not depend on product domains. Therefore, we think the proposed system has a good performance regarding precisions. The recalls are spread in the range between 27.8% and 100%, which shows the current limitation of the proposed system. Main reason of the unstable recalls is that most of the pages called blogs[8], which belong to “personal pages”, are misidentified as “advertisements or news pages”. We think it is required for improving recalls that the five

3 http://www.google.com.
page categories are subdivided and new features are introduced to identify the new page categories. Since the proposed system adopts a simple classification method, it is easy to add new page categories and new features, which suggest that there is room for improving precision and recall.

5 Conclusions

In this paper, we have proposed and evaluated a system for retrieving product reputations from the WWW. We have illustrated the proposed system can classify web pages into the five page categories using the six features. We have shown that the system has a good performance regarding precisions, however, the recalls have sometimes becomes low. We have discussed that the main reason of the low recalls, and we have reached the conclusion that recalls can be improved by subdividing the five page categories and adding new features.

Since the classification method in our system is simple and domain-independent, the framework can be applied to various reputation retrievals, e.g., retrieval of sightseeing spot reputations, hot spring reputations, hotel reputations, restaurant reputations and so on. As future works, we would like to construct these retrieval systems which has high precision and recall.

References

1. Bayus, B.L.: Word of Mouth: The Indirect Effect of Marketing Efforts, Journal of Advertising Research, Vol.25(3), (1985) 37-51
2. Shardanand, U. and Mases, P.: Social Information Filtering: Algorithms for Automating "Word of Mouth", Proceedings of the SIGCHI conference on Human Factors in Computing Systems, (1995) 210-217
3. Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T.: Mining Product Reputations on the WEB, Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discover and Data Mining, (2002) 341-349.
4. Linden, G., Smith, B. and York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, (2003) 76-80
5. Frosini, A., Gori, M., and Priami, P.: A Neural Network-Based Model for Paper Currency Recognition and Verification, IEEE Transactions on Neural Networks, vol.7, no.6(1996) 1482-1490
6. Gori, M., Lastrucci, L., and Soda, G.: Autoassociator-Based Models for Speaker Verification, Pattern Recognition Letters, Vol.17, (1995) 241-250
7. Ranghavan, V.V., Bollmann, P. and Jung, G.S.: Retrieval System Evaluation using Recall and Precision: Problems and answers, Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1989) 59-68
8. Searls, D. and Sifry, D.: Building with Blogs, Linux Journal, Vol.2003(107), (2003) 4
A Logic-Based Approach for Matching User Profiles

Andrea Cali¹, Diego Calvanese², Simona Colucci³, Tommaso Di Noia³, and Francesco M. Donini⁴

¹ Dip. di Informatica e Sistemistica, Università di Roma “La Sapienza”, Via Salaria 113, I-00198 Roma, Italy
ac@andreacali.com
² Faculty of Computer Science, Free University of Bolzano/Bozen, Piazza Domenicani, 3 I-39100 Bolzano, Italy
calvanese@inf.unibz.it
³ Dip. di Elettrotecnica ed Elettronica, Politecnico di Bari, Via Re David 200 I-70125 Bari, Italy
{s.colucci, t.dinoia}@poliba.it
⁴ Università della Tuscia, Facoltà di Scienze Politiche, Via San Carlo 32, I-01100 Viterbo, Italy
donini@unitus.it

Abstract. Several applications require the matching of user profiles, e.g., job recruitment or dating systems. In this paper we present a logical framework for specifying user profiles that allows profile description to be incomplete in the parts that are unavailable or are considered irrelevant by the user. We present an algorithm for matching demands and supplies of profiles, taking into account incompleteness of profiles and incompatibility between demand and supply. We specialize our framework to dating services; however, the same techniques can be directly applied to several other contexts.

1 Introduction

The problem of matching demands and supplies of personal profiles arises in the business of recruitment agencies, in firms’ internal job assignments, and in the recently emerging dating services. In all scenarios, a list of descriptions of demands is to be matched with a list of descriptions of supplies. In electronic commerce, the general problem is known as matchmaking, although here we do not consider any exchange of goods or services.

In matchmaking, finding an exact match of profiles is not the objective; instead, best possible matches to be provided; in fact, such a match is very unlikely to be found, and in all cases where an exact match does not exist, a solution to matchmaking must provide one or more best possible matches to be explored. Non-exact matches should consider both missing information — details that could be positively assessed in a second phase — and conflicting information — details that should be negotiated if the proposed match is worth enough pursuing. Moreover, when several matches are possible, a matchmaker should list them in a most-promising order, so as to maximize the probability of a
successful match within the first trials. However, such an order should be based on transparent criteria — possibly, logic — in order for the user to trust the system.

Profiles matchmaking can be addressed by a variety of techniques, ranging from simple bipartite graph matching (with or without cost minimization) [9], to vector-based techniques taken from classical Information Retrieval [11][13][12], to record matching in databases, among others. We now discuss some drawbacks of these techniques when transferred to solve matchmaking.

Algorithms for bipartite graph matching find optimal solutions when tying to maximize the number of matches [8][10]. However, such algorithms rely on some way of assigning costs to every match between profiles. When costs are assigned manually, knowledge about them is completely implicit (and subjective), and difficult to revise. Moreover, in maximizing the number of matches a system may provide a bad service to single end users: for example, person $P_1$ could have a best match with job profile $J_1$, but she might be suggested to take job $J_2$ just because $J_1$ is the only available job for person $P_2$. Hence, from end user’s viewpoint, maximizing the number of matches is not the feature that a matchmaker should have.

Both Database techniques for record matching (even with null values), and information retrieval techniques using similarity between weighted vectors of stemmed terms, are not suited for dealing with incomplete information usually present in matchmaking scenarios. In fact, information about profiles is almost always incomplete, not only because some information is unavailable, but also because some details are simply considered irrelevant by either the supplier or the demander — and should be left as such. Imposing a system interface for entering profiles with long and tedious forms to be filled in, is the most often adopted “solution” to this incompleteness problem — but we consider this more an escape for constraining real data into an available technique, than a real solution. For example, in a job posting/finding system, the nationality could be considered irrelevant for some profiles (and relevant for others); or in a dating service, some people may find disturbing (or simply inappropriate) the request to specify the kind of preferred music, etc. In such situations, missing information can be assumed as an “any-would-fit” assertion, and the system should cope with this incompleteness as is.

To sum up, we believe that there is a representation problem that undermines present solutions to matchmaking: considering how profiles information is represented is a fundamental step to reach an effective solution, and representations that are either too implicit, or overspecified, lead to unsatisfactory solutions.

Therefore, our research starts with proposing a language, borrowed from Artificial Intelligence, that allows for incomplete descriptions of profiles, and both positive and negative information about profiles. In particular, we propose a Description Logic [1] specifically tailored for describing profiles. Then, we model the matching process as a special reasoning service about profiles, along the lines of [5][6]. Specifically, we consider separately conflicting details and missing details, and evaluate how likely is the match to succeed, given both missing and conflicting details. Our approach makes transparent the way matches are evaluated — allowing end users to request justifications for suggested matches. We devise some special-purpose algorithms to solve the problem
for the language we propose, and evaluate the possible application scenarios of a dating service.

2 A Description Logic for Representing Profiles

We use a restriction of the \(\mathcal{ALC}(D)\) Description Logic, that, besides concepts and roles to represent properties of (abstract) objects, also allows one to express quantitative properties of objects, such as weight, length, etc., by means of concrete domains \([2]\). Each concrete domain \(D\), e.g., the real numbers \(\mathbb{R}\), has a set of associated predicate names, where each predicate name \(p\) denotes a predicate \(p^D\) over \(D\). For our purpose, it is sufficient to restrict the attention to unary predicates, and we assume that among such unary predicates we always have a predicate \(\top\) denoting the entire domain, and predicates \(\geq \ell(\cdot)\) and \(\leq \ell(\cdot)\), for arbitrary values \(\ell\) of \(D\). We also assume that the concrete domains we deal with are admissible, which is a quite natural assumption, satisfied e.g., by \(\mathbb{R}\) (see [2] for the details). Besides roles, the logic makes use of features. Each feature has an associated concrete domain \(D\) and represents a (functional) relation between objects and values of \(D\).

Starting from a set of concept names (denoted by the letter \(A\)), a set of role names (denoted by \(R\)), a set of unary predicate names (denoted by \(p\)), and a set of features (denoted by \(f\)), we inductively define the set of concepts (denoted by \(C\)) as follows.

- Every concept name \(A\) is a concept (atomic concept), and for \(C_1\) and \(C_2\) concepts, \(R\) a role name, \(f\) a feature with associated domain \(D\), and \(p\) a unary predicate of \(D\), the following are concepts:
  - \(C_1 \cap C_2\) (conjunction), \(C_1 \cup C_2\) (disjunction), and \(\neg C\) (negation);
  - \(\exists R.C\) (existential restriction) and \(\forall R.C\) (universal restriction); \(p(f)\) (predicate restriction).

To express intentional knowledge about concepts, we make use of a concept hierarchy, which is a set of assertions of the form \(A_1 \sqsubseteq A_2\) and \(A_1 \sqsubseteq \neg A_2\), with \(A_1\) and \(A_2\) concept names. The former assertion expresses an inclusion, while the latter expresses a disjointness. For example, football \(\sqsubseteq\) sport and male \(\sqsubseteq \neg\) female could be assertions that are part of a concept hierarchy.

Formally, the semantics of concepts is defined by an interpretation \(\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})\), consisting of an abstract domain \(\Delta^\mathcal{I}\) and an interpretation function \(\cdot^\mathcal{I}\) that assigns to each concept name \(A\) a subset \(A^\mathcal{I}\) of \(\Delta^\mathcal{I}\); to each role name \(R\) a binary relation \(R^\mathcal{I}\) over \(\Delta^\mathcal{I}\), and to each feature name \(f\), associated with the concrete domain \(D\), a partial function \(f^\mathcal{I}: \Delta^\mathcal{I} \rightarrow D\). The interpretation function can be extended to arbitrary concepts as follows:

\[
\begin{align*}
(C_1 \cap C_2)^\mathcal{I} & = C_1^\mathcal{I} \cap C_2^\mathcal{I} \\
(C_1 \cup C_2)^\mathcal{I} & = C_1^\mathcal{I} \cup C_2^\mathcal{I} \\
(\neg C)^\mathcal{I} & = \neg C^\mathcal{I} \\
(\exists R.C)^\mathcal{I} & = \{c \in \Delta^\mathcal{I} \mid \text{there exists } d \in \Delta^\mathcal{I} \text{ s.t. } (c, d) \in R^\mathcal{I} \text{ and } d \in C^\mathcal{I}\} \\
(\forall R.C)^\mathcal{I} & = \{c \in \Delta^\mathcal{I} \mid \text{for all } d \in \Delta^\mathcal{I} \text{ s.t. } (c, d) \in R^\mathcal{I} \text{ we have } d \in C^\mathcal{I}\} \\
(p(f))^\mathcal{I} & = \{c \in \Delta^\mathcal{I} \mid f^\mathcal{I}(c) \in p^D\}
\end{align*}
\]

An assertion \(A_1 \sqsubseteq A_2\) is satisfied by an interpretation \(\mathcal{I}\) if \(A_1^\mathcal{I} \subseteq A_2^\mathcal{I}\). An assertion \(A_1 \sqsubseteq \neg A_2\) is satisfied by an interpretation \(\mathcal{I}\) if \(A_1^\mathcal{I} \cap A_2^\mathcal{I} = \phi\). We call an interpretation...
that satisfies all assertions in a hierarchy $\mathcal{H}$ a model of $\mathcal{H}$. A concept $C$ is satisfiable in $\mathcal{H}$ if $\mathcal{H}$ admits a model $\mathcal{I}$ such that $C^\mathcal{I} \neq \phi$. A hierarchy $\mathcal{H}$ logically implies an assertion $C_1 \subseteq C_2$ between arbitrary concepts $C_1$ and $C_2$ if $C_1^\mathcal{I} \subseteq C_2^\mathcal{I}$, for each model $\mathcal{I}$ of $\mathcal{H}$.

### 3 Representing User Profiles

We describe how to represent user profiles using the Description Logic presented in Section 2. The user profiles are tailored for dating services, though the same framework can be used, with small modifications, for different applications. We do not use the full expressive power of the Description Logic. In particular, we use a single role hasInterest, to express interest in topics, and we make a limited use of the constructs. We assume the set of features to represent physical characteristics such as age, height, etc. Additionally, we use a special feature level that expresses the level of interest in a certain field. The concrete domain associated to level is the interval $\{\ell \in \mathbb{R} | 0 < \ell \leq 1\}$.

A user profile $P$ consists of the conjunction of the following parts:

- A conjunction of atomic concepts, to represent atomic properties associated to the user. We denote the set of such concepts as $\text{Name}(P)$.

- A conjunction of concepts of the form $p(f)$, to represent physical characteristics. The (unary) predicate $p$ can be one of the predicates $\geq_{\ell}(\cdot)$, $\leq_{\ell}(\cdot)$, $=\ell(\cdot)$, where $\ell$ is a value of the concrete domain associated to $f$, or any logical conjunction of them. We denote the set of such concepts as $\text{Features}(P)$. Since $(p_1 \land p_2)(f)$ is equivalent to $p_1(f) \cap p_2(f)$, in the following, we can assume w.l.o.g. that $\text{Features}(P)$ contains at most one concept of the form $p(f)$ for each feature $f$.

- A conjunction of concepts of the form $\exists \text{hasInterest}.(C \cap \geq_{x}(\text{level}))$, where $C$ is a conjunction of concept names, and $0 \leq x \leq 1$. Each such concept represents an interest in a concept $C$ with level at least $x$. We denote the set of such concepts as $\text{Interests}(P)$.

- A conjunction of concepts of the form $\forall \text{hasInterest}.(\neg C \sqcup \leq_{x}(\text{level}))$, where $C$ is a conjunction of concept names, and $0 \leq x \leq 1$. Each such concept represents the fact that the interest in a concept $C$ has level at most $x$. Note that, to represent the complete lack of interest in $C$, it is sufficient to put $x = 0$. We denote the set of such concepts as $\text{NoInterests}(P)$.

**Example 1.** A supplied profile describing, say, a 35-years-old male, 1.82 cm tall, with strong interests in fantasy novels and japanese comics, fair interest in politics and no interest in football, could be expressed as follows:

$$\text{male} \land =_{35}(\text{age}) \land =_{1.82}(\text{height}) \land$$
$$\exists \text{hasInterest}.(\text{fantasyNovels} \cap \geq_{0.8}(\text{level})) \land$$
$$\exists \text{hasInterest}.(\text{japaneseComics} \cap \geq_{0.8}(\text{level})) \land$$
$$\exists \text{hasInterest}.(\text{politics} \cap \geq_{0.4}(\text{level})) \land$$
$$\forall \text{hasInterest}.(\neg \text{football} \sqcup \leq_{0}(\text{level}))$$

---

1 For modeling profiles in different contexts, additional roles could be added to this language. For example, hasSkill for expressing skills in certain fields.
where we suppose that interests are organized in a hierarchy including \( \text{fantasyNovels} \subseteq \text{novels} \), \( \text{japaneseComics} \subseteq \text{comics} \), and \( \text{male} \subseteq \neg \text{female} \).

Observe that, when a profile is demanded, usually features like \( \text{age} \) and \( \text{height} \) will be used with range predicates (e.g., \( (\geq 30 \land \leq 70)(\text{age}) \)), instead of equality predicates as in the above example.

The following property follows immediately from the semantics of existential restriction. For every pair of concepts \( C_1 \) and \( C_2 \), role \( R \), feature \( f \) with associated concrete domain \( D \), and \( p \) a predicate of \( D \):

\[
\text{if } H \models C_1 \sqsubseteq C_2 \text{ then } H \models \exists R.(C_1 \cap \geq \ell(f)) \sqsubseteq \exists R.(C_2 \cap \geq \ell(f))
\]

For example, if \( \text{football} \subseteq \text{sport} \), then someone with a level of interest \( \ell \) in football has at least the same level of interest in sport. This property is exploited in the matching algorithm provided in Section 4.

### 4 The Matching Algorithm

We present the algorithm for matching user profiles. The matching is performed over two profiles: the demand profile \( P_d \) and the supply profile \( P_s \). The algorithm is not symmetric, i.e., it evaluates how \( P_s \) is suited for \( P_d \), which is different from how \( P_d \) is suited for \( P_s \) \([7]\); of course, in order to determine how \( P_d \) is suited for \( P_s \), we can simply exchange the arguments of the algorithm.

From a logical point of view, we extend the non-standard inferences \textit{contraction} and \textit{abduction} defined in \([4]\). In particular, our contraction either removes or weakens conjuncts from \( P_d \) so as to make \( P_d \cap P_s \) satisfiable in \( H \); abduction, instead, either adds or strengthens conjuncts in \( P_s \) so as to make \( H \models P_s \subseteq P_d \). The algorithm is based on structural algorithms for satisfiability and subsumption \([3]\). Since it is reasonable to assume that users do not enter contradicting information, we assume that the profiles \( P_d \) and \( P_s \) are consistent.

The result of the match is a \textit{penalty} in \( \mathbb{R} \): the larger the penalty, the less \( P_s \) is suited for \( P_d \). In particular, partial penalties are added to the overall penalty by matching corresponding conjuncts of the two profiles; this is done in two ways.

**Contraction.** When a conjunct \( C_d \) in \( P_d \) is in contrast with some conjunct \( C_s \) in \( P_s \), then \( C_d \) is removed and a penalty is added. Intuitively, since the supplier has something the demander does not like, in order to make the profiles match the demander gives up one of her requests. For example, let \( C_d = \forall \text{hasInterest}.(\neg \text{sport} \sqcup \leq 0.2(\text{level})) \) and \( C_s = \exists \text{hasInterest}.(\text{football} \sqcap \geq 0.4(\text{level})) \), where we have \( \text{football} \subseteq \text{sport} \) in \( H \). In this case the demander looks for someone who does not like sports very much, while the supplier likes football and therefore he likes sports. In this case, pursuing the match would require the demander to give up his/her request about sports, so the algorithm adds a penalty \( \Pi_{\ell_{\ell_d}}(0.4, 0.2) \) that depends on the gap between the lower bound (0.4) of the supply and the upper bound (0.2) of the demand. Similarly, for a feature \( f \) with
contrasting predicates $p_d$ and $p_s$, a penalty $\Pi_{cf}(p_d(f), p_s(f))$ is added to take into account the removal of $p_d(f)$ from $P_d$. In case a concept $A_d$ representing an atomic property has to be removed, the algorithm makes use of another penalty function $\Pi_c(\cdot)$, whose argument is the concept $A_d$.

**Abduction.** When a conjunct $c_d$ in $P_d$ has no corresponding conjunct in $P_s$, we add a suitable conjunct $c_s$ in $P_s$ that makes the profiles match, and add a corresponding penalty. Intuitively, the demander wants something which the supplier does not provide explicitly; in this case we assume that the supplier may or may not satisfy the demander’s request, and as a consequence of this possibility of conflict we add a penalty. This is done by means of a penalty function $\Pi_a(\cdot)$, whose argument is a concept $C$, that takes into account the addition of $C$ to $P_s$. When the level of interest must be strengthened, we use a function $\Pi_{a\ell}(\cdot)$, that takes into account the gap between bounds. Similarly, a penalty function $\Pi_{af}(\cdot)$ takes into account the addition of features.

**Algorithm** CalculatePenalty

**Input** demand profile $P_d$, supply profile $P_s$, concept hierarchy $\mathcal{H}$

**Output** real value penalty $\geq 0$

penalty := 0;

// Contraction
foreach $A_d \in \text{Names}(P_d)$ do
    if there exists $A_s \in \text{Names}(P_s)$ such that $\mathcal{H} \models A_d \subseteq \neg A_s$ then remove $A_d$ from $P_d$
    penalty := penalty + $\Pi_c(A_d)$

foreach $p_d(f) \in \text{Features}(P_d)$ do
    if there exists $p_s(f) \in \text{Features}(P_s)$ such that $\exists x. p_d(x) \land p_s(x)$ is unsatisfiable in the domain associated to $f$ then remove $p_d(f)$ from $P_d$
    penalty := penalty + $\Pi_{cf}(p_d(f), p_s(f))$

foreach $\exists \text{hasInterest}.(C_d \sqsubseteq_{x_d} \text{level}) \in \text{Interests}(P_d)$ do
    foreach $\forall \text{hasInterest}.(\neg C_s \sqsubseteq_{x_s} \text{level}) \in \text{NoInterests}(P_s)$ do
        if $\mathcal{H} \models C_d \sqsubseteq C_s$ and $x_d \geq x_s$
            then replace $\exists \text{hasInterest}.(C_d \sqsubseteq_{x_d} \text{level})$ in $P_d$
                with $\exists \text{hasInterest}.(C_d \sqsubseteq_{x_s} \text{level})$
            penalty := penalty + $\Pi_{c\ell}(x_d, x_s)$

foreach $\forall \text{hasInterest}.(\neg C_d \sqsubseteq_{x_d} \text{level}) \in \text{NoInterests}(P_d)$ do
    foreach $\exists \text{hasInterest}.(C_s \sqsubseteq_{x_s} \text{level}) \in \text{Interests}(P_s)$ do
        if $\mathcal{H} \models C_s \sqsubseteq C_d$ and $x_d \leq x_s$
            then replace $\forall \text{hasInterest}.(\neg C_d \sqsubseteq_{x_d} \text{level})$ in $P_d$
                with $\forall \text{hasInterest}.(\neg C_d \sqsubseteq_{x_s} \text{level})$
            penalty := penalty + $\Pi_{c\ell}(x_s, x_d)$
Abduction

foreach $A_d \in \text{Names}(P_d)$ do
  if there does not exist $A_s \in \text{Names}(P_s)$ such that $\mathcal{H} \models A_s \sqsubseteq A_d$
    then add $A_d$ to $P_s$
      penalty := penalty + $\Pi_a(A_d)$

foreach $p_d(f) \in \text{Features}(P_d)$ do
  if there exist $p_s(f) \in \text{Features}(P_s)$
    then if $\forall x. p_s(x) \Rightarrow p_d(x)$ is false in the domain associated to $f$
      then add $p_d(f)$ to $P_s$
        penalty := penalty + $\Pi_{af}(p_d(f), p_s(f))$
    else add $p_d(f)$ to $P_s$
      penalty := penalty + $\Pi_{af}(p_d(f), \top(f))$

foreach $\exists \text{hasInterest.}(C_d \sqsupseteq x_d)(\text{level}) \in \text{Interests}(P_d)$ do
  if there does not exist $\exists \text{hasInterest.}(C_s \sqsupseteq x_s)(\text{level}) \in \text{Interests}(P_s)$
    such that $\mathcal{H} \models C_s \sqsubseteq C_d$ and $x_s \geq x_d$
    then if there exists $\exists \text{hasInterest.}(C_s \sqsupseteq x_s)(\text{level}) \in \text{Interests}(P_s)$
      such that $\mathcal{H} \models C_s \sqsubseteq C_d$
        then let $\exists \text{hasInterest.}(C_s \sqsupseteq x_s)(\text{level})$ be the concept in $\text{Interests}(P_s)$
          with maximum $x_s$ among those for which $\mathcal{H} \models C_s \sqsubseteq C_d$
          holds
          penalty := penalty + $\Pi_{af}(x_d, x_s)$
        else penalty := penalty + $\Pi_a(\exists \text{hasInterest.}(C_d \sqsupseteq x_d)(\text{level}))$
          add $\exists \text{hasInterest.}(C_d \sqsupseteq x_d)(\text{level})$ to $P_s$
  foreach $\forall \text{hasInterest.}(-C_d \sqsubseteq x_d)(\text{level}) \in \text{NoInterests}(P_d)$
    if there does not exist $\forall \text{hasInterest.}(-C_s \sqsubseteq x_s)(\text{level}) \in \text{NoInterests}(P_s)$
      such that $\mathcal{H} \models C_d \sqsubseteq C_s$ and $x_d \geq x_s$
      then if there exists $\forall \text{hasInterest.}(-C_s \sqsubseteq x_s)(\text{level}) \in \text{NoInterests}(P_s)$
        such that $\mathcal{H} \models C_d \sqsubseteq C_s$
          then let $\forall \text{hasInterest.}(-C_s \sqsubseteq x_s)(\text{level})$ be the concept in $\text{Interests}(P_s)$
            with maximum $x_s$ among those for which $\mathcal{H} \models C_d \sqsubseteq C_s$
            holds
            penalty := penalty + $\Pi_{af}(x_s, x_d)$
          else penalty := penalty + $\Pi_a(\forall \text{hasInterest.}(-C_d \sqsubseteq x_d)(\text{level}))$
            add $\forall \text{hasInterest.}(-C_d \sqsubseteq x_d)(\text{level})$ to $P_s$
  return penalty

It is easy to check that all subsumption tests $\mathcal{H} \models C_1 \sqsubseteq C_2$ in the algorithm can be done in polynomial time in the size of $\mathcal{H}, C_1,$ and $C_2$. Hence, it can be straightforwardly proved that the complexity of the algorithm is polynomial w.r.t. the size of the input.

The following theorem establishes the correctness of the above algorithm w.r.t. the computation of contraction and abduction. We denote with $P_d^a$ the profile $P_d$ after contraction, and with $P_s^a$ the profile $P_s$ after abduction.

**Theorem 1.** Given a concept hierarchy $\mathcal{H}$, a demand profile $P_d$, and a supply profile $P_s$, the following properties hold: (i) $P_d^a \sqcap P_s$ is satisfiable in $\mathcal{H}$; (ii) $P_s^a$ is satisfiable
in $\mathcal{H}$; (iii) $\mathcal{H} \models P^a_d \sqsubseteq P^a_s$; (iv) there does not exist a profile $P'_s$ more general than $P^a_s$ (i.e., $\mathcal{H} \models P^a_s \sqsubseteq P'_s$ and $\mathcal{H} \not\models P'_s \sqsubseteq P_s$) such that $\mathcal{H} \models P'_s \sqsubseteq P_s$ and $\mathcal{H} \models P_s' \sqsubseteq P^c_d$.

5 Conclusions

In this paper we have addressed the problem of matching user profiles, when the demander’s and supplier’s profiles can have missing or conflicting information. In such a case, we have to take into account that the demander may need to give up some of her requests, and/or she may need to make assumptions on unspecified properties of the supplier’s profile. We have proposed a DL-based framework for expressing user profiles in this setting, and a language suited for dating services. We have proposed an ad-hoc structural algorithm for matching profiles that, given a demander’s and a supplier’s profile, returns a penalty: the higher the penalty, the less the two profiles are compatible. As a future work, we want to test the algorithm in real cases with a prototype that is currently under development: we believe that promising applications of our techniques can be dating, recruitment, and service discovery systems.

Acknowledgments. The first two authors were partly supported by MIUR under FIRB (Fondo per gli Investimenti della Ricerca di Base) project “MAIS: Multichannel Adaptive Information Systems” in the context of the Workpackage 2 activities.

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2003.
2. F. Baader and P. Hanschke. A schema for integrating concrete domains into concept languages. In Proc. of IJCAI’91, pages 452–457, 1991.
3. A. Borgida and P. F. Patel-Schneider. A semantics and complete algorithm for subsumption in the CLASSIC description logic. J. of Artificial Intelligence Research, 1:277–308, 1994.
4. S. Colucci, T. Di Noia, E. Di Sciascio, F. M. Donini, and M. Mongiello. Concept abduction and contraction in description logics. In Proc. of DL 2003. CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-81/ 2003.
5. S. Colucci, T. Di Noia, E. Di Sciascio, F. M. Donini, M. Mongiello, and M. Mottola. A formal approach to ontology-based semantic match of skills descriptions. J. of Universal Computer Science, Special issue on Skills Management, 2003.
6. T. Di Noia, E. Di Sciascio, F. M. Donini, and M. Mongiello. Abductive matchmaking using description logics. In Proc. of IJCAI 2003, pages 337–342, 2003.
7. T. Di Noia, E. Di Sciascio, F. M. Donini, and M. Mongiello. A system for principled matchmaking in an electronic marketplace. In Proc. of WWW 2003, pages 321–330, May 20–24 2003.
8. Z. Galil. Efficient algorithms for finding maximum matching in graphs. ACM Computing Surveys, 18(1):23–38, 1986.
9. F. S. Hillier and G. J. Lieberman. Introduction to Operations Research. McGraw-Hill, 1995.
10. J. Kennington and Z. Wang. An empirical analysis of the dense assignment problem: Sequential and parallel implementations. ORSA Journal on Computing, 3(4):299–306, 1991.
11. D. Kuokka and L. Harada. Integrating information via matchmaking. *J. of Intelligent Information Systems*, 6:261–279, 1996.

12. K. Sycara, S. Widoff, M. Klusch, and J. Lu. LARKS: Dynamic matchmaking among heterogeneous software agents in cyberspace. *Autonomous agents and multi-agent systems*, 5:173–203, 2002.

13. D. Veit, J. P. Müller, M. Schneider, and B. Fiehn. Matchmaking for autonomous agents in electronic marketplaces. In *Proc. of AGENTS ’01*, pages 65–66. ACM, 2001.
Pose Classification of Car Occupant Using Stereovision and Support Vector Machines

Min-Soo Jang\textsuperscript{1}, Yong-Guk Kim\textsuperscript{2}, Hyun-Gu Lee\textsuperscript{1}, Byung-Joo Lee\textsuperscript{1}, Soek-Joo Lee\textsuperscript{3}, and Gwi-Tae Park\textsuperscript{1}

\textsuperscript{1}Dept. of Electrical Engineering, Korea University, Seoul, Korea
gtpark@korea.ac.kr
\textsuperscript{2}School of Computer Engineering, Sejong University, Seoul, Korea
ykim@sejong.ac.kr
\textsuperscript{3}Hyundai Autonet Co., Korea

Abstract. Airbag in the cars plays an important role for the safety of occupants. However, Highway traffic safety report shows that many occupants are actually killed by wrong deployment of the airbags. For reducing risk caused by airbag, designing a smart airbag is an important issue. The present study describes an occupants’ pose classification system, by which triggering and intensity of the airbag deployment can be controlled. The system consists of a pair of stereo cameras and a SVM (Support Vector Machine) classifier. Performance of the system shows its feasibility as a vision-based airbag controller.

1 Introduction

Airbags are important role for preventing life-threatening and debilitating head and chest injuries by avoiding direct impact contact at the dashboard. Although airbags have saved many lives, the National Highway Traffic Safety Administration (NHTSA) reported that 153 occupants have been killed by the deployment of the airbag itself (2000.4). Particularly, 89 occupants (58.2\%) of the death were children or infants in rear-facing infant seats \cite{1}. For solving this problem, NHTSA has recently issued a set of regulations mandating low risk deployment of the airbag.

In this paper, we present a new system that consists of a stereo vision and occupants’ pose classification system using a SVM (Support Vector Machine) classifier. The advantages of using the vision sensor for this application include the ability to capture diverse information about occupant, such as class, pose, distance from occupant to dashboard, and so on. In addition, it can be installed almost anywhere within the car and it has higher accuracy compared to other sensors based on weight, ultrasound, infrared, and electronic fields \cite{8, 9}. Our vision system is a pair of CCD cameras mounted rear-view mirror position and it is pointed towards the occupant. We first extract a disparity map using a fast stereo algorithm \cite{4, 5}, and then normalized horizontal and vertical projection histograms of the disparity map are used as input for the SVM classifier \cite{6}. SVM is relatively simple, but is powerful classification technique initially proposed by Vapnik and his colleagues at AT&T Bell Laboratory \cite{2, 3}. The well trained SVM is able to separate between two target classes using a
hyperplane. Or it can be extend to be a multiple classifier. Here, we classify the poses of the occupants into five classes (i.e. crawling pose, crawling-right pose, crawling-left pose, standard pose, and laying down pose). When such classification is possible, the airbag system will regard the crawling, crawling-right, and crawling-left posed occupants who are too close to the airbag and will deploy the airbag weakly. On the other hand, in case of standard posed occupants, the system deploys the airbag normally. Finally, it will not deploy the airbag for laying down posed occupants.

In section 2, the details of our stereo vision system are described. The SVM classification method is discussed in section 3. Result of experiments is described in section 4. Finally, we summarize our results, and discuss the performance of the whole system in section 5.

2 Stereo Vision

In this section, we describe our stereovision system and how the system is operated. Figure 1(a) is a schematic illustration of the system. It consists of two CCD cameras with a jig for controlling rotation and tilt of the cameras. Figure 1(b) shows a pair of stereo images captured with the vision system.

![A stereo vision system (a) and a pair of stereo images (b)](image)

Figure 2(a) shows the concept of disparity. Disparity of arbitrary point, \( P \) is defined as follow:

\[
d = u - u'
\]  

(1)

where \( f \) is focal length of the camera, \( B \) is the distance between two cameras, and \( (u,v) \), \( (u',v') \) are coordinates of the point, \( P \) on the left and right images, respectively. To calculating a disparity of an arbitrary point \( P \), we need to a stereo matching method that computes the disparity between two images of the pair. For this purpose, we adopt a fast SAD (Sum of the Absolute Difference) algorithm [4, 5]. Figure 2(b) shows a disparity map generated by the fast SAD algorithm. Since the disparity map contains 3D information of an occupant, we can recognize the pose of occupant by analyzing the disparity map.
3 Support Vector Machine as a Classifier

The fundamental idea of SVM is to construct a hyperplane as the decision line, which separates the positive (+1) classes from the negative (-1) ones with the largest margin [2, 3]. In a binary classification problem, let us consider the training sample \( \{ (x_i, d_i) \}_{i=1}^{N} \), where \( x_i \) is the input pattern for the i-th sample and \( d_i \) is the corresponding desired response (target output) with subset \( d \in \{-1, +1\} \). The equation of a hyperplane that does the separation is

\[
wx + b = 0
\]  

(2)

where \( x \) is an input vector, \( w \) is an adjustable weight vector, and \( b \) is a bias. Figure 3 shows an optimal hyperplane for the linearly separable case and margin, \( \gamma \). The aim of the SVM classifier is to maximize the margin.

The margin, distance of the nearest point to the typical hyperplane, equals to \( 1/\|w\| \). So, the problem turns into a quadratic programming:

![Figure 3](image-url)
minimize \frac{1}{2} \|w\|^2

s.t. \quad y_i (w \cdot x_i + b) \geq 1, \quad i = 1, 2, 3, \ldots, N

(3)

Introducing Lagrange multipliers \alpha_i \geq 0, i = 1, \ldots, n, one for each of the constraints in (3), we get the following Lagrangian [10]:

L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \left( y_i (w \cdot x_i + b) - 1 \right)

(4)

We get the dual quadratic optimization problem:

\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j)

subject to \alpha_i \geq 0, i = 1, \ldots, n,

\sum_{i=1}^{n} \alpha_i y_i = 0

(5)

Thus, by solving the dual optimization problem, one obtains the coefficients \alpha_i. This leads to the following decision function.

f(x) = \text{sgn} \left( \sum_{i=1}^{n} y_i \alpha_i (x \cdot x_i) + b \right)

(6)

The other important property of SVM is that it can use diverse kernels in dealing with the input vectors. The standard kernels are polynomial, Gaussian and sigmoid kernels.

4 Image Database and System Performance

4.1 Stereo Image Database

The performance of the pose classification algorithm is evaluated with our stereo image database, which consists of many stereo images captured within a car with 10 people with 64 different poses. The total of the captured images are 1280 and the resolution of an image is 320 × 240.

Table 1 catalogs the diverse composition of the poses with variation of legs, hands, body, background and seat. Figure 4 depicts several examples of the database such as crawling pose (a), crawling-right pose (b), crawling left pose (c), standard pose (d), and laying down pose (e), respectively.

Figure 5 shows the normalized horizontal and vertical projection histogram of a disparity map and the histograms will be used as an input of SVM [6]. Note that the whole disparities on same horizontal or vertical line are added. We calculate only disparities on odd lines for reducing dimension of the data. Therefore, a data set for one disparity map image consists of 250 values (horizontal: 105, vertical 145), since
Table 1. Composition of the different poses

| Variation        | Legs | Hands   | Body                  | Background | Seat          |
|------------------|------|---------|-----------------------|------------|---------------|
|                  | Open | Straight down | Crawling pose | with pattern | Standard position |
|                  | Close| On knee  | Standard pose        | without pattern | Laying down position |
|                  |      |          | Laying down pose      |            |                |

Fig. 4. Stereo image pairs for each pose (a) Crawling pose (b) Crawling right pose (c) Crawling left pose (d) Standard pose (e) Laying down pose

disparities on odd lines for reducing dimension of the data. Therefore, a data set for one disparity map image consists of 250 values (horizontal: 105, vertical 145), since the outline of a disparity map image (15 pixels), shown as black band in Fig. 5, is not used by the stereo matching algorithm.

Figure 6 shows average vertical and horizontal projection histogram of each poses. In the horizontal projection histogram of Fig. 6(a) and vertical projection histogram of Fig. 6(b), solid line, dashed line, dotted line, dash dot line, and cross marker line means the crawling pose, standard pose, laying down pose, crawling right pose, and crawling left pose, respectively.
4.2 Performance of the System

The total number of data sets within the DB is 640 with 250 values for each image. We randomly split the data sets into a training set (50%) and a test set (50%). We used the public domain implementation of SVM, called LibSVM and two standard kernels (i.e. polynomial kernel and radial basis function (or Gaussian) kernel) [7]. The experiment was accomplished for 10 times and average results are shown in table 2. In case of polynomial kernel, the correction rate was 92.47%, and in case of RBF kernel, it was 99.78%.

Table 2. Classification rate for each kernels

| Kernel   | Classification rate |
|----------|---------------------|
| Polynomial | 92.47%             |
| RBF      | 99.78%             |
5 Conclusions and Discussion

As an initiative to develop a smart airbag system, we have designed a stereovision system combined with a SVM classifier. For the stereo vision system, the fast SAD algorithm is adopted to calculate the disparity map of the stereo images. As the disparity map image contains 3-D information of the occupant, it is possible to classify the pose of the occupant by analyzing the disparity map image. SVM is used for classifying occupants’ poses, since it is known that the SVM classifier is relatively simple and yet powerful. The stereo image DB was constructed for verifying the performance of the system. Result shows that the performance of the system is satisfactory, suggesting that the vision-based airbag control has a potential. We plan to carry out further experiment for classifying child from adult based upon the present vision system.

Acknowledgments

This work was supported by Hyundai Autonet Co.

References

1. http://www.nhtsa.dot.gov
2. V. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, NY, USA, pp.45-98, 1995
3. S. Haykin, “Neural Network”, Prentice Hall, 1999
4. C. L. Zitnick and T. Kanade, “A Cooperative Algorithm for Stereo Matching and Occlusion Detection”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, Issue 7, pp. 675-684, 2000
5. L. Di Stefano, M. Marchionni, S. Mattoccia, and G. Neri, “A Fast Area-Based Stereo Matching Algorithm”, 15th International Conference on Vision Interface, 2002
6. ZI. Haritaoglu, D. Harwood, and L. S. Davis, “Ghose: A Human Body Part Labeling System Using Silhouettes”, 14th international conference on Pattern Recognition, vol. 1, pp. 77-82, 1998
7. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
8. W. Fultz, D.Griffin, S.Kiselewich, M. Murphy, C. Wu, Y. Owechko, N. Srinivasa, and P. Thayer, “Selection of a Sensor Suite for an Occupant Position and Recognition System”, Proc. Of SAE TOPTEC conference on Vehicle Safety Restraint Systems, 1999
9. Y. Owechko, N. Srinivasa, S. Medasani, and R. Boscolo, “Vision-Based Fusion System for Smart Airbag Applicaions”, IEEE, Intelligent Vehicle Symposium, vol. 1, pp. 245-250, 2002
10. K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, “An Introduction to Kernel-Based Learning Algorithms”, IEEE Trans. on Neural Networks, vol. 12, no. 2, pp. 181-201, 2001
A Fully Automatic System Recognizing Human Facial Expressions

Yong-Guk Kim¹, Sung-Oh Lee², Sang-Jun Kim², and Gwi-Tae Park²*

¹School of Computer Engineering, Sejong University, Seoul, Korea
ykim@sejong.ac.kr
²Department of Electrical Engineering, Korea University, Seoul, Korea
*gtpark@korea.ac.kr

Abstract. The facial expression recognition system normally consists of three cascade stages: face detection, normalization and classification of facial expressions. Recent studies in this area often concentrate on how to develop a fast and accurate classifier, which corresponds to the last stage. However, it is essential to automate the first two stages. This paper describes a fully automatic facial expression system in which above three stages are carried out without any human intervention. In particular, we focus on how the normalization stage impacts upon the overall performance of the facial expression recognition system. Recognition performance of the automatic case is compared with that of the manual normalization case.

1 Introduction

The human face is a great communication device, because facial expressions of humans are so diverse and subtle, and yet are immediately recognized by observers. Long ago Charles Darwin [3] establishes that there are six prototypical (or basic) facial expressions, which are universal across human ethnicities and cultures: surprise, fear, sadness, anger, disgust and happiness. Such foundation provides a convenient framework, by which we test and compare different facial expression recognition systems. Accurate measurement of facial actions has many potential applications such as intelligent human-computer interaction, entertainment devices, psychiatry [5] and neurology.

In the face recognition system, several preprocessing operations such as detecting a facial area from the given image, finding the locations of two eyes, and warping the face image by adjusting the locations of the eyes are necessary. Similarly, the facial expression recognition system also includes a geometrical normalization stage [2, 4]. In this case, however, the location of the mouth as well as those of two eyes are included for normalizing the input image, since it seems that the muscle movements around the mouth provide important information in extracting facial emotions.

To represent facial expression images effectively, several methods have been proposed such as PCA [2, 4], ICA [4], Gabor representation [4, 7], Optic flow, and geometrical tracking method. Among them, the Gabor representation has been favored
among many researches, because of its better performance and biological implication [1, 4]. In this paper, it is shown that an enhanced Fisher linear discrimination model (EFM) classifier combined with the Gabor representation has performed better than the others. Moreover, we present a fully automated facial expression recognition system where face and facial feature detection are performed without any human intervention.

The rest of this paper is organized as follows: Section 2 describes the overall system architecture. The facial expression image database is introduced in section 3. And normalization method of the image is described. Section 4 contains face and facial feature detection method. Section 5 shows the recognition process including feature representations and classifiers. In section 6, we compare the performance between several classifiers and automatic registration and manual registration for the facial expression recognition. Discussion and conclusions will be given in section 7.

![Fig. 1. A facial expression recognition system consisting of three cascade stages](image)

## 2 The Cascade Stages of the Facial Expression Recognition System

Our facial expression recognition system, as illustrated in Fig. 1, consists of three cascade stages: face detection, normalization, and facial expression recognition, including the facial expression database input images. In the face region detection stage, SVM (Support Vector machine) extracts the face candidates and determine the face region by choosing a center among those candidates. Then, the facial features within the detected face are also localized by the similar SVM method used in the face detection task. In the face normalization stage, the face-warping process is performed based upon the detected facial feature points as the control information. The histogram equalization is used to normalize the irregular illumination intensity of the facial image. Finally, the facial expression recognition component classifies the facial expressions by analyzing feature vectors that represent the facial image such as pixel intensity or Gabor representation. For the purpose of performance comparisons, we developed each stage as an independent module.
3 The Facial Expression Image Database and Normalization of the Images

There are several facial expression databases released by researchers. We have used Cohn-Kanade facial expression database [6]. The benefit of using such standard database is that it allows us to compare the facial expression recognition performances between the results from the different laboratories. In the database, each facial expression is recorded as a video sequence. For this study, we chose a maximally expressed image (e.g. happiness) as shown in Fig. 2(a).

![Fig. 2. Illustration of face detection (a) from an image, clipping facial area (b), normalization based on the locations of two eyes (c) and by two eyes and mouth (d), respectively](image)

Normalization in a conventional face recognition system is a necessary step, since the input image may contain a face with the different scales or a rotated face. In that case, the locations of two eyes are used for face normalization as shown in Fig. 2(c). However, in the facial expression recognition system, not only the locations of the two eyes but also the location of the mouth is used in warping the face for the purpose of normalization as shown in Fig. 2(d).

4 Support Vector Machines for Face and Facial Feature Detection

As SVM is a supervised learning machine, it requires a group of training images [10]. Normally, a set of images is prepared for the positive answers, and the other set for the negative answers. For the face detection task, the former corresponds to the face images, the later the non-face images. Notice that each face image in our case contains a facial expression. Following the training session, a window trained with positive and negative sets is sliding over the input image from upper-left to bottom-right to search potential face region to detect any facial area within the given image. When the window finds several face candidates, a central point among those candidates is determined as the center of the face. Because the central point of the face is located around the nose, the upper part of the point could contain two eyes and the lower part of the same point should have a mouth region. Similar to the face detection case, the training sets for the eye and mouth detection contain both the positive and negative answers. The detection accuracy of the face can be measured using the Receiver Operating Characteristic (ROC) curve analysis. In Fig. 3(a), the ROC curves were obtained by varying the number of training images. The detection accuracy of the face is improved
as the number of the training images increases. Similarly, three ROC curves in Fig.
3(b) show how the detection accuracy of the eye within the detected face is improved
as the number of the training images increases.

![ROC curves for locating face (a) and eyes (b), respectively](image)

**Fig. 3.** ROC curves for locating face (a) and eyes (b), respectively

For the purpose of performance comparisons, we construct 3 different types of geo-
metric normalization based on selected facial area only (Fig. 2(b)), locations of two eyes
(Fig. 2(c)), and locations of two eyes and mouth (Fig. 2(c)), respectively as described in
section 6. After the geometrical normalization, an illumination normalization step is
carried out using the histogram equalization method as illustrated in Fig. 1.

## 5 Facial Expression Recognition

### 5.1 Gabor Wavelets Representation for the Facial Expression Images

Daugman has shown an effective way of representing 2D image: Gabor wavelets
representation [1]. Since then, it has been popular method for representing faces,
fingerprints and facial expressions. As the Gabor wavelets consists of orientation
component as well as spatial frequency components, they have an advantage in repre-
senting any object in an image, compared the convectional edge detectors, which
tends to focus on orientation of the object. Note that spatial frequency representation
has translation invariant property that is important in case of slight translation error
exists. In our study, we used five frequencies (from 0 to 4) and eight orientations at
each image point within the grid (22×22), that drawn over the normalized facial
image (88×88 pixels).

### 5.2 EFM Classifier

The EFM classifier determines the discriminative features for the reduced image space.
We introduce it to improve on the generalization ability of the standard Fisher linear
discriminant (FLD) based methods. The EFM applies first Principle component analysis
A Fully Automatic System Recognizing Human Facial Expressions

(PCA) for dimensionality reduction before proceeding with FLD type of analysis. Then, it discriminate the reduced PCA subspaces [8].

6 Performance Evaluation

We compare the facial expression recognition rates between four different methods of facial expression recognition: the PCA based method (PCA), the EFM based method (EFM), PCA based method with Gabor representation (Gabor+PCA), and the EFM based method with Gabor representation (Gabor+EFM) as shown in Fig. 4 and Fig. 5. Each horizontal axis represents the number of features (or components) and each vertical one indicates the recognition rate. In this experiment, we first sampled the

![Face detection only (no registration)](image-a)

![Manual registration (Eyes+Mouth)](image-b)

![Fig. 4. Performance comparison between no-registration (a) and the manual registration case (b)](image-fig-4)
half of the image database for the training purpose. The test was carried out using the remaining half of the images. Then, we repeated these tests for 20 times and averaged the results.

For the manual registration, although we used the same database set, three locations of two eyes and mouth were selected by clicking the mouse. In Fig. 4, recognition accuracy between two cases of geometric normalization is compared: the face detection only case without any facial feature registration and the manual registration case of eyes and mouth. For the most part, the manual registration case outperforms than the no-registration case in a great deal. It is clear that the normalization with facial features results in higher accuracy of the facial expression recognition. Notice that the Gabor representation boosts up the performance at least 10%. And the EFM method is much better than the PCA method.
Fig. 5(a) plots how the performance of the facial expression recognition is varied depending on classification methods. Similar to the previous case, the EFM classifier performs well in the automatic registration case. And the best result was obtained when the EFM was combined with the Gabor representation. For instance, when the 64 features were used, its correction rate was around 90%. This is slightly lower than the best case of the manual registration case, which is 93%, as shown in Fig. 5(b). Note that the overall performance of the system was the best when the registration was done by hand. The next best one was when both the locations of eyes and mouth were accounted in registration stage. And, the poor result came from the no-registration case. All these results suggest that the normalization is a very important stage within the facial expression recognition system.

7 Discussion and Conclusions

Our main contribution in this paper is the automation of face detection as well as facial feature detection for the facial expression recognition system. Although the performance of the automatic case is slightly lower than that of the manual one, it is still similar to that of the average human observer for the same task [2]. We also have confirmed that the Gabor wavelet is an effective way of representing the facial expression image. Moreover when such representation was combined with a high performing classifier, namely, the EFM method, the best outcome was obtained. Examining the results of different geometric normalization cases, we have found that the normalization stage is a crucial step to obtain a high performance for the facial expression recognition system.

References

1. J. Daugman, Uncertainty relationship for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America A, 2: 1160 – 1169, 1985.
2. M. Dailey, G. Cottrell, and C. Padgett, EMPATH: A neural network that categorizes facial expressions, Journal of Cognitive Neuroscience, 13, 513 – 532, 2002.
3. C. Darwin, The expression of emotions in man and animals. John Murray, London, 1872.
4. G. Donato, M. Bartlett, J. Hager, P. Ekman and T. Sejnowski, Classifying facial actions, IEEE PAMI, 21, 10, 974 – 989, 1999.
5. P. Ekman and W. Friesen, Unmasking the Face. A guide to recognizing emotions from facial clues. Palo Alto. Consulting Psychologists Press, 1975.
6. T. Kanade, J. Cohn, and Y. Tian, Comprehensive database for facial expression analysis, Proc. Int’l Conf Face and Gesture Recognition, 46 – 53, 2000.
7. M. Lades, J. Vorbruggen, J. Buhmann, L. Lange, C. von der Malsburg, R. Wurz, and W. Konen, Distortion invariant object recognition in the dynamic link architecture, IEEE Computers, 42, 300 – 311, 1993.
8. C. Liu and H. Wechsler, Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition, IEEE Image Processing, 11, 4, 467 – 476, 2002.
9. Y. Tian, T. Kanade and J. Cohn, Recognizing action units for facial expression analysis, IEEE PAMI, 23, 2, 97 – 115, 2001.
10. V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
A Study of the Radial Basis Function Neural Network Classifiers Using Known Data of Varying Accuracy and Complexity

Patricia Crowther, Robert Cox, and Dharmendra Sharma
School of Information Sciences and Engineering, University of Canberra, ACT 2601, Australia

Abstract. Neural networks are increasingly used in a wide variety of applications such as speech recognition, diagnostic prediction, income prediction and credit screening. This paper empirically compares the performance of Radial Basis Function (RBF) and Multilayer Perceptron (MLP) neural networks using artificially generated data sets, enabling us to accurately chart the effectiveness of each network type and to provide some guidance to practitioners as to which type of network to use with their data. We find that when the discriminator is simple, RBF and MLP network performances are similar; when the number of data points is relatively small the MLP outperforms the RBF; when the discriminator is complex the RBF outperforms the MLP; and when the data has an unrelated input and the underlying discriminator is simple, the MLP outperforms the RBF.

1 Introduction

It has been claimed that Radial Basis Function (RBF) networks have been experimentally and theoretically proved to be better generalisers than Multilayer Perceptron (MLP) [1]. We have no argument with the theoretical and mathematical basis for this assertion. However, in a real life context, the problem domains are limited, and experimental results are unlikely to match the theory. We show that in some circumstances, MLP networks will outperform RBF networks.

Various studies have compared the performance of MLP and RBF networks in areas such as speaker recognition [2], sensor validation of flight control systems [3], mass identification in mammograms [4], facial expression recognition [5] and deviation signals for on-line identification of a synchronous generator [6]. Only [5] found that an MLP network outperformed an RBF; the others found that RBF networks outperform MLP networks for their applications.

This paper empirically compares the performance of RBF and MLP neural networks using artificially generated data sets. This enables us to accurately chart the effectiveness of each network type and to provide some guidance to practitioners as to which type of network to use with their data, particularly when they have confidence in their knowledge.
of the nature of that data. When choosing what type of neural network to use in evaluating a problem domain, a practitioner must consider the nature of the data.

This paper draws on a larger study of results of artificially generated data sets available at the University of Canberra website.

2 Characteristics and Accuracy

We generate training data that has known characteristics, vary those characteristics and test the effects on accuracy. In addition, we compare the results for RBFs to those from MLP networks. We use a ‘base line’ data set for comparison with the subsequent experiments, varying one or more characteristics of the base set.

One of the characteristics varied was the discriminator. The discriminator is an imaginary line (or lines) described by a mathematical function that is used to put the data into its classifications. The neural networks used in this paper provide two outputs, so values on one side of the discriminator represent one of the outputs, values on the other side, the other output. For a real life data set, it is not possible to know the formula for the ‘correct’ discriminator, but because we are generating our data sets artificially, we are able to define what the discriminator should be (these are the underlying functions we use to generate our data sets). We can then compare the values that a trained neural network would give with the correct values. Figure 1 shows one of the discriminators used in this study. The two different data points are shown by ‘+’ and ‘x’. The black line shows the discriminator.

![Fig. 1. Data in the left graph contains no errors (all x data points are above the discriminator; all + data points are below). The right hand graph is a dataset with 30% bad data. Bad data points are represented by a square (+ is in the x zone or x is in the + zone)](image)

Neural networks are considered to be good generalisers - they deal well with bad data, effectively being able to ignore at least some badly classified data points. Two terms are used widely in this paper, observed accuracy (OAcc) and true accuracy (TAcc). Observed accuracy is accuracy of the network as it would be observed by someone who knew little about the data. True accuracy is based on the underlying mathematical function that is used to generate the data. To determine TAcc, the output of the ANN is compared against the data function used to generate the data.
Both Oacc and TAcc are expressed as a percentage of correctly classified data points. For each experiment we ran 20 runs and averaged the results.

3 Data and Networks

Most experiments had a data dimension of 2 and were split into 2 classification outputs. The classification outputs were given values of 1 and 2. A bad data point is defined as a point that is misclassified in training, testing and validation data. A good data point is defined as a point that is correctly classified in training, testing and validation data. The domain of the data was such that: $0 \leq x, y \leq 1$

For most data sets a continuous discriminator function was used and points were classified depending on whether their x input was below the discriminator function threshold. For one experiment (the three circles data set) a discontinuous data function was used. Data was generated with the following known parameters; number of data points in the training and testing data sets; number of data points in the validation set (usually 1000); underlying data distribution; error distribution; error rate expressed as a percentage (100 points with an error of 10% would contain 90 correct points and 10 incorrect points). No measure of fuzzy accuracy was used – a given point was classified as being either correct or incorrect.

In this paper, we use two types of neural network; MLP - traditional multilayer perceptron neural networks with one hidden layer trained with traditional back propagation and RBF - classifiers that use a radial basis function as an activation function, with the positioning of the nodes generated by a k-means algorithm. The RBF output layer is trained using a linear function.

For both MLP and RBF networks, training was done by an automated trial and error mechanism. This training mechanism gives good results when compared with networks where a human has determined these parameters. It does not guarantee the best possible network but tests show it produces a consistently ‘good’ network.

4 Results

4.1 Similar Results for a Simple Discriminator

RBF and MLP neural networks produce very similar results when the underlying discriminator is simple, whether or not the distribution of errors is skewed. The underlying discriminator (shown in figure 1) has the formula:

$$y = \frac{1}{2} \sin \frac{3\pi x}{2} + \frac{1}{2}$$

(1)

Figure 2 shows the results of the base line experiment comparing a RBF to and MLP. The RBF network exhibits very close to the same results as an MLP for the same data set. Its accuracy is very slightly less, but not enough to be significant when looking at observed accuracy. The difference for the true accuracy is slightly greater. When the distribution of errors was skewed to concentrate towards the point (0,0), the RBF and MLP networks still performed very closely.
4.2 MLP Outperforms RBF When There Are Fewer Data Points

An MLP network appears to outperform an RBF network when there are fewer data points. Table 1 compares RBF and MLP networks with 500 and 250 testing and training data points. It can be seen clearly that the difference between the MLP and RBF networks performance is greater for 250 than for 500 data points. This suggests that as the size of a data set reduces, the MLP will perform better than the RBF.

Table 1. Selected experimental results for MLP and RBF networks using data sets with 250 data points and 500 data points

| Error | Observed Accuracy | True Accuracy |
|-------|-------------------|---------------|
| %     | 500 Points | 250 Points | 500 Points | 250 Points |
|       MLP  | RBF   | MLP  | RBF  | MLP  | RBF  | MLP  | RBF  |
| 0     | 98.78 | 97.71 | 98.15 | 96.57 | 98.78 | 97.71 | 98.15 | 96.57 |
| 5     | 93.83 | 92.81 | 93.03 | 91.28 | 98.68 | 97.56 | 97.85 | 95.93 |
| 10    | 88.76 | 87.65 | 88.01 | 86.66 | 98.35 | 97.16 | 97.53 | 95.85 |
| 15    | 83.81 | 82.53 | 82.75 | 81.06 | 98.18 | 96.47 | 96.72 | 94.36 |
| 35    | 63.02 | 61.29 | 61.13 | 59.99 | 92.08 | 87.52 | 86.84 | 81.92 |
| 40    | 57.23 | 55.89 | 56.24 | 55.11 | 87.32 | 81.33 | 82.18 | 76.65 |
| 45    | 52.42 | 51.96 | 51.70 | 51.22 | 75.88 | 68.54 | 67.02 | 64.51 |

4.3 RBF Outperforms MLP When There Is a Complex Discriminator

When the discriminator is more complex, the RBF network will outperform the MLP network. We found that as the complexity of the discriminator increased, the RBF networks outperformed the MLP networks, while at the same time the ability of the
neural network to classify the data diminished. For this set of experiments, the underlying functions were the Chebyshev polynomial orders 1 to 9 (see Figure 3).

![Image](https://via.placeholder.com/150)

**Fig. 3.** Underlying functions – Chebyshev polynomial. Orders 1, 2, 3, 5 and 9 are shown

We ran four experiments here, each with a different error rate in the training data set. (specifically 0%, 10%, 35%, 45%) Figure 4 shows the comparisons for 10% and 45% with the same experiments run for an MLP network.

![Image](https://via.placeholder.com/150)

**Fig. 4.** Chebyshev experiment COL7b 10% errors; COL7d 45% errors

![Image](https://via.placeholder.com/150)

**Fig. 5.** A discontinuous discriminator (RBF vs MLP) with discriminator function shown

It can be seen that as the discriminator becomes more complex, even with perfect training data (0% errors) the RBF outperformed the MLP to a significant extent once the complexity of the polynomial exceeded order 5. At higher complexity levels with more errors (order 8 or 9 with an error rate of 45%) the MLP outperforms the RBF,
but only by a small amount. A possible explanation for the convergence of accuracy at order 9 would be that the networks have trained to a simpler linear discriminator, which would slightly favour the performance of the MLP over the RBF.

Not all 2-dimensional discriminators can be described by a continuous function. As an RBF network is based on a clustering algorithm, we would expect the RBF to perform well with a discontinuous discriminator, e.g., the three circles data set (fig. 5). We see from figure 5 that the RBF network performs considerably better than the MLP network for this particular discontinuous discriminator.

4.4 MLP Outperforms RBF on Introduction of an Unrelated Input

A neural network can be configured to accept a large number of inputs, and can have more than two possible outputs. Most of the experiments in this paper have two inputs (x and y co-ordinates) and two outputs – above or below the discriminator. With real life data however, it is common to have many more inputs. Fisher’s iris data, for example, has four inputs (petal and sepal widths and lengths of various iris flowers) and three outputs (the variety - setosa, virginica or versicolor) [7]. Often a neural network practitioner is not sure whether or not one or more of these inputs is related to the solution being sought. We have found that if one of your inputs is unrelated to the data, then an MLP network outperforms an RBF network.

The data for the first two inputs is generated in the same way as the base line data. The third input, however, is generated completely randomly; it has no relationship to the first two columns. We show that the extra, unrelated input reduces the accuracy of the RBF network. The left hand graph in figure 6 compares the 250 point data set produced in the second finding with a 250 point data set with an extra, unrelated input, showing that an extra input reduces the effectiveness of the RBF network.

The right hand graph in figure 6 shows the results of the comparison between RBF and MLP networks with an unrelated third input. This clearly shows that an MLP network is a better classifier where unrelated inputs are included, particularly if
the amount of data available is relatively small. This result is particularly relevant to practitioners working with biological data, where large numbers of inputs are used, often with little convincing evidence that they are in any way related to the output.

### 5 Conclusions and Future Directions

A neural network practitioner must carefully consider the nature of their data when choosing a neural network classifier. From the reported experiments conducted on RBFs and MLPs on known data, it is concluded that if a large number of possibly unrelated inputs is used, an MLP would be a much safer classifier to use. Also, if you have a relatively small number of data points, an MLP would again be a safer choice. If, on the other hand, you are reasonably convinced that all your input data is related to the problem at hand, the underlying relationships are complex (leading to a complex discriminator) and you have sufficient data points, then the use of an RBF network would be a better approach. Future work will expand on the types of data sets and error distributions used, along with the inclusion of other neural network types. Our aim is to eventually build a new and better type of neural network, using the data accumulated in these and further experiments to validate its effectiveness.

### References

1. Kasabov, Nikola K.: Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. The MIT Press Cambridge, Massachusetts (1996)
2. Finan, R.A, Sapeluk, A.T., & Damper, R.I: Comparision of Multilayer and Radial Basis Function Neural Networks for Text-Dependent Speaker Recognition. Proceedings International Conference on Neural Networks (ICNN'96) 4 1992—1997
3. Giampiero Campa, Mario L. Fravolini, Marcello R. Napolitano, Brad Seanor, Neural Networks-Based Sensor Validation for the Flight Control System of a B777 Research Model. American Control Conference 2002, May 8-10, (2002).
4. Keir Bovis, Sameer Singh, Jonathan Fieldsend, Chris Pinder. Identification of masses in digital mammograms with MLP and RBF nets. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000 Vol 1 IEEE 2000 p 342-347.
5. Gargesha, M., Kuchi, P., Facial Expression Recognition using Artificial Neural Networks. EEE 511 – Artificial Neural Computation Systems, Spring 2002.
6. Park, Jung-Wook, Harley, R.G., Venayagamoorthy, G.K., Comparison of MLP and RBF Nerual Networks Using Deviation Signals for On-Line Identification of a Synchronous Generator. Proc. 2002 IEEE PES Winter Meeting, New York, Vol. 1, pp 274-279, January 2002.
7. Fisher, Robert: The Use of Multiple Measures in Taxonomic Problems. Ann. Eugenics, vol. 7, (1936) pp. 179-188
8. Machine learning repository: http://www.ics.uci.edu/~mlearn/MLSummary.html
9. Pandya, Abhijit S., Macy, Robert B.: Pattern Recognition with Neural Networks in C++. CRC Press (1996)
10. Ceccarelli, M., Hounsou, Joel T.: RBF Networks vs Mulitlayeer Perceptrons for Sequence Recognition. Neural Networks Theory, Technology, and Applications IEEE Technical Activities Board ed Patrick K. Simpson (1996)
Top Down Modelling with Genetic Programming

Daniel Howard

Biocomputing and Developmental Systems Group,
Computer Science and Information Systems, University of Limerick, Ireland
danielhoward@sunrisemalvern.freeserve.co.uk

Abstract. This paper explores the connection between top down modelling and the artificial intelligence (AI) technique of Genetic Programming (GP). It provides examples to illustrate how the author and colleagues took advantage of this connection to solve real world problems. Following this account, the paper speculates about how GP may be developed further to meet more challenging real world problems. It calls for novel applications of GP to quantify a top down design in order to make rapid progress with the understanding of organizations.

1 Top Down Modelling

One simple definition of ‘top down modelling’ is using intuition and experience to arrive at one or more astute subdivisions of the solution. When considering a computer based solution these astute subdivisions correspond to components of the architecture of a computer program (the organization of the main program, the number and re-use of: memory; iteration; recursion; conditional statements; and subroutines). The intention of top down modelling is greater efficiency: efficiency in finding the solution; efficiency of the resulting solution.

Intuition and experience can cause the investigator to prescribe many of these subdivisions a-priori. In the context of GP the investigator can accomplish this in a plethora of ways. For example, the investigator can prescribe how many parameterized subroutines, also known as Automatically Defined Functions (ADFs), the solution will be allowed to invoke, and what number of parameters the ADFs will take. The investigator may assign a certain task to each ADF. When separating amino acid chains into those which have trans-membrane domains and those which do not, Koza deliberately forced the computer program to use an iteration step to populate a memory that was then acted upon by a results producing branch [1]. A simple way to prescribe the nature of the solution a priori is by the choice of functions and terminals that are offered to GP. In image analysis problems the image features (GP terminals) may be determined a priori rather than discovered by GP. In summary, these and other methods enable the combination of inductive learning and of analytical knowledge in solution search.

However, the investigator may refrain from specifying the ‘advantageous architecture’ and allow GP the freedom to discover it. Then the top down architectural sub-divisions that are advantageous for finding a good solution will be automatically discovered by a GP run (using fitness directed trial and error).
ADFs and other constructs that are manipulated by GP aid solution search (pattern exploration and pattern exploitation [3]) automatically discovering an advantageous solution architecture.

Working on the solution architecture has intended and unintended effects. It may result in efficient data processing by the final solution. It may also encourage a more human-readable (understandable, comprehensible) solution (model).

Typically, a ‘top down’ model will establish a hierarchy in terms of that which is most stable (less likely to change) and that which is most likely to change. This can be exploited by the designer or by the GP process. The persistent structure (parameterized subroutines that are discovered by GP) can be stored in function libraries [2]. These functions can be called upon to solve a different problem. Re-use of readily available functions in libraries to solve a different problem has a ‘bottom up’ modelling quality.

2 Genetic Programming

GP is a flexible technology and ideal candidate for top down modelling. Why? Because GP easily manipulates computer software components such as variables, arrays, memories, functions, function sub-programs, parameterized subroutines, iterations and recursions. And additionally, GP has numerous advantages and qualities that make it a method of choice for solving real world problems:

- GP manipulates high level computer programming constructs and this contributes to its transparency and to the ease of interpretation of the resulting expressions (‘white box’ quality).
- GP can input a large number of variables (terminals) or introduce them by point mutation during the course of a run [3] and a considerable body of empirical evidence suggests that it does a good job at reducing this input set to that which provides the highest information value.
- GP can manipulate a number of incompatible building blocks (function types) and enforce rules about how to combine them together [4] [5].
- The solution to the problem may be evolved in one representation (using a tree structure) but implemented in another representation (the resulting expression is simplified and implemented as compiled C or optimized code in Assembler).
- GP variables (population size, selection strategy) can be set up to result in solution search that is as greedy or as gentle as necessary (unlike backpropagation, evolution is not a greedy search).
- GP can use silent (also known as neutral) mutation to search the solution space [5].
- The concept of working with a population can be readily exploited. It can implement different selection strategies (ranking vs roulette wheel, migration between populations, speciation, or methods of maintaining diversity in the population) as well as graduating individuals to different fitness levels [6], and thus it can be said that evolutionary computation in general and GP in particular offer unprecedented flexibility in terms of solution search.
– GP is a natural fit for multi-objective optimization; in a single run GP can use its population and its selection strategies to evolve the receiver characteristics curve (rather than a single solution).
– The population concept enables GP to produce more than one solution to a problem, which is important for problems that require a number of different but equivalent solutions.
– GP may facilitate the discovery of solutions using the co-evolutionary paradigm; co-evolution can change the landscape of the search or it can be applied to a game theory type of problem [7].
– GP can implement supervised as well as unsupervised learning methods.
– GP can implement separation of the genotype (blue print) from the phenotype (the solution); this could be important as there is now empirical evidence that supports the claim that very simple computer programs can generate complex (random) behaviour [8]. So that the ability to evolve a genotype (the simple genotype that may lead to a complex phenotype) arguably is a very powerful tool for modelling complexity.
– GP can exploit parallel computations either as parallel independent runs, parallelization of the GP algorithm and its fitness function, or an extreme parallel computation as an algorithmic chemistry [9].
– GP can evolve a library of functions as well as the solution to a problem. It can be organized into stages and produce explicit building blocks or explicit subroutines for posterior reuse [2] [3].

3 Examples

Machine Vision concerns itself with the recognition of objects in the field of vision. The research described in this section used GP to learn to detect object from non-object. The object was one of large variability and of poor definition (find evidence of a ‘vehicle’ in airborne IR surveillance imagery [10]). Three methods of top-down modelling are now illustrated.

By requirement, the GP solution would need to process vast amounts of imagery in a pixel by pixel fashion. Also the GP evolution process would need to take-in a plethora of examples of ‘non vehicle’ and a far smaller number of examples of ‘vehicle’. Therefore, the evolution was arranged into two evolution stages. The first stage learns to detect all of the examples of ‘vehicle’ w.r.t. a random subset of non vehicle image regions. The GP evolved detector is then applied to the entire training set, and the resulting false alarms are recorded. A second evolution stage then learns to detect all of the examples of ‘vehicle’ from the false alarms that are produced by the first detector. However, the requirement for speed was addressed as follows. The first evolution stage was constrained to work with fewer pixel data and to develop a small structure, while the second evolution stage remained unconstrained (the second stage had the tougher job of learning to discriminate like from like). The resulting detection was very fast because the first detector is applied to millions of pixels, whereas the second and more computationally expensive detector is applied to thousands of pixels (the positive returns from the first detector) [12].
Although the method of stages \cite{11} demonstrates how top down modelling can be combined with GP, any method of AI can be arranged in stages in a similar way. However, the second methodology is more specific to GP. The number of false alarms was unacceptably high, and it became obvious that information concerning the context of detection would be useful to improve the results. A top down modelling decision was taken to imitate a theory of saccades in human vision (see for example the introduction in \cite{13}). By this theory, the eye quickly discovers a number of features and the mind then projects or imagines what it sees. The system constantly reinforces or abandons the belief with real time input from sensors (vision, smell, touch, sound). Therefore, GP was arranged as follows. A data crawler would commence its travels from the pixel that was indicated as residing on a vehicle. The data crawler would feel its way in the surroundings to verify the claim \cite{14}. The multi-part GP program consisted of a results producing branch that moved the crawler; a subroutine that could be consulted to turn the crawler; another subroutine that could be consulted to determine whether to mark the pixel below the crawler as a ‘feature’; finally, another result producing branch examined the resulting binary map of features and non-features and used statistics based on the distribution of features to determine whether the starting point of the crawl was on a ‘vehicle’ or not. These four GP branches used different input terminals (averages over pixels, textural statistics, memories concerning the last few turns of the crawler or position of previously identified features). In summary, top down modelling decides upon the architecture of the resulting solution but GP has the freedom to evolve the important details such as “what is a feature?” “where do I look next?” and “what does this distribution of features mean?” Although other methods of AI could be constructed to perform in a similar way, these would require a lot of thought whereas with GP, implementation was very straightforward \cite{14}.

The third example concerns finding the orientation of a vehicle. Vehicles in IR imagery are roughly rectangular but their edges are very fuzzy and at close inspection look like oblong blobs. It is also the case that edges are not always complete and that the human eye fills in the information to see them. Traditional methods (Hu moments) proved useless to determine the orientation of these vehicles automatically. A top down decision was taken to find the orientation of the vehicle by comparing the standard deviation across a vehicle to that along a vehicle \cite{15}. The former should be higher than the latter. With the analytical knowledge of the typical width of a vehicle an equilateral triangle was constructed with its corners missing. The triangle was rotated in 5 degree angle intervals and the average and standard deviations computed on each of the three line segments. At each angle of rotation GP used this information to evolve a decision rule to determine the orientation of the vehicle. Again, GP could have been replaced by another AI technique but unlike other methods, GP provides clear formulae that can be understood by humans and which can be simplified and implemented as a general purpose and an economical procedure.

An example from the field of Bioinformatics provides another illustration of the top down modelling concept. The expression of the genome is analogous to
the execution of a computer program. Depending on the chemical context of the cell, the genome will execute in different ways to achieve different results. This is the reason why with the same DNA code, there exist 200 types of cells in the human body. Arguably this is the reason why the proteins in human and chimpanzee are nearly identical, although we are in many respects very different living creatures. The expression of the genes is triggered by areas of DNA known as *cis*-acting sites that attract proteins. It is believed that there is a significant difference in those parts of the genome that determine when and where will these genes become expressed (the *cis*-acting sites). From copious *in vitro* experimentation it is known that these *cis*-acting sites are very short motifs, and *in vivo* observation tells us that they can be separated by thousands of nucleotides.

This information was sufficient to develop a top-down design of the problem of finding these *cis*-acting sites in human DNA [16]. A finite state machine was combined with genetic programming motifs at each state of the machine, and the entire thing was evolved using evolutionary strategies with a mutation operator. Whenever a tour visits a state of the machine, the GP motif is applied at that point in the DNA. If there is a match the machine transitions to another state indicated by an arc of the directed graph. Otherwise it visits another state indicated by another arc. Each state contains a number such as ‘100’ that commands a ‘nucleotide jump’ prior to the motif being applied. The machine allows recursion (a state can transition to itself). Furthermore, the tour produces a logical statement (each arc has a boolean function such as XOR, OR, AND, NOR, etc. associated with it). When an individual is selected, each of the many possible mutations (mutate a link, a GP tree, replace a boolean function, modify the nucleotide jump, and so on) is consulted in turn to determine whether it should be applied. Although the procedure worked well [16], confirming the presence of promoters that were identified from *in vitro* experiments, its ability to produce large jumps gives the scheme wide application, and this is owed to the top-down modelling analysis.

4 Future Directions: Socio-Technical Applications

This section speculates on how to develop the concept of Top Down modelling and GP one step further to model more challenging problems such as socio-technical problems. Such problems are subject to the laws of cybernetics that were developed several decades ago to understand the behaviour of machines, animals, and organizations. One idea would be to model such systems by using these Cybernetic Laws or principles as the grammatical rules of a Genetic Programming system. This could be implemented using Grammatical Evolution [5]. Another way to implement this is as a multi-part computer program (similar to that of the data crawler in the previous section). The execution of such a program would be complex as it would involve interactions in the multi-part program such as between a manager and a production unit.

For illustration consider one of these Cybernetic top-down models from the 1970s. Beer postulated that a top down model of his creation known as the Viable
Systems Model (VSM)\cite{17,18} could be applied to any organization to diagnose it, to understand its internal and external workings and to help to improve its viability (its survival). Beer proposed an abstraction into five functional systems which make up the VSM, (see\cite{19} for examples and illustrations). When implementing the VSM as a multi-part computer program using GP, the manager (system 3) and the operational unit (system 1) would agree a ‘resource bargain’\cite{17} and an auditor (system 3*) would check on the honesty of the unit by a random audit. Other parts of the program would set policy (system 5), carry out more long term research (system 4), check for the interaction of the various units and the environment (system 2), see\cite{18}, and so on. The fitness function for the multi-part program would be composed of a number of objectives and measures, considering quality and efficiency of each part and of the whole. Reverse engineering the details of an organization that has been postulated and studied in much greater detail at a higher level of granularity is very relevant to Biology. Practical needs abound in this and in other fields and three of these needs are:

1. Animals are often organized in groupings to hunt. Although the broad architecture of such organizations has been studied, the details are not yet understood. Reverse engineering the details of animal organizations from field data and other assumptions would provide helpful insights. These are required for unravelling the genetics of behaviour.

2. Reverse engineering the organization of mental functions within the confines of a top down model provided by neuro-psychologists\cite{20} (clinical observation of patients) could advance the study of the brain, the understanding of cognition, psychoanalysis, and also Robotics\cite{21}.

3. Systems such as Oracle and SAP capture a plethora of transaction data temporally. This data could be used not only to understand relationships between transactions and customers (as is done at present) but also whether the organization that is responsible for the transactions could be better organized.

Research into application of the VSM as a starting point for GP solutions must answer questions of uniqueness, solvability, and complexity. Uniqueness: is the VSM any different from any other abstraction for modelling a viable system in a complex environment? Solvability: what examples can be solved or not solved most appropriately with this model? Complexity: what is the minimum number of rules required to call a system ‘viable’? For example, Cellular Automata systems with a tiny number of rules appear to be ‘viable’, but then again, the real-world environment is much more complex and changing in time. Will solutions find the break down of tasks into these five systems helpful?

**References**

1. J.R. Koza (1994) Genetic Programming II, MIT Press.
2. C. Ryan, M. Kreijzer and M. Cattolico (2004), Favourable Biasing of Function Sets Using Run Transferable Libraries, Genetic Programming Theory and Practice 2004, Kluwer (edited by Rick Riolo, Tina Yu, Una-May O’Reilly and Bill Worzel).
3. Daniel Howard (2003), Modularization by Multi-Run Frequency Driven Subtree Encapsulation, Genetic Programming Theory and Practice 2003, Kluwer (edited by Rick Riolo and Bill Worzel).
4. D. J. Montana (1995), Strongly Typed Genetic Programming, Evolutionary Computation, vol 3, no. 2, pp. 199–230.
5. M. O’Neill and Conor Ryan (2001), Grammatical Evolution, IEEE Trans. Evolutionary Computation vol 5, no. 4, pp. 349–358.
6. E. D. Goodman and Jianjun Hu (2003), The Continuous Hierarchical Fair Competition Model for Sustainable Innovation in Genetic Programming, Genetic Programming Theory and Practice 2003, Kluwer (edited by Rick Riolo and Bill Worzel).
7. Jordan B. Pollack and Alan D. Blair (1998), Co-Evolution in the Successful Learning of Backgammon Strategy. Machine Learning vol. 32, no. 3, pp. 225-240.
8. Stephen Wolfram (2002), *A New Kind of Science*, Wolfram Media Inc., 2002.
9. W. Banzhaf and C. Lasarczyk (2004), Genetic Programming of an Algorithmic Chemistry, Genetic Programming Theory and Practice 2004, Kluwer (edited by Rick Riolo, Tina Yu, Una-May O’Reilly and Bill Worzel).
10. Simon C. Roberts and Daniel Howard (1999) Evolution of Vehicle Detectors for Infrared Linescan Imagery. Lecture Notes in Computer Science 1596, pp. 110–125, Springer.
11. Daniel Howard and Simon C. Roberts (1999) A Staged Genetic Programming Strategy for Image Analysis. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 1999), pp. 1047-1052, Morgan Kaufmann.
12. Daniel Howard, Simon C. Roberts and Richard Brankin (1999) Target Detection in SAR imagery by Genetic Programming. Advances in Engineering Software, 30, pp. 303-311.
13. B. Laeng and D-S. Teodorescu (2002), Eye Scanpaths During Visual Imagery Reenact Those of Perception of the Same Visual Field, Cognitive Science, vol. 26, pp. 207–231.
14. Daniel Howard, Simon C. Roberts and Conor Ryan (2002). Machine Vision: Analysis of Context with Genetic Programming. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2002), pp. 756-763, Morgan Kaufmann.
15. Simon C. Roberts and Daniel Howard (2000), Genetic Programming for Image Analysis: Orientation Detection, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000), pp. 651–657, Morgan Kaufmann.
16. Daniel Howard and Karl Benson (2003), Evolutionary Computation Method for Pattern Recognition of Cis-Acting Sites. Biosystems, vol. 72, issues 1-2 , pp. 19-27. (special issue on Bioinformatics and Computational Intelligence).
17. S. Beer, *The Heart of the Enterprise*, Wiley, 1979.
18. S. Beer, *Diagnosing the system for organizations*, Wiley, 1985.
19. http://www.bogacki.co.uk/C7TER_fig_1.htm, http://www.greybox.uklinux.net/visma_2.2/pdf/visma_2.2.pdf
20. http://www.cs.bham.ac.uk/~axs/fig/your.mind.gif
21. http://www.nesc.ac.uk/esi/events/Grand_Challenges/proposals/ArchitectureOfBrainAndMind.pdf
A Two Phase Genetic Programming Approach to Object Detection

Mengjie Zhang, Peter Andreae, and Urvesh Bhowan

School of Mathematical and Computing Sciences
Victoria University of Wellington,
P. O. Box 600, Wellington, New Zealand
{mengjie,pondy}@mcs.vuw.ac.nz

Abstract. This paper describes two innovations that improve the efficiency and effectiveness of a genetic programming approach to object detection problems. The first innovation is to break the GP search into two phases with the first phase applied to a selected subset of the training data, and a simplified fitness function. The second phase is initialised with the programs from the first phase, and uses the full set of training data to construct the final detection programs. The second innovation is to add a program size component to the fitness function. Application of this approach to three object detection problems indicated that the innovations increased both the effectiveness and the efficiency of the genetic programming search.

1 Introduction

Object detection tasks arise in a very wide range of applications, such as detecting faces from video images, finding tumours in a database of x-ray images, and detecting cyclones in a database of satellite images. Given the amount of data that needs to be detected, automated object detection systems are highly desirable. However, creating such automated systems that have sufficient accuracy and reliability turns out to be very difficult.

Genetic programming (GP) is a relatively recent and fast developing approach to automatic programming [1,2]. In GP, solutions to a problem are represented as computer programs. Darwinian principles of natural selection and recombination are used to evolve a population of programs towards an effective solution to specific problems. The flexibility and expressiveness of computer program representation, combined with the powerful capabilities of evolutionary search, make GP an exciting new method to solve a great variety of problems.

There have been a number of reports on the use of genetic programming in object detection [3,4,5,6,7,8,9]. The approach we have used in previous work [8,9] is to use a single stage approach (referred to as the basic GP approach here), where the GP is directly applied to the large images in a moving window fashion to locate the objects of interest. Past work has demonstrated the effectiveness of this approach on several object detection tasks.
However, this genetic programming approach has several problems — the training time was often very long, even for relatively simple object detection problems and the evolved programs are often hard to understand or interpret. We believe that the large size and the redundancy of the programs contributes to the long training times and may also reduce the quality of the resulting detectors by unnecessarily increasing the size of the search space. Also, evaluating the fitness of a candidate detector program in the basic GP approach involves applying the program to each possible position of a window on all the training images, which is expensive. An obvious solution is to apply the program to only a small subset of the possible window positions, but it is not obvious how to choose a good subset.

This paper describes two innovations on the basic GP approach to address these problems. The first is to split the GP evolution into two phases, using a different fitness function and just a subset of the training data in the first phase. The second is to augment the fitness function in the second phase by a component that biases the evolution towards smaller, less redundant programs. We consider the effectiveness and efficiency of this approach by comparing it with the basic GP approach and a neural network approach. We also examine the comprehensibility of the evolved genetic programs.

2 The Approach

In the first phase of the approach, the genetic programs were initialised randomly and trained on object examples cut out from the large images in the training set — an object classification task which is simpler than the full object detection task. This phase uses a fitness function which maximises classification accuracy on the object cutouts.

In the second phase, the GP process is initialised with the programs generated by the first phase, and trained on the full images in the training set by applying the programs to a square input field (“window”) that was moved across the images to detect the objects of interest. This phase uses a fitness function that maximises detection performance on the large images in the training set.

Because the object classification task is simpler than the object detection task and the fitness function is cheaper, we expect the first phase to be able to find good genetic programs much more rapidly and effectively than the second phase. Although simpler, the object classification task is closely related to the detection task, so we believe that the genetic programs generated by the first phase are likely to be very good starting points for the second phase, allowing the more expensive evolutionary process to concentrate its effort in the more optimal part of the search space.

Since the difficulty of finding an optimal program increases with the size of the programs, in the second phase, we added a program size component to the fitness function to bias the search towards simpler functions. We expect this to increase both the efficiency and the effectiveness of the evolutionary search and have a tendency to remove redundancy, making the programs more comprehensible.
3 Image Data Sets

We used three data sets of increasing difficulty in the experiments. Figure 1 shows example images. Data set 1 (Shape) consists of artificially constructed circles and squares against a uniform background. Data set 2 (Coins) consists of scanned images of New Zealand 5 and 10 cent coins, showing both heads and tails, at random orientations. The classes of interest are the 5 cent coins and the 10 cent coins. Data set 3 (Heads/tails) consists just of 5 cent coins on a non-uniform background, and the two classes of interest are the heads and the tails. Given the low resolution of the images, this detection task is very difficult—even humans cannot distinguish the classes perfectly. In the experiments, we used one, three, and five images as the training set and used five, ten and ten images as the test set for the Shape, Coins, and Heads/tails data sets, respectively.

![Figure 1. Object Detection Problems](image)

4 Experiment Configuration

In this system, we used tree structures to represent genetic programs. The terminals (features) consisted of the mean and standard deviation of local square and circular regions, as shown in figure 2. The function set consisted of the four standard arithmetic operators and a conditional operator. The division operator represents a “protected” division in which a divide by zero gives a result of zero. The conditional operator returns its second argument if its first argument is positive, and otherwise returns its third argument.

The output of a genetic program is a floating point number. For multiple class object detection problems, this value must be mapped to a class name. We used a program classification map [10] to map the output, \( v \), of an evolved genetic program to the class of the object located in the current input field. The output is mapped to background if \( v < 0 \), and to class \( i \) if \((i - 1) T < v \leq i T\), where \( T \) is a threshold defined by the user.
We used two fitness functions for the two learning phases. In the first phase, we used the classification accuracy directly as the fitness function to maximise object classification accuracy. In the second phase, we used a multi-objective fitness function. In previous work [9], we used a weighted sum of three components: the Detection Rate (DR: the number of objects correctly classified by a detection system as a percentage of the total number of objects in the image), the False Alarm Rate (FAR: the number of incorrectly classified objects and non-objects as a percentage of the total number of objects in the images), and the False Alarm Area (FAA: the number of pixels incorrectly reported as object centres).

A problem with this fitness function is that it has no bias towards smaller programs. When a short program and a long program produce the same detection rate and the same false alarm rate (perhaps because one is the same as the other but with additional, redundant elements), the GP system will choose both programs equally. Selecting longer programs will tend to produce further long programs, which will slow the evolution down and possibly decrease the chance of finding a good program. Also, longer programs are usually difficult to interpret. Therefore, for the experiments in this paper, we have augmented above fitness function with a program size component:

$$fitness = K_1 \cdot (1 - DR) + K_2 \cdot FAR + K_3 \cdot FAA + K_4 \cdot ProgSize$$

where progSize is the number of terminals and non-terminals in the program, and $K_i$ are the weightings on the components.

Notice that adding the program size constrain to the fitness function is a kind of parsimony pressure technique [11, 12, 13]. Early work on this issue resulted in diverse opinions on its effectiveness, and a fear that it could lead to premature convergence [13]. Therefore, we used a very small weight for the program size component.

The ramped half-and-half method [1, 2] was used for the initial population and for the mutation operator. The proportional selection mechanism and the reproduction [10], crossover and mutation operators [1] were used in the learning process. We used reproduction, mutation, and cross over rates of 2%, 28%, and 70%, respectively. Table 1 shows the other important experimental parameter values.

---

1 See [10] for details of the moving window detection algorithm.
The evolutionary process was run for a fixed number \((\text{max-generations})\) of generations, unless it either found a program that solved the problem perfectly (100% detection rate and no false alarms), or there was no increase in the fitness for 10 generations, at which point the evolution was terminated early.

### Table 1. Parameters used for GP training for the three data sets

| Parameter name       | Shape | Coins | Heads/tails |
|----------------------|-------|-------|-------------|
| population-size      | 800   | 1000  | 1600        |
| initial-max-depth    | 2     | 2     | 5           |
| max-depth            | 6     | 7     | 8           |
| max-generations      | 50    | 150   | 200         |
| input-size           | 20×20 | 72×72 | 62×62       |

| Parameter name       | Shape | Coins | Heads/tails |
|----------------------|-------|-------|-------------|
| T                    | 100   | 80    | 80          |
| K_1                  | 5000  | 5000  | 5000        |
| K_2                  | 100   | 100   | 100         |
| K_3                  | 10    | 10    | 10          |
| K_4                  | 1     | 1     | 1           |

## 5 Results and Discussion

### 5.1 Object Detection Results

The detection results of the two phase GP approach for the three image data sets are compared with the basic GP approach [8, 14] and a neural network approach [15] using the same set of features. For all cases, the experiments were repeated 10 times and the average results on the test set were presented. All three approaches achieved ideal results for the Shape and Coin data sets, and were able to achieve 100% detection rate for the Heads/tails data set. Table 2 shows the false alarm rates for the Heads/tails data set. Clearly, the two phase GP approach achieved the best performance. Notice also that both GP approaches achieved better results than the neural network approach using the same set of features.

### Table 2. False Alarm rates on Heads/tails for different approaches

|                      | heads | tails |
|----------------------|-------|-------|
| Two-phase GP Approach | 0     | 55%   |
| Basic GP Approach    | 0     | 100%  |
| Neural Networks      | 9.4%  | 134.1%|

### 5.2 Training Time and Program Size

Although both of the GP approaches achieved better results than the neural networks overall, the time spent on the training/refining process are quite different. For the Coins data set, for example, the two phase GP approach used 11 hours on average, whereas the basic GP approach required 17 hours on average. For the Heads/tails data set, the two phase GP approach found good programs after 23 hours on average, in contrast to 45 hours for the basic GP approach. The first phase typically required only minutes because the size of the training
data set is small, and evaluating the fitness function is very fast. However, the programs it finds appear to be very good starting points for the more expensive second phase, which enables the evolution in the second phase to concentrate its search on a much more promising part of the search space.

In addition, the programs evolved by the two phase GP approach were smaller than those evolved by the basic GP approach. For the Coins data set, for example, the program size in the two phase GP approach averages 56 nodes, in contrast to 107 nodes for the basic GP approach.

5.3 Comprehensibility of Genetic Programs

To check the effectiveness of the new fitness function at improving the comprehensibility of the programs, we considered one of the evolved programs for the shape data set:

\[
\text{if } ((F_{4\mu} - T)/F_{4\mu}) \text{ then } F_{3\mu} \text{ else } ((F_{4\mu} - F_{2\mu}) \times F_{1\sigma}) / (F_{4\mu}/F_{4\mu})
\]

where \(F_{i\mu}\) and \(F_{i\sigma}\) are the mean and standard deviation of the square region \(i\) (see figure 2 left) of the window, respectively, and \(T\) is the predefined threshold. This program still has some redundancy, but can easily be simplified to:

\[
\text{if } (F_{4\mu} - T) \text{ then } F_{3\mu} \text{ else } (F_{4\mu} - F_{2\mu}) \times F_{1\sigma}
\]

The condition of the if \((F_{4\mu} > T)\) can be straightforwardly interpreted as testing whether a large fraction of the central region of the sweeping window is over an object. The output then depends either on the average value of the 2nd smallest region, which will be higher for the circle objects than for square objects, or the difference between the average of the central and the 2nd largest regions.

While this program detector can be relatively easily interpreted and understood, the programs obtained using the old fitness function are generally hard to interpret due to the length of the programs and the redundancy. The new fitness function appears to make the evolved genetic programs more comprehensible.

6 Conclusions

To improve the effectiveness and efficiency of genetic programs for an object detection task, we developed a two phase GP approach, and augmented the fitness function with a program size component. We tested the approach on three object detection problems of increasing difficulty. Our results suggest that the two phase approach is more effective and more efficient than the basic GP approach and more effective than a neural network approach using the same set of features. The augmented fitness function resulted in genetic program detectors that were better quality and more comprehensible. It also reduced the search computation time.

Although this approach considerably shortens the training times, the training process is still relatively long. We intend to explore better classification strategies and add more heuristics to the genetic beam search to the evolutionary process.
While the programs evolved by this approach are considerably shorter than
the basic GP approach, they usually still contain some redundancy. We suspect
that this redundancy reduces the efficiency and the effectiveness of the evolu-
tionary search, but it may play an important role in the search. We are currently
experimenting with simplification of the programs during the evolutionary pro-
cess to remove all redundancy.

References

1. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D.: Genetic Programming: An
Introduction on the Automatic Evolution of computer programs and its Applica-
tions. Morgan Kaufmann Publishers (1998)
2. Koza, J.R.: Genetic programming : on the programming of computers by means
of natural selection. Cambridge, Mass. : MIT Press, London, England (1992)
3. Tackett, W.A.: Genetic programming for feature discovery and image discrimi-
nation. In Forrest, S., ed.: Proceedings of the 5th International Conference on
Genetic Algorithms, Morgan Kaufmann (1993) 303–309
4. Benson, K.: Evolving finite state machines with embedded genetic programming
for automatic target detection within SAR imagery. In: Proceedings of the 2000
Congress on Evolutionary Computation, IEEE Press (2000) 1543–1549
5. Graae, C.T.M., Nordin, P., Nordahl, M.: Stereoscopic vision for a humanoid robot
using genetic programming. In Cagnoni, S., et al. eds.: Real-World Applications
of Evolutionary Computing, Volume 1803 of LNCS., Springer-Verlag (2000) 12–21
6. Howard, D., Roberts, S.C., Ryan, C.: The boru data crawler for object detection
tasks in machine vision. In Cagnoni, S., et al. eds.: Applications of Evolutionary
Computing, Proceedings of EvoWorkshops2002. Volume 2279 of LNCS. Springer-
Verlag (2002) 220–230
7. Lindblad, F., Nordin, P., Wolff, K.: Evolving 3d model interpretation of images us-
ing graphics hardware. In: Proceedings of the 2002 IEEE Congress on Evolutionary
Computation, IEEE Press (2002)
8. Zhang, M., Ciesielski, V.: Genetic programming for multiple class object detection.
In Foo, N., ed.: Proceedings of the 12th Australian Joint Conference on Artificial
Intelligence, LNAI Vol. 1747, Springer-Verlag (1999) 180–192
9. Zhang, M., Andreae, P., Pritchard, M.: Pixel statistics and false alarm area in
genetic programming for object detection. In Cagnoni, S., ed.: Applications of
Evolutionary Computing, Lecture Notes in Computer Science, LNCS Vol. 2611,
Springer-Verlag (2003) 455–466
10. Zhang, M., Ciesielski, V., Andreae, P.: A domain independent window-approach
to multiclass object detection using genetic programming. EURASIP Journal on
Signal Processing, Special Issue on Genetic and Evolutionary Computation for
Signal Processing and Image Analysis 2003 (2003) 841–859
11. Smith, P.W.H.: Controlling code growth in genetic programming. In John, R.,
Birkenhead, R., eds.: Advances in Soft Computing, De Montfort University, Le-
icester, UK, Physica-Verlag (2000) 166–171
12. Dallaway, R.: Genetic programming and cognitive models. Technical Report CSRP
300, School of Cognitive & Computing Sciences, University of Sussex,, Brighton,
UK (1993)
13. Soule, T., Foster, J.A.: Effects of code growth and parsimony pressure on populations in genetic programming. Evolutionary Computation 6 (1998) 293–309
14. Bhowan, U.: A domain independent approach to multi-class object detection using genetic programming. BSc Honours research project/thesis, School of Mathematical and Computing Sciences, Victoria University of Wellington. (2003)
15. Ny, B.: Multi-class object classification and detection using neural networks. BSc Honours research project/thesis, School of Mathematical and Computing Sciences, Victoria University of Wellington (2003)
Mapping XML Schema to Relations Using Genetic Algorithm

Vincent Ng, Chan Chi Kong, and Stephen Chan

Department of Computing,
Hong Kong Polytechnic University
cstyng@comp.polyu.edu.hk

Abstract. As web-applications grow in number and complexity, there is a need for efficient mappings from XML schemas to the flat relational tables so that existing functions in relational database systems can be utilized. However, many of the existing mapping methods, such as the model-based or the structure-based methods, do not exploit query history for better query performance. In this paper, we propose the use of genetic algorithm (GA) in a cost-based approach for converting a XML schema to relational tables. By formulating the mapping problem as a cost optimization task with respect to a set of weighted frequent queries, we can obtain an efficient mapping that minimizes the queries execution time. In our experiments, we show that the mapping obtained by GA is superior to other non-cost-based approaches. In particular, the GA approach has out-performed the greedy heuristic in the browsing queries where the accessed attributes are many and scattered.

Keywords: XML schema, relational database, genetic algorithms

1 Introduction

XML has become a popular data exchange format in the World Wide Web in recent years. The data they access, on the other hand, are still frequently stored using well established relational database technology, and there is no trivial way to find efficient mappings from the semi-structured XML schemas to the flat relational ones. In many commercial database system such as [1] and [2], manual mapping is needed.

Proposals for storing XML documents using relational model come in two main research directions. In the model-based approaches [3][4], relationships between nodes and paths in XML documents are stored in tables regardless of the node type, whereas in the structure-based approaches, the schemas of the XML documents are used to derive corresponding relations. Examples of the latter approach include [5], where the XML DTDs are first simplified and then, starting from the document root, each descendent of the root that has a 1-1 correspondence are inlined into the same relation. This method also forms the basis of several other works such as [6], which formulates the mapping using regular tree grammar, and [7][8], which define rules to preserve some of the semantic information present in the DTD.

The structure-based approaches attempt to inline as many elements into the same relation as possible, without taking any access patterns or statistics into account. The
effectiveness of the resulting relational schemas are therefore often sub-optimal. Applying this inlining approach to the Medline DTD [11], for instance, will result in a schema that is consisted of one relation with 70 attributes that are not necessarily closely related, and 36 others smaller relations with 10 attributes or less. It would be therefore beneficial if relevant access statistics can be used for fine-tuning the conversion process.

Data and queries statistics are used in some approaches such as STORED[15] and XStorM[16]. In both works, the semi-structured data and frequent queries are analyzed using data-mining techniques and the most frequently occurred patterns are mapped into relations. However, not every node is mapped into relations, as the rest of the data are still stored in a semi-structured overflow section.

A cost-based approach that stores XML using entirely RDBMS can be found in [9], where, instead of inlining as many elements as possible as in structure-based approaches, a greedy algorithm is used to decide the set of elements to be inlined by minimizing the access cost of a set of important queries. However, the greedy algorithm may not be the best choice of different types of queries and it had been experimented with a small dataset. Hence, we are interested to adopt the genetic algorithm, which is proven to be one of the useful optimization tools, to assist the relational mapping of XML data.

In this paper, we present an approach for converting XML schemas to relations using genetic algorithm. A set of important queries is assumed to be available and their frequency and access pattern are used in guiding the mapping process. This method is adapted from a technique we used in database vertical partitioning [10]. By considering the queries access patterns during the conversion process, our task share some similarity with the vertical partitioning problem. As with the original vertical partitioning algorithm in [10], we assume the element access patterns, and some simple statistic such as frequency of each sample query and the average branch factor of each node in the DTD-tree are available, and are represented by an Attribute Usage Matrix (AUM). The AUM is then processed by a GA based search routine to generate a relational schema.

The remaining of this paper is organized as follow. Section 2 describes the access cost model. Section 3 describes the setting of the genetic algorithm we used for the conversion. Section 4 describes an alternative approach that uses greedy algorithm. Section 5 presents the experiment results for various access patterns from two available databases. Section 6 concludes the paper.

2 The Access Cost Model

We start with the DTD-tree presentation of the XML schema, together with a set of important representative queries for the XML document we interested in. In the DTD-tree we choose not to distinguish element nodes from attributes nodes but instead we treat as another element node and marked it the symbol ‘a’. Also we assume that the DTD-tree is normalized such that each node can have one parent node only. For simplicity, DTDs with recursive/cyclic definitions are not handled in this study. Regarding the representative queries, instead of limiting our study to any particular query languages, we summarize the queries using an AUM. The AUM, as shown in
figure 1, is a matrix where each row corresponds to the access pattern of a query, and each column corresponds to a node (element or attribute) in the DTD-tree. Each entry $A_{ij}$ of the AUM represents the frequency that a node $j$ in the DTD-tree is accessed during the retrieval of a single XML record resulted from an execution of query $i$. It is assumed that the required number of accesses to each resultant, conditional or intermediate node is coded this fashion. An example of this encoding is given in figures 2 to 4. The DTD in Figure 2 can be represented using a corresponding DTD-tree as given in Figure 3.

| Attribute | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------|---|---|---|---|---|---|---|---|---|----|
| Query 1   | 1 | 0 | 0 | 0 | 10| 0 | 10| 0 | 15| 0  |
| Query 2   | 0 | 15| 15| 0 | 0 | 0 | 0 | 15| 15| 0  |
| Query 3   | 0 | 0 | 0 | 20| 0 | 20| 0 | 0 | 0 | 20 |
| Query 4   | 0 | 10| 0 | 0 | 0 | 0 | 10| 10| 0 | 0  |
| Query 5   | 1 | 10| 10| 0 | 10| 0 | 20| 20| 20| 0  |
| Query 6   | 1 | 0 | 0 | 0 | 1 | 0 | 10| 0 | 0 | 0  |
| Query 7   | 0 | 0 | 1 | 10| 0 | 0 | 0 | 1 | 0 | 0  |
| Query 8   | 0 | 0 | 10| 10| 0 | 15| 0 | 0 | 10| 10 |

Fig. 1. Attribute Usage Matrix

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT A (B, C+)>  
<!ELEMENT B (D+, E)>  
<!ELEMENT C (F)>  
<!ATTLIST C        
    G CDATA #REQUIRED >  
<!ELEMENT D (#PCDATA)>  
<!ELEMENT E (#PCDATA)>  
<!ELEMENT F (#PCDATA)>  
```

Fig. 2.

Suppose that in this example the average branch factor for the edge AC in the document is 10, that is, there are 10 occurrences of node C, F and G for each occurrence of the root node A. The corresponding XPath query
//A[B/E='value']/C would be represented in the AUM as Query1 in Figure 4 (node F and G are returned as the children of node C).

| Attribute | A | B | C | D | E | F | G |
|-----------|---|---|---|---|---|---|---|
| Query 1   | 1 | 0 | 10| 0 | 1 | 10| 1 |
| Query 2   | ...| ...| ...| ...| ...| ...| ...|

Fig. 4.

The AUM, together with the frequency and the selectivity (the percentage of records in the document that match the query’s selection criteria) of each sample query, will be fed into the GA search routine described in section 3. The output is a set of Boolean values $\text{InliningSchema} = \{O_1, O_2, O_3...O_n\}$, where $n$ is the number of nodes.
in the DTD-Tree, and each InliningStatus $O_i$ is a Boolean deciding whether the $i^{th}$ node will be inlined (a value of 1 means the node is inlined, whereas 0 means the node, and its descendents, are stored in different relations as its parent).

In our approach, the value of each InliningStatus $O_i$ is determined by a GA search with the goal of minimize the total access cost for the frequent queries. Exception to that are the set-type nodes whose parent link are marked by * or +, in which case the corresponding value for $O_i$ is fixed at 0, and the attribute nodes whose parent link are marked by $a$, in which case the corresponding value for $O_i$ must be 1. As an example, suppose the InliningSchema for the DTD-tree in figure 3 is $\{1, 1, 0, 0, 1, 0, 1\}$, then the corresponding relational schema will contain 4 relations as shown in figure 5. Note that under this representation, AUM entries for nodes that are mapped to the same relation would have the same value. In addition to the nodes assigned, each relation will contain an $id$ column which serves as the primary key, and, for each relation not containing the root node, a foreign key column referencing the parent relation as shown in figure 6.

The heart of the search routine is a cost formula for evaluating the InliningSchemas. The database operating cost for any schema is comprised of page access cost, update cost, logging cost, integrity checking cost and storage cost. In this study, we consider only the page access cost because it dominates all the other internal processing costs. Hence, the cost of transaction processing is measured only in terms of the number of pages accessed in order to satisfy the transaction. In order to calculate the access cost, a simulation is performed using a Tuple Usage Matrix (TUM) which is a $t \times q$ matrix where $t$ is the number of records in the document and $q$ is the number of queries. Each element $T_{ij}$ is a Boolean representing whether the $t^{th}$ record satisfies the selection constraint of the $j^{th}$ query and hence some of its nodes will be returned as specified by the AUM. Each $T_{ij}$ will be randomly initialized at the being of the mapping procedure such that each $T_{ij}$ will be set to 1 (meaning the record is retrieved) with a probability $S_j$, where $S_j$ is the selectivity of query $j$.

For simplicity, we assume each attribute in the resulting relations is to be accessed through an unclustered index. We further assume a translation table exists for mapping between DTD nodes and the attributes of the resulting relations. In this way, the number of accesses to the database files can be simplified into the number of data blocks in each relation that contain at least one node that is, as specified by the AUM and the TUM, required to be accessed.

Under these assumptions, the access cost will be expressed as the number of page accesses incurred. Since all disk accesses are assumed to be through unclustered index scan, the prefetch blocking factor can be ignored. Hence, the cost of processing a query $q$ using a storage structure with InliningSchema $P$, where the accessed attributes are stored in $m_p$ relations, can be expressed in terms of the number of disk page accessed as :

$$C_p = \sum_{q=1}^{Q} f_q \sum_{j=1}^{m_p} \sum_{i=1}^{b_{pj}} B X_{pqij}$$

where $Q$ is the total number of queries, $f_q$ is the frequency of query; $B$ is the page block Size; $m_p$ is number of resulting relations in the InliningSchema $P$; $X_{pqij}$ is 1 if any tuple in the block $i$ in relation $j$ is retrieved by query $q$ in the InliningSchema
$P$, and is 0 otherwise; $Q$ is the total number of queries; $b_{pj}$ is the number of block required for a relation $j$; $C_P$ is the Total cost of database access using storage structure with InliningSchema $P$. The number of blocks required for a relation $j$ is given by

$$b_{pj} = \lceil C_R / [B / L_{pj}] \rceil$$

where $L_{pj}$ is the length of a relation $j$ in an InliningSchema $P$; $C_R$ is the cardinality of the relation which is given by $N * A$, where $N$ is the number of records in the XML document and $A$ is anyone of the equal-value entries in the AUM for representing the frequency of the nodes in that relation.

The $C_P$ expression calculates the total sum of the number of pages required by all queries. Based on the previous assumptions, all other overheads incurred during the query resolution are ignored. The decision variable $X_{Pqij}$ can be determined from the input Attribute Usage Matrix and the Tuple Usage Matrix (TUM). An example of TUM is given in Figure 7.

| Record | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|--------|---|---|---|---|---|---|---|---|
| 1      | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 2      | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 3      | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 4      | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 5      | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6      | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |

Fig. 7. Tuple Usage Matrix

A penalty cost will be imposed when multiple relations are required to satisfy a query. It is because, as in vertical partitioning problems, accesses to multiple relations incur greater overhead cost than accesses to a single relation [12]. Therefore, another objective is to minimize the number of relations that will be accessed by each transaction while reducing the number of irrelevant attributes retrieved. To simplify the calculation, we assume that the penalty cost increases linearly with the additional number of relations required. Hence, the revised cost for our scheme is as follows:

$$C_P = \sum_{q=1}^{Q} f_q (1 + (m_{pqu} - 1) W) \sum_{j=1}^{m_p} \sum_{i=1}^{b} X_{Pqij}$$

where $W$ = Penalty Cost Factor that is configurable in the cost model, $m_{pqu}$= number of relations accessed by query $q$ in InliningSchema $P$. Our experiment suggests that a small amount of penalty cost, i.e. 10% of the original cost, will serve well for the secondary objective.
3 Applying Genetic Algorithm

In this section we outline the setting of the genetic algorithm we used in the conversion process. As with traditional genetic algorithms representation, each chromosome is a binary string of genes, with each gene either being one or zero [13]. In our case the number of genes is equal to the number of nodes in the DTD-tree such that a value of 1 in a gene would mean the corresponding node will be inlined with respect with its parent. Each gene in a chromosome thus corresponds to the InliningStatus of a node whose value is to be decided by GA search, with the exception that genes corresponding to nodes whose parent link are marked by ‘*’ or ‘+’ or ‘a’ are fixed at 1, 1, and 0 respectively throughout the process, and are not affected by the GA crossovers or mutations operations. In our implementation, the order of the nodes in the chromosomes is determined by a Breadth-First-Search (BFS) in the DTD-tree. For example, in the figure 3 case, a chromosome will have the format \{C_A, C_B, 0, 0, C_E, C_F, 1\}, where C_i is the InliningStatus of a node labeled I. The chromosome \{1, 1, 0, 0, 1, 0, 1\}, for instance, respresents a relational schema that creates separate relations for the subtrees of node C, node D and node F, as shown in figure 6.

Apart from BFS, we also looked at other ordering options including Depth-First-Search and random ordering in other experiments. No significant difference is found in the resulting InliningSchemas they produce however. BFS nodes ordering is assumed for the rest of this paper.

3.1 Initialization

Our initial population is generated by randomly assigning each non-fixed gene in each of the chromosome to either zero or one. Each chromosome of the initial population will be evaluated by a fitness function which maps each chromosome in the population to a real value cost.

3.2 Evaluation

The total cost of a InliningSchema is the sum of disk data block accesses incurred by each transaction plus the penalty cost incurred when multiple relations are accessed by the transaction. Thus, the objective of the search process is to minimize the total cost.

3.3 Selection

In order to produce new offsprings, chromosomes from current generation are selected as parents with a probability proportional to the distance from the best cost in the current generation. In each generation, the highest (i.e. worst) cost, \(成本_{max}\), and the lowest (i.e. best) cost, \(成本_{min}\) among the current chromosomes is identified. For each chromosome \(i\), a relative fitness score, \(f_i\) is calculated as

\[
f_i = (成本_{max} - cost_i) / (3 * (成本_{max} - 成本_{min}))
\]
Here, \( \text{cost}_i \) is the total access cost of the chromosome \( i \). The multiple of 3 in the denominator is a value chosen after fine-tuning to avoid the selection process being too biased toward the leading chromosomes.

Each time a parent is needed, each chromosome \( i \) would have a probability \( \text{prob}_i \) to be selected, where

\[
\text{prob}_i = \frac{f_i}{\sum_j f_j}
\]

### 3.4 Reproduction

One point crossover is selected as the reproduction operator. In order to prevent a loss of diversity, each offspring will undergo mutation with a probability \( \mu_a \) which is a configurable parameter. This operation is performed after the crossover. The mutation operator randomly selects a non-fixed gene and then inverse the gene from 0 to 1 for from 1 to 0.

### 3.5 Replacement

Two replacement schemes are studied. In the first approach, only the best chromosome is allowed to survive to the next generation, and the rest of the next generation is to be generated by reproduction. In each reproduction two parents are selected from the population and they reproduce two children, and the child with the lower cost will be added to the next generation.

In the second approach, we assign each newly created chromosome with a \( \text{life} \), which is initially set to 1. In each generation, each chromosome \( i \) would have a chance of dying as calculated by

\[
\text{ProbDie}_i = (1 - f_i) \times (1 - \text{life}_i)
\]

Each chromosome that does not survive will be replaced by a new one generated by reproduction. Each chromosome that survive will has its \( \text{life} \) factor decreased by a predefined \text{aging-factor} of value between 0 and 1

\[
\text{life}_i = \text{life}_i \times \text{aging-factor}
\]

Our experiment shows that the result obtained using both approaches are very close, which suggests the more simple approach, the first one, is already sufficient for the task, and is therefore preferred.

### 3.6 Termination

The GA search terminates when there is no more improvement of the best-fit chromosome for a consecutive number of generations (50 in our experiments), or after the maximum number of chromosomes has been evaluated (20000 in our experiment setting).

In summary, the algorithm in applying the genetic algorithm in finding an effective relational mapping is shown in Figure 8.
initialize population $P(0)$
evaluate $P(0)$
g = 1
repeat
  $P(g) = \{ \text{chromosome with max. fitness in } P(g-1) \}$
  for $i = 1$ to population_size do
    $x(t) = \text{select-parent}(P(g-1))$
    $y(t) = \text{select-parent}(P(g-1))$
    $P(g) = P(g) \cup \text{crossover}(x(t), y(t))$
    mutate($P(g), \mu_a$)
  evaluate($P(g), AUM, TUM$)
  $g = g + 1$
until (termination criteria are met)
return Schema

Fig. 8. Algorithm for GA to find a relational mapping

4 Greedy Search

Apart from GA, we also implemented a greedy-search adapted from [9] for comparison. There are two variants in the approach. In the first variant, labeled Greedy_In, all non-fixed nodes in the InliningSchema are initialized to 1, meaning all inlinable nodes are inlined; whereas in the other Greedy_Out variant, all non-fixed

minCost = $\infty$
inliningSchema = GetInitialSchema(DTD)
cost = GetSchemaCost(inliningSchema, AUM, TUM)
while (cost < minCost) do
  minCost = cost
  candidateSchemaList = ApplyFlippings(inliningSchema)
  for each candidateSchema $\in$ candidateSchemaList, do
    cost’ = GetSchemaCost(candidateSchema, AUM, TUM)
    if cost’ < cost then
      cost = cost’
inliningSchema = candidateSchema
  endif
endfor
endwhile
return Schema

Fig. 9. Algorithm for Greedy Heuristic to finding a relational Mapping
node are initially outlined (inliningStatus=0). In both cases, the nodes are then “flipped” greedily: Within each iteration, the inliningStatus of one of the nodes is flipped from 0 to 1 or from 1 to 0. The node that is flipped is the one that would produce the highest reduction in cost. The greedy algorithm is shown in figure 9. The greedy search ends when no more cost-reduction can be obtained by flipping.

Experiments for both the GA and greedy approaches are described in the next section.

5 Experiments

Our experiments are designed to evaluate the performance of GA and Greedy Search under typical web-application scenarios. In web applications, there are, in general, two types of queries. A user typically first spends time in browsing through the web information to look for something interesting. In this phase we expect the issued queries are more scatter in term of accessed attributes and there are few identifiable patterns. After that, when a user knows what to look for, he can then navigate to obtain the desired information by issuing more specific queries with more meaningful access pattern. Both the browsing and navigating modes are scenarios to be studied in our experiments.

For the navigating mode experiments, we used DTD from two real-life applications:

1. The Internet Movies Database DTD (IMDB) [14]
2. Medline DTD [11]

For the 34 nodes IMDB case, we followed the 20 sample queries defined in [9]. In [9], 15 of the 20 queries are described as “lookup” queries, which are simple select-join-project (SJP) queries accessing relatively few (less than 10) nodes with low selectivity (5%). They represent the type of enquiry frequently occur in web applications. The remaining 5 queries are described as “publishing” queries, which are requests that access a large number of nodes and returning a large amount of data (selectivity ranging from 20% to 100%). In our experiment, the performance of the algorithms is tested using a series of different mix of the two query types ranging from 100% lookup to 100% publishing queries. This is also the approach used in [9] and would enable us to examine the performance of the algorithm under different workloads.

For the 102 nodes Medline case, we defined 10 lookup type SJP queries with less than 20 nodes each, and 5 publishing queries each returning complete sections of the documents with 30 nodes or more. The algorithm is tested using different mix of the two query types as in the previous case. The selectivity for all queries is fixed at 10%.

For the browsing mode experiments, six tests are performed using the Medline data. In order to model the more scatter nature of the browsing mode queries, the AUM is used to directly generate 10 uniform distributed random access patterns in each test instead of starting with sample SJP queries and represent them using AUM.

1 Actually we simplified the DTD slightly. The original DTD contains more than 102 nodes.
Each pattern accesses between 15 and 30 nodes and it corresponds to a browsing query where the selectivity is fixed at 10%.

In each experiment, we tested the performance of the following approaches:

- **All-Inlined**: This is the non-cost-based approach that forms the basis of the algorithm in [5][6][7]. All nodes, except the ones connected by * or + edges, are inlined.
- **All-Outlined**: The opposite of All-Inlined. In this approach all nodes in the DTD-tree are outlined, meaning each node will form its own relation, with the exception of attribute nodes, which are indicated by ‘a’ edges.
- **Greedy_In**: This is one of the two variants of a greedy search approach proposed in [9] which is described in section 4.
- **Greedy_Out**: The other variant of the greedy approach as described in section 4.
- **GA**: Our approach as described in section 3.

The results are shown in figure 10 to figure 12. In each case we calculated the improvement of the heuristic $h$ (where $h = \text{GA, Greedy}_\text{In or Greedy}_\text{Out}$ ) over that of the all-inlined or all-outlined approach, whichever is better, as follow:

$$\text{Improvement}_h = \frac{\min(\text{Cost}_{\text{all in}}, \text{Cost}_{\text{all out}}) - \text{cost}_h}{\min(\text{Cost}_{\text{all in}}, \text{Cost}_{\text{all out}})}$$

Where $\text{Cost}_{\text{all in}}, \text{Cost}_{\text{all out}}$ and, $\text{Cost}_h$ are the costs for the all-inlined, all-outlined and the heuristic (GA, Greedy_In, Greedy_Out) respectively. A positive value improvement indicates a more efficient mapping is obtained compared with that of the non-costed-based approaches.

![Fig. 10](image-url)
Figure 10 and 11 show the results for navigation mode experiments for IMDB and Medline respectively. In figure 10, both GA and the greedy heuristics show improvements between 7% to 25% over that of the non-cost-based approaches, especially for the query-mixes that contain more publishing queries. The results are close as GA obtained better mapping than both Greedy_In and Greedy_Out in 2 of the 11 cases, while obtained the same result in the remaining 9 cases. One interesting point is that, contrary to the finding in [9], we found that in our experiments, the end result costs obtained from Greedy_In and Greedy_Out are not the always same. For cases when they differed, the better one is always used when comparing to GA and the non-cost-based approaches.

![Navigating Mode: Medline](image)

**Fig. 11.**

Figure 11 shows the results for Greedy_In and GA for various query mixes for the larger medline DTD. The result of Greedy_Out actually turned out to be worse than that of the non-cost-based results in all cases and are therefore not shown because of the figure scale. The results show that GA out-performed Greedy_In significantly in 3 of the 9 cases, while the remaining six cases are tied.

Figure 12 shows the results for the browsing mode experiments using Medline. In this set of experiments GA is the clear winner in all the cases. In fact, GA turns out to the only heuristic that out-performed the non-cost-based approaches consistently in all test cases.

6 Conclusion

We studied the use of genetic algorithm in the mapping of XML schemas to relational schemas. By representing a set of sample queries as an Attribute Usage Matrix, we could use GA to select the set of nodes in the DTD-Tree that are to be outlined in
separate relations. The resulting improvement of GA over other non-cost-based approaches is significant as it is the only heuristic that consistently out-performs the non-cost-based approaches in all cases. Comparing the with another cost-based approach that uses greedy-algorithm, the result of GA is matching and sometimes better for small data sets, and is much better for data sets where the number of accessed attributes is large, for both everyday browsing and navigating type web-accesses.

References

1. Josephine Cheng and Jane Xu, XML and DB2, In Proceedings of 16th International Conference on Data Engineering (2000) pp. 569-576
2. Oracle Corporaion, XML Support in Oracle 8 and beyond, Technical white paper, http://otn.oracle.com/tech/xml/htdocs/xml_twp.html
3. Florescu D. and D. Kossman, Storing and Querying XML Data Using an RDBMS, IEEE Data Engineering Bulletin 22(3), (1999) pp. 27-34
4. Tadeusz Pankowski, XML-SQL: An XML Query Language Based on SQL and Path Tables, Lecture Notes in Computer Science 2490 (2002) pp. 184-209
5. Jayavel Shanmugasundaram, Kristin Tufte, Chun Zhang, Gang He, David J. DeWitt, Jeffrey F. Naughton: Relational Databases for Querying XML Documents: Limitations and Opportunities. VLDB 1999: pp. 302-314
6. Murali Mani and Dongwon Lee, XML to Relational Conversion using Theory of Regular Tree Grammars, Lecture Notes in Computer Science 2590 (2003) pp. 81-103
7. Dongwon Lee, Murali Mani, Wesley W. Chu, Efficient Schema Conversions between XML and Relational Models (Invited Paper), Workshop on Knowledge Transformations for the Semantic Web (2002)
8. Dongwon Lee and Wesley W. Chu, CPI: Constraints-Preserving Inlining Algorithm for Mapping XML DTD to Relational Schema, *Data and Knowledge Engineering (DKE), Vol. 39, No 1* (2001) pp. 3-25

9. P. Bohannon, J. Friere, P. Roy, J. Simeon. From XML Schema to Relations: A Cost-based Approach to XML Storage. In *Proceedings of the 2002 International Conference on Data Engineering, Feb* (2002) pp. 64 - 80

10. Vincent Ng, D.M. Law, N. Gorla. and C. K. Chan, Applying Genetic Algorithms in Database Partitioning, In *Proceedings of the 2003 ACM Symposium on Applied Computing* (2003) pp. 544-549

11. Medline DTD http://www.nlm.nih.gov/databases/dtd/nlmmedline_citation_021101.dtd

12. B. Niamir, Attribute partitioning in a Self-Adaptive Relational Database System, PhD Dissertation, MIT Lab. for Computer Science (1978)

13. Lance Chambers, *Practical Handbook of Genetic Algorithms, Vol. 1*, CRC Press (1995)

14. Internet Movies Database. http://www.imdb.com

15. Deutsch, M. Fernandez, D. Suciu. Storing semistructured data with stored, *SIGMOD Int'l Conf. on Management of Data*, (1999) pp. 431-442

16. Wen Qiang Wang, Mong-Li Lee, Beng Chin Ooi, Kian-Lee Tan: XStorM: A Scalable Storage Mapping Scheme for XML Data, Poster Proceedings of the Tenth International World Wide Web Conference, (2001) pp. 176-177
Diagnosing the Population State in a Genetic Algorithm Using Hamming Distance

Radu Belea¹, Sergiu Caraman¹, and Vasile Palade²

¹ “Dunarea de Jos” University of Galati, Domneasca Str., No. 47
800008 Galati, Romania
{Radu.Belea, Sergiu.Caraman}@ugal.ro
² Oxford University, Computing Laboratory, Parks Road,
OX1 3QD, Oxford, UK
Vasile.Palade@comlab.ox.ac.uk

Abstract. In the literature, the term premature convergence of the entire population is used with the meaning of closing the evolution before reaching the optimal point. It can be emphasized only on a test function with known landscape. If the function landscape is unknown, one can notice the population convergence only. This paper aims to answer to the question: “how can we influence the control parameters of the genetic algorithm so that the exploration time of the parameter space be longer and the risk of premature convergence be reduced?”. The answer to the above question implies choosing a crossover operator with good performances in the landscape exploration and the use of two performance indicators for the detection of the population convergence. In choosing the control parameters of the genetic algorithm, the fitness function landscape must be taken into consideration.

Keywords: genetic algorithm, binary-coded genes, real-coded genes, chromosomes, Hamming distance, uniform crossover, arithmetic crossover.

1 Introduction

In a genetic algorithm the population has the natural tendency to concentrate in an ecological niche. Once the population settled in a restricted area, the individuals are slightly differentiated, the evolution stops and the genetic algorithm degenerates into a local searching algorithm. If the fitness function is not static, then the population convergence prevents the population to adapt to the new environment conditions.

The Hamming distance between two chromosomes was initially used in the early studies on the crossover genetic operator, e.g. in Booker’s study (Booker, 1987), where the uniform crossover was suggested, or in Eshelman’s research (Eshelman, 1991), where the HUX crossover (highly disruptive form of crossover) is proposed. In (Rana, 1999), the author presented a survey on the crossover operator programming methods using the following concepts: crossover mask, Hamming mask and Hamming distance. Also Rana suggests the use of Hamming distances histogram between parents and offsprings, in order to appreciate the quality of the crossover operator.
The aim of this study is to analyse the situations where the genetic algorithm population degenerates, and the methods by which the degeneration of the population can be avoided. The paper has the following structure:

Section 2 presents, in brief, the gene coding technique proposed in (Belea and Beldiman, 2003). The representation of the real parameters through the already known techniques - binary-coded genes and real-coded genes - is replaced by a single representation where we can use the genetic operators defined for the binary-coded genes as well as genetic operators defined for real-coded genes.

Section 3 presents the Hamming distance between two genes, and the Hamming distances histogram is used to compare the binary crossover performances with the arithmetic crossover performances. This analysis is necessary because the exploring function of the parameter space depends mainly on the crossover operator quality used by the genetic algorithm.

In Section 4 we emphasize the definitions for the degenerated population and for the degenerated genetic algorithm. As a consequence, two statistical indicators, the average Hamming distance between individuals and the spread of the Hamming distances are defined. They are used to signal the risk of the convergence.

Section 5 presents a case study based on the following principle: “if the mutation probability is bigger, the genetic algorithm is similar to the local searching algorithm”. A difficult problem was chosen for the searching algorithm: “a crest inclined by an angle of $\pi/4$ in relation with the coordinate axes”. Some conclusions are presented in Section 6.

2 Unified Representation of Real Parameters

A real floating point number is stored as a string of binary figures like that:

$$b_1b_2b_3b_4 \cdots b_{8n_b}$$  \hspace{1cm} (1)

where $n_b$ is the number of bytes used for representing the number. The value of the floating-point real number is:

$$V = (-1)^S \cdot 2^{N-n_{off}} \cdot (b_{msb} \cdot F)$$  \hspace{1cm} (2)

where $S$ is the sign bit, $N$ is the string of bits which stores the exponent, $n_{off}$ is the exponent offset, $b_{msb}$ is the most significant bit of the mantissa and $F$ is a string of bits which stores the fractional part of the mantissa.

As in the representation technique of real numbers the mantissa is normalised, all numbers have $b_{msb} = 1$, and then $b_{msb}$ is not necessary to be stored. It is easy to notice that the real number 0 cannot be represented as a real number with normalised mantissa. This is the reason why number 0 is codified by storing of all null bits. The natural number $n_{off}$ is necessary for storing both the positive exponents and the negative ones, in situation where $N$, the string of bits that stores the exponent, cannot have only null bits for a non-zero real number.
For the genetic algorithms, the floating-point representation of the real number has an inconvenience: the bits in the same position in two different real numbers have different weights. However, in the real number floating-point representation with normalised mantissa, we have intervals like:

\[ x \in [2^k, 2^{k+1}), \quad k \in [-n_{off}, n_{off}) \]  

(3)

where all real numbers in this interval have the same sign and exponent. The mantissa bits in the same position have the same weight. Consequently, both crossover genetic operators and mutation operators can be used in the same way as they have been defined for binary-coded genes or for real-coded genes.

In (Belea and Beldiman, 2003), the authors used the IBM representation of a real number in single precision using 4 bytes, where: \( b_1 \) is the sign bit, \( b_2 \cdots b_9 \) are the exponent bits, \( n_{off} = 127 \) and \( b_{10} \cdots b_{32} \) are the mantissa bits. A gene is a real number \( g \in [1, 2) \) and the decoding function \([1, 2) \to [x_{\min}, x_{\max})\) of the genetic information is linear:

\[ x = x_{\min} + (x_{\max} - x_{\min})(g - 1) \]  

(4)

where \( x \in X \) is the value of the real parameter, \([x_{\min}, x_{\max})\) is the parameters space \( X \). The numbers in the interval \([1, 2)\) have \( S = 0 \), \( N = 01111111 \), thus the exponent is \( N - 127 = 0 \) and the mantissa may have values from combination \( 000 \cdots \) up to combination \( 111 \cdots 1 \). This representation has the following advantages:

- the chromosome is a vector of real numbers;
- both genetic operators defined for the binary-coded genes and genetic operators defined for real-coded genes can be used;
- the decoding of the genetic information is made with a linear function (4);
- all the arithmetic calculations are made by the arithmetic coprocessor.

In the genetic operators programming, a constraint should be taken into consideration: the genetic operators will not modify the bits that transmit the sign and the exponent.

For the previously presented gene codifying technique, the random initialisation of the genes is made with (5):

\[ g = \text{Random} + 1 \]  

(5)

where the \textit{Random} function generates a real random number, uniformly distributed within the \([0, 1)\) interval. In (Belea and Beldiman, 2003), there were presented the statistic tests for checking the random initialisation of genes and the programming appropriateness of the genetic operators.
3 The Hamming Distance

The Hamming distance between two genes $g_1$ and $g_2$ is the natural number:

$$d_h(g_1, g_2) = bc(g_1 \oplus g_2)$$  \hspace{1cm} (6)

where $bc(\cdot)$ is the function *bit count*, which returns the number of bits 1 of a bit string, and $\oplus$ is the logic *exclusive or* operator applied to the binary figures in the same position. For example, taking two four-bit genes $g_1 = 0111$ and $g_2 = 0110$, the Hamming distance between these two genes is:

$$d_h(g_1 \oplus g_2) = d_h(0111 \oplus 0110) = bc(0001) = 1$$  \hspace{1cm} (7)

Using the Hamming distance, we can make a statistical analysis of the randomly initialised genes. The following criterion has been proposed to appreciate the quality of the crossover operator:

**Criterion:** The crossover operator exploring the parameter space is more efficient if the Hamming distances between the parent chromosomes and the offspring chromosomes are bigger.

Let us consider a species with a single gene, $p_1$ and $p_2$ - two parents, and $o_1$ and $o_2$ their offsprings resulted using the binary crossover. Only the bits which differ will be affected by the crossover, so we deduce the following relations:

$$d_h(p_1, o_1) = d_h(p_2, o_2)$$
$$d_h(p_1, o_2) = d_h(p_2, o_1)$$  \hspace{1cm} (8)

Figure 1 comparatively presents the Hamming distances between parents and offsprings histograms for the one point crossover and the uniform crossover. The test followed this procedure: two different crossover operators have been applied to a randomly initialised gene pair, the operation was repeated 1000 times, then the distances $d_h(p_1, o_1)$ and $d_h(p_1, o_2)$ histogram was graph-bar represented.

![Fig. 1. The comparison between one point crossover and uniform crossover](image)
The real parameters representation method presented in section 2 allows the comparison of the genetic operators designed for binary-coded genes with the genetic operators designed for real-coded genes. In figure 2, the uniform crossover is compared to the arithmetic crossover.

![Figure 2. The comparison between the uniform crossover and the arithmetic crossover](image)

The statistic test has been done under the same conditions as in the case presented in figure 2. The arithmetic crossover was implemented by the equations:

\[
\begin{align*}
o_1 &= r \cdot p_1 + (1-r) \cdot p_2 \\
o_2 &= (1-r) \cdot p_1 + r \cdot p_2
\end{align*}
\]

where \( r \in [0,1] \) is a random real number uniformly distributed. However the arithmetic crossover has a major disadvantage: it cannot explore the entire searching space. Figure 3 presents a problem where the parameter space is \((x_1, x_2) = [0,1] \times [0,1]\). The crossover implemented by equations (9) will produce two offsprings located in the rectangle bound by points \((x_{1,A}, x_{2,A})\) and \((x_{1,B}, x_{2,B})\) which represents the positions of A and B parent chromosomes in the searching space.

![Figure 3. The space effectively explored by the arithmetic crossover](image)

4 Indicators of the Population State

The term *premature convergence* is used with the meaning of evolution closing before reaching the optimal point. But the term is not sufficient to explain what occurs in the population and how the genetic algorithm works in this situation. For this we define the followings:

**Definition 1.** A degenerated population is a population with insignificant differences between individuals.

**Definition 2.** A degenerated genetic algorithm is a genetic algorithm that operates on a degenerated population.

A degenerated population is approximately uniform and all the individuals have about the same fitness function value. Any selection method applied to a degenerate
population is arbitrary. Likewise, by matching two individuals that are different only by a few bits, the crossover operator is no longer decisive in transmitting the genetic information from parents to offsprings. The genetic algorithm with binary-coded genes degenerates in the *bit climbing* algorithm and the genetic algorithm with real-coded genes degenerates in the optimisation technique named *relaxation method*.

Let two individuals’ chromosomes \( c_i \) and \( c_j \), and \( g_i[k] \) and \( g_j[k] \), \( k = 1, \ldots, K \), the genes of the two chromosomes. If the genes are codified by the same number of bits, the Hamming distance (6) can be extended to the distance measurement between the two chromosomes:

\[
d_{h}(c_i, c_j) = \frac{1}{K} \sum_{k=1}^{K} d_h(g_i[k], g_j[k])
\]

(10)

Subsequently two state indicators are proposed, which will be able to measure the degree of the population degeneration. Let consider \( P_t \) the population from generation \( t \). The average Hamming distance between the chromosomes of the population is:

\[
\mu(t) = \frac{2}{N(N-1)K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{K} d_{h}(g_i[k], g_j[k])
\]

(11)

where \( N \) is the size of the population. The spread of the Hamming distances between individuals in population \( P_t \) is:

\[
\sigma(t) = \sqrt{\frac{2}{N(N-1)K} \left[ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{K} d_{h}^2(g_i[k], g_j[k]) \right] - \mu^2(t)}
\]

(12)

For population \( P_t \), the average Hamming distance between individuals and the spread of the Hamming distances between the individuals are the two state indicators that allow us to appreciate the degeneration degree of the population. For any given species, the state indicators \( \mu(t) \) and \( \sigma(t) \) do not depend on the number of genes of individuals and on the number of individuals in the population.

5 Experimental Results

It is well known that the premature convergence risk depends on the fitness function landscape, the selection method, the performances of the genetic operator “crossover” and the mutation probability. In order to check the crossover operator performances, a test function of crest type, inclined by an angle of \( \pi / 4 \) in relation with the coordinate axes, has been chosen (Figure 4).

The peak is \( f(x_1, x_2) = 1 \), and is obtained for \( x_1=2/3 \) and \( x_2=5/7 \). In this case, the mutation plays a minimum role. The local searching algorithms are efficient if there is
a path which, through steps parallel to the axes, connects any point from the parameter space to the optimum point. In case the crest is inclined (Figure 4), the path mentioned above is difficult to find.

![Figure 4. The landscape and the level curves of the test function](image)

The following problems make use of a genetic algorithm running under the following conditions: the parameter space is $X \in [0,1] \times [0,1]$, the population is made of 100 individuals, out of which 50 individuals are replaced in each generation. The elitism surviving method was not used. The proportional selection method is used because it is the most exposed to an early convergence risk. The probability of mutation has been used as recommended by Goldberg in (Goldberg, 1991). The mutation probability is different from one problem to another.

**Problem 1:** to run the genetic algorithm using the arithmetic crossover. Mutation probability is 0.

Figure 5 illustrates the genetic algorithm applied to problem 1. In the upper time diagram, the fitness function value of the best-adapted individual in the current generation was plotted with a thick line, while the thin line represents the average fitness value of the population. In the lower time diagram, the function $\mu(t)$ (the average Hamming distance between individuals) was plotted with a thick line, and $\sigma(t)$ (the spread of the Hamming distances between individuals) was plotted with a thin line.

![Figure 5. Convergence test for uniform crossover](image)  ![Figure 6. Convergence test for hybrid crossover](image)
Problem 1 considered $p_m = 0$, so that the mutation does not affect the population evolution. It can be noticed a typical case of premature convergence $f(X_{\text{max}}) < 1$. As seen in the two figures, the population convergence occurred after 30-40 generations. The previous test was resumed using only the uniform crossover. Similar results were reached, except that the population convergence occurred 5-10 generations later.

From relation (9) and Figure 3, it follows that the arithmetic crossover fails to explore the entire searching space. There is not such a disadvantage with respect to the uniform crossover. Subsequently we named hybrid crossover a combination between the uniform and the arithmetic crossover.

**Problem 2:** to run the genetic algorithm using the hybrid crossover. Mutation probability is 0.

Figure 6 illustrates the genetic algorithm applied to problem 2. We looked for the best combination between the two types of crossover. The longest time of population convergence was reached by the hybrid crossover with the probability of the uniform crossover 0.7 and the probability of the arithmetic crossover 0.3.

**Problem 3:** to run the genetic algorithm with hybrid crossover of probability 0.7 and 0.3, respectively, and mutation probability $p_m = 0.05$.

Figure 7 shows a realisation of the genetic algorithm running for problem 3. On the right hand figure, it can be noticed that, after 100 generations, an optimum point was reached and part of the population is concentrated in the ecological niche where the optimum is to be found, while another part is dispersed within the searching space. However exploring the searching space is not good, because the mutations displace the individuals from the ecological niche only along the coordinate axes.

**Problem 4:** to run the genetic algorithm with hybrid crossover of probability 0.7 and 0.3, respectively, and mutation probability of $p_m = 0.05$. Every generation, a number of 10 individuals is randomly initialised.
Figure 8 illustrates a realisation of the genetic algorithm for problem 4. On the right-hand figure, it can be seen that, after 100 generations, the optimum point was reached and the population is more clearly diversified than in problem 3. Looking on the left-hand figure, the exploring of the searching space seems to be better than in problem 3.

6 Conclusions

Section 2 deals with a new gene encoding technique that allows the use of genetic operators designed for both the binary-coded genes and real-coded genes. This method has been used in section 3 for comparing the two types of genetic operators. Using the criterion defined in section 3, it results that the arithmetic crossover has the biggest Hamming distances between parents and offsprings, but it doesn’t explore the whole parameter space. In section 5, the use of hybrid crossover is proposed. The longest exploration time of the parameter space has been obtained for a uniform crossover probability of 0.7 and an arithmetic crossover probability of 0.3.

The results shown in figure 2, in general, are not valid, except for an initial population that has been randomly generated. As the population develops, its composition changes too, and it is only the state indicators $\mu(t)$ and $\sigma(t)$ (defined in section 4) that can provide an indication on the degree to which the population converges toward a certain point. The two indicators have been used as a performance criterion in problems 1-4 from section 5. In section 4, the notions of degenerated population and degenerated genetic algorithm have been defined.

Figures 5 and 6 show two cases of population degeneration of a genetic algorithm. Such undesired situations are caused by the null mutation probability. Figure 7 presents an accurate running of the genetic algorithm with no early convergence, despite low mutation probability. That happens because the exploration of the searching space is not properly done. Figure 8 shows an improved situation, due to the random reset of 10% of the population in each generation.
References

1. Booker, L.: Improving Search in Genetic algorithms. in L. Davis (ed.), Genetic Algorithms and Simulated Anealing, Chapter 5, pp. 61-73. Morgan Kaufman Publishers, Los Altos, CA, 1987.

2. Belea, R. and Beldiman, L.: A New Method of Gene Coding For a Genetic Algorithm Designed for Parametric Optimization. The Annals of “Dunarea de Jos” University of Galați, Fascicle III (2003), pp. 66-71.

3. Eshelman L. J.: The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. “Foundations of genetic Algorithms”, edited by Gregory Rowlins. Morgan Kaufmann Publishers, Los Altos, CA (1991), pp. 265-283.

4. Goldberg D. E.: Genetic Algorithms. Addison-Wesley USA, (1991). French translation, Copyright © June 1994, Editions Addison-Wesley France, S. A.

5. Michalewicz Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag (1994).

6. Rana S.: Examining the Role of Local Optima and Schema Processing in Genetic Search. PHD Thesis, Colorado State University, Fort Collins, Colorado (1999).
Optimizing a Neural Tree Using Subtree Retraining

Wanida Pensuwon¹,², Rod Adams², and Neil Davey²

¹ Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, 40002, Thailand
Wanida.Pe@Ubu.ac.th

² Department of Computer Science, Faculty of Engineering and Information Science, University of Hertfordshire, Hatfield, Herts, AL10 9AB, United Kingdom
{R.G.Adams, N.Davey}@Herts.ac.uk

Abstract. Subtree retraining applied to a Stochastic Competitive Evolutionary Neural Tree model (SCENT) is introduced. This subtree retraining process is designed to improve the performance of the original model which provides a hierarchical classification of unlabelled data. The effect of subtree retraining on the network produces stable classificatory structures by repeatedly restructuring the weakest branch of the classification tree based on internal relation between members. An experimental comparison using well-known real world data sets, chosen to provide a variety of clustering scenarios, showed the new approach produced more reliable performances.

1 Introduction

The standard Neural Network Competitive Learning algorithm [6] may be modified by the addition of dynamic node creation and the imposition of a tree structure on the classificatory ordering of the nodes. This brings two main advantages: the number of clusters that the neural network will identify does not need to be predefined, and the hierarchical tree structure improves the interpretability of the results. In addition, the use of a tree structure allows a more efficient search for the classifying node so increasing the speed of the model. A basic neural network hierarchical clusterer has been introduced in [1,2]. The latest version of which is called the SCENT model [7,8].

The Stochastic Competitive Evolutionary Neural Tree (SCENT) attempts to combine the advantages of using a tree structure, and of growing the network. The SCENT model shows promise in its ability to produce stable yet flexible hierarchical structures. The nature of the hierarchical structure and the quality of the final classification produced by the networks is, however, very strongly influenced by the values given to externally set parameters. This network is described in Section 2. A subtree retraining process applied to the SCENT model is introduced in Section 3. The comparative results of running the SCENT model with subtree retraining and other neural classifiers over a variety of real world data sets are presented in Section 4. Finally, conclusions of this study are presented in Section 5.
2 Stochastic Competitive Evolutionary Neural Tree

In SCENT, the tree structure is created dynamically in response to structure in the data set. The neural tree starts with a root node with its tolerance (the radius of its classificatory hypersphere) set to the standard deviation of input vectors and its position is set to the mean of input vectors. It has 2 randomly positioned children. Each node has two counters, called inner and outer, which count the number of occasions that a classified input vector is within or outside tolerance, respectively. These counters are used to determine whether the tree should grow children or siblings once it has been determined that growth is to be allowed.

2.1 Main Algorithm

At each input presentation, a recursive search through the tree is made for a winning branch of the tree. Each node on this branch is moved towards the input using the standard competitive neural network update rule. Any winning node is allowed to grow if it satisfies 2 conditions. It should be mature (have existed for an epoch), and the number of times it has won compared to the number of times its parent has won needs to exceed a threshold. A finite limit is put on the number of times a node attempts growth. When a node is allowed to grow, if it represents a dense cluster, then its inner counter will be greater than its outer counter and it creates two children. Otherwise, it produces a sibling node. The process of growth is illustrated in Figure 1.

![Fig. 1. The winning leaf node can grow, (a) children nodes are growth, called downgrowth and (b) sidegrowth in which sibling node is growth](image)

To improve the tree two pruning algorithms, short and long term, are applied to delete the insufficiently useful nodes. The short-term pruning procedure deletes nodes early in their life, if their existence does not improve the classificatory error. The long term pruning procedure removes a leaf when its activity is not greater than a threshold, illustrated in Figure 2.
2.2 Stochasticity

Interest in stochastic method, from the fact that in real world problems, the cost function seldom succeed in defining precisely the real optimality of the solution. The optimality condition is far too complex to be understood by any particular formula, in such cases it is desirable to have methods, which can suggest a set of goods demonstrate such behaviour. Due to the randomise involved, stochastic method the can handle the cost function and can be considered as promising candidates for solving real world problems.

One of the main goals of neural network researches has been to build the system the mimic the ability of humans to solve the real world problem. Curiously, tradition approaches in solving have been based on the two valued logic. This is a form of hard computing which based on precision and certainty. It is in sharp contrast to real human reasoning or what is known as commonsense which is based on approximate, rather than precise computing methods.

Adding stochasticity to a deterministic version (CENT) [1, 2] have some benefit in helping the model avoid local minima in its implicit cost function. There are two different ways in which stochasticity can be added to the model [8]. Firstly the deterministic decisions relating to growth and pruning can be made probabilistic, it is called Decision Based Stochasticity. The Decision Base Stochasticity has 3 procedures which decisions growth allow, grow type (downgrowth or sidegrowth) and short-term pruning. Secondly the attributes inherited by nodes when they are created can be calculated with a stochastic element, it is called Generative Stochasticity by changing tolerance values from deterministic function that new tolerance is same every node are created, to stochastic values is randomness effect to new node more flexibility.

To both of these approaches a simulated annealing process can be added to mediate the amount of non-determinism in a controlled way, so that a decreasing temperature allows for less randomness later in the life of the network. The Decision
will be deterministic which integration of stochastic decision based and simulated annealing. Deterministic version performed well on artificial training set and adding stochasticity still allowed high quality trees to be produced.

3 Subtree Retraining Process and SCENT

All dynamic networks have to overcome the problem of creating too many nodes and over classifying the data set. There can potentially be a node produced to classify each input vector. Sometimes it is possible for the SCENT model to produce a poor subtree. In order to deal with this scenario, we examined all of the level one subtrees produced by the SCENT algorithm. The worst of these subtrees, as identified by the hierarchical correlation measure (described later in Section 3.1) is removed, See Figure 3. The subset of the data previously classified by the aforementioned subtree is now used as a new data set and is represented to the SCENT model. A new tree is therefore generated to classify this subset of the data. Finally, this new tree is reattached to the original tree to replace the previously pruned subtree. This resulting tree is our final classification.

![Fig. 3. The process of subtree replacement](image)

3.1 Cluster Measures

The general goal in many clustering applications is to arrive at clusters of objects that show small within cluster variation relative to the between-cluster variation [4]. Clustering is difficult as many reasonable classifications may exist for a given data set, moreover it is easy for a clusterer to identify too few or too many clusters. Suitable cluster criterion measures are therefore needed [3, 5]. There are two types of
clustering measure: ones that measure the flat clustering performance of the leaf nodes and ones that measure the hierarchical structure.

Hierarchical Correlation (HC): measures the correlation between the dissimilarity matrix $d_{ij}$ and hierarchical separation matrix $h_{ij}$. A measure of the quality of the hierarchical structure in the tree and then be found using the following formula. HC gives a value between 0 and +1, where +1 is optimal.

$$HC = \frac{\sum (d_{ij} - \bar{d}) (h_{ij} - \bar{h})}{\sqrt{\sum (d_{ij} - \bar{d})^2 \sum (h_{ij} - \bar{h})^2}}$$

$d_{ij}$ is a dissimilarity between objects i and j. In this study, the dissimilarity is computed using Euclidean distance.

$h_{ij}$ is the relative height, which is the number of steps in the tree to the closest node that has i and j as descendants.

### 4 Experiments and Results

In order to know how well the SCENT model with subtree retraining performs, comparative results of the SCENT model with subtree retraining, the SCENT and CENT (deterministic model) models are presented. The three models were applied to four real world data sets: IRIS, Egyptian, FGlass, and Wine data set [8] three times; the resulting trees were evaluated using the aforementioned cluster measure.

Table 1 gives the average and standard deviation of the hierarchical correlation. In Table 1, the overall performance of the SCENT model with subtree retraining was clearly better than the other networks in terms of the hierarchical correlation measure. Figure 4 shows an example where the new SCENT model produced a better hierarchical clustering than other two versions of the SCENT model.

| Model                  | Iris  | Egyptian | Fglass | Wine  | Average | Standard Deviation |
|------------------------|-------|----------|--------|-------|---------|-------------------|
| CENT                   | 0.825 | 0.805    | 0.608  | 0.403 | 0.660   | 0.197             |
| SCENT                  | 0.798 | 0.846    | 0.714  | 0.442 | 0.700   | 0.180             |
| SCENT-Subtree retraining | 0.833 | 0.905    | 0.825  | 0.404 | 0.742   | 0.228             |

Table 1. Average hierarchical correlation of the classifications produced by the 3 neural network models tested over 4 real world data sets. A value close to 1 representing the best performance.
5 Conclusion

This paper presents a method for subtree retraining applied to a dynamic hierarchical neural clusterer. The key idea that we introduce here is that the worst subtree should be identified, removed from the main tree, and a new subtree is created in its place by using the SCENT algorithm on the subset of data originally classified by the poor subtree.

The performance of the SCENT model with subtree retraining over a range of data sets has been presented and has been shown to provide a good interpretation of the real world data set, and to produce better classifications compared to the SCENT and the CENT models over the four real world data sets examined here. Finally, it
should be noted that the subtree retraining method may also be a good approach to apply to any neural tree models not just the SCENT model.

References

1. R.G. Adams, K. Butchart, and N. Davey, “Classification with a Competitive Evolutionary Neural Tree,” Neural Networks Journal, Vol. 12, (1999) 541-551
2. N. Davey, R.G. Adams, and S. G. George, “The Architecture and Performance of a Stochastic Competitive Evolutionary Neural Tree,” Journal of Applied Intelligence, Vol. 12, No. 1/2, (2000) 75-93
3. A. D. Gordon, Classification, Chapman & Hall, London, (1999)
4. J. A. Hartigan, Clustering Algorithms, John Wiley & Sons, USA, (1975)
5. G. W. Milligan and M. C. Cooper, “An Examination of Procedures for Determining the Number of Clusters in a Data Set,” Psychometrika Journal, Vol. 50, No. 2, (1985) 159-179
6. J. Hertz, A. Krogh, and R. G. Palmer, An Introduction to the Theory of Neural Computation, Addison Wesley, USA, (1991)
7. W. Pensuwon, R. G. Adams, and N. Davey, “Comparative Performances of Stochastic Competitive Evolutionary Neural Tree (SCENT) with Neural Classifiers,” Proceedings of the 8th International Conference on Neural Information Processing - ICONIP-2001., (2001) 121-126
8. W. Pensuwon, R. G. Adams, and N. Davey, “The Analysis of the Addition of Stochasticity to a Neural Tree Classifier,” Journal of Applied Soft Computing, Vol. 1, No.3, (2001) 189-200
Cluster Analysis of Gene Expression Profiles Using Automatically Extracted Seeds

Miyoung Shin and Seon-Hee Park

Bioinformatics Team, Future Technology Research Division, ETRI,
161 Gajeong-dong, Yuseong-gu, Daejeon, Korea 305-350
{shinmy, shp}@etri.re.kr

Abstract. This paper addresses the problem of clustering gene expression profiles based on automatically extracted seeds which are obtained by our proposed method. Specifically, we introduce a new clustering methodology that consists of three stages: seed extraction, cluster generation, and its evaluation. Performance analysis of the proposed methodology is done with a real dataset, and its results are reported in detail. Overall, based on our empirical studies, the proposed clustering methodology seems to be very favorable for gene expression data analysis, as alternatives to current clustering methods.

1 Introduction

In recent years, the problem of analyzing large volume of gene expression profiles, produced by DNA microarray technology, has been greatly emphasized to help biologists find new biological knowledge of significance in an efficient way from drastically accumulated gene expression data. Cluster analysis of gene expression profiles is to find biologically relevant groups of genes (or samples) by investigating their expression levels. Generally there is an agreement that co-expressed genes are functionally related to each other. Thus, various clustering methods have been investigated to identify groups of co-expressed genes based on gene expression profiles.

Current clustering approaches [1,2,3,4] have their own strengths and limitations. For example, hierarchical clustering method generates a tree-like structure, called dendrogram. The dendrogram helps users to intuitively define appropriate clusters by visual inspection, but due to the need of visual inspection, this method does not scale well for large datasets. K-means clustering method is relatively simple to use and works well for most cases, but its clustering results tend to be very sensitive to the choice of initial values. Self-organizing maps are well suited for multi-dimensional data analysis and provide easy visualization, but they are somewhat complicated to use. Principal component analysis is, generally, very helpful for exploring multi-dimensional data visually by projecting them onto two or three dimensional space, but it does not improve the quality of clustering results.

In this paper we introduce a new clustering methodology for gene expression data analysis, which consists of a three-stage procedure: seed extraction, cluster generation and cluster evaluation. The details of our proposed clustering methodology are
presented in Section 2. For performance analysis, our experiments were made with a real dataset, and its results are reported in detail in Section 3. Finally, some concluding remarks are presented in Section 4.

2 Proposed Clustering Methodology

Our proposed clustering methodology is basically a three-step procedure: seed extraction, cluster generation, and cluster evaluation. The first step of seed extraction is, for a user specified value of $k$, to automatically find $k$ good seeds of genes by computational analysis of gene expression profiles. After seed extraction, the second step of cluster generation proceeds to generate $k$ clusters with the chosen seeds, i.e., by using either seed-nearest method or seed-kmeans method. The resulting clusters are then evaluated in the third phase. This three step process can be iterated for different choices of $k$.

2.1 Seed Extraction

For a user specified value of $k$, the process of seed extraction is to identify $k$ seeds of genes (or samples) by computational analysis of gene expression profiles, in such a way that each seed would potentially represent a prototype expression profile for one of the $k$ clusters. The goodness of the seeds here is defined by how well distributed they are in capturing the entire gene expression space while not being very similar to one another, i.e., far away enough to avoid information redundancy. To implement this idea, we employ the mathematical machinery of singular value decomposition based subset selection procedure [5] for transformed gene expression profiles. A detailed description of seed extraction is given below.

Let $G$ be the gene expression profiles for $n$ genes, $G = \{g_i\}_{i=1}^n$, where an expression profile of $g_i = (g_{i1}, g_{i2}, \cdots, g_{id})$ consists of its expression levels over $d$ conditions or samples. For seed extraction, the gene expression profiles $G$ need to be first transformed into $G' = \{g'_i\}_{i=1}^n$ where the transformation of $g_i$ is defined as $g'_i = (\varphi_1(g_i), \varphi_2(g_i), \cdots, \varphi_n(g_i))$. The function $\varphi_j(\cdot)$ is Gaussian function having $g_j$ and $\sigma$ as its center and width, respectively. That is, for a given gene $g_i$, $\varphi_j(g_i) = \exp(-\| g_i - g_j \|^2 / 2\sigma^2)$. Thus, it should be noted that the width $\sigma$ of Gaussian is a control parameter which defines an exact functional form for transformation and so needs to be specified by users.

Next, from the transformed expression profiles of $G'$, $k$ sufficiently independent column vectors should be chosen by applying QR factorization with column pivoting [5] of the right singular matrix of $G'$. Once we have the $k$ chosen columns $\varphi_{s_1}(\cdot), \varphi_{s_2}(\cdot), \cdots, \varphi_{s_k}(\cdot)$ of sufficient independence, the genes of $s_1, s_2, \ldots, s_k$ are then extracted by identifying the centered genes of Gaussian functions $\varphi_{s_1}(\cdot), \varphi_{s_2}(\cdot), \cdots, \varphi_{s_k}(\cdot)$, and these become our choice of seeds.
2.2 Cluster Generation

With the $k$ seeds chosen in the previous stage, cluster generation can be performed in two different ways: seed-nearest method and seed-kmeans method.

Cluster generation by the seed-nearest method is to consider the chosen seeds $s_1, s_2, \ldots, s_k$ each as a prototype expression vector for one of the $k$ target clusters $C_1, C_2, \ldots, C_k$, and thus to assign a gene $g_i$ to the cluster $C_{\tilde{k}}$ of which prototype expression vector $s_{\tilde{k}}$ is the nearest to the gene $g_i$. That is, for each gene of $g_i$, $i = 1, \ldots, n$, its cluster membership $C_{\tilde{k}}$ is defined by $\tilde{k} = \arg \min_j (\| g_i - s_j \|)$.

Cluster generation by the seed-kmeans method is to take the $k$ chosen seeds as initial centers for $k$ clusters to be generated. Based on these initial centers, each gene is assigned to the cluster of which center is the closest to the gene. After initial assignments are made for all the genes, the cluster centers are then refined by taking the averaged gene expression vectors of the respective clusters. This refinement process is iterated until the $k$ cluster centers are stable, eventually optimizing the centers in terms of minimal within-cluster dispersion. That is, if there is no refinement in the $k$ cluster centers, current cluster memberships become our final result of $k$ clusters.

2.3 Cluster Evaluation

Our clustering results are assessed by adjusted rand index, which is a statistical measure to assess the agreement between two different partitions and has been used in some previous research on gene expression data analysis [6]. The adjusted rand index is defined as in Formula (1), whose a value closer to 1 implies that the two partitions are closer to perfect agreement.

Suppose that $U = \{u_1, \ldots, u_n\}$ is the true partition and $V = \{v_1, \ldots, v_c\}$ is a clustering result. Then, according to [6], the adjusted rand index is defined as follows:

$$\frac{\sum_{u_i} \sum_{j} \left(\begin{array}{c} n_j \\ 2 \end{array}\right) - \left(\begin{array}{c} \sum_{j} \left(\begin{array}{c} n_j \\ 2 \end{array}\right) \right) \left(\begin{array}{c} \sum_{j} \left(\begin{array}{c} n_j \\ 2 \end{array}\right) \right) \left(\begin{array}{c} n \\ 2 \end{array}\right) \right)}{\frac{1}{2} \left(\begin{array}{c} \sum_{j} \left(\begin{array}{c} n_j \\ 2 \end{array}\right) \right) + \left(\begin{array}{c} \sum_{i} \left(\begin{array}{c} n_i \\ 2 \end{array}\right) \right) - \sum_{j} \left(\begin{array}{c} \sum_{i} \left(\begin{array}{c} n_i \\ 2 \end{array}\right) \right) \left(\begin{array}{c} n_j \\ 2 \end{array}\right) \right) \left(\begin{array}{c} n \\ 2 \end{array}\right) \right)}$$

where $n$ is the total number of genes in the dataset, $n_y$ is the number of genes that are in both class $u_i$ and cluster $v_j$, and $n_i$ and $n_j$ are the number of genes in class $u_i$ and cluster $v_j$, respectively.

3 Experiments

Our experiments were performed with gene expression profiles regarding yeast cell cycles, for which the target clusters are already known. For the dataset, we first determine $k$ seeds by our systematic method (see Sect. 2.1), employing a value of the control parameter $\sigma$ approximately in the range of $0 < \sigma < \sqrt{d}/2$, where $d$ is the number of conditions. The range of $\sigma$ was heuristically determined based on our previous studies. After $k$ seeds are identified, seed nearest method and seed $k$-means method
are applied for cluster generation, respectively. The clustering results are then evaluated with the adjusted rand index, as a validation measure to quantify the closeness to the true solution. By selecting an optimal value of the control parameter $\sigma$ in terms of the adjusted rand index, the clustering solutions are obtained for a given $k$.

For different values of $k$, the above procedure can be iterated, and an optimal number of clusters can be identified by sensitivity analysis of clustering solutions to $k$. Euclidean distance was used as a distance metric in our analyses. For comparative analysis, we also generated $k$ clusters using four other methods; centroid-linkage, complete-linkage, single-linkage, and k-means clustering with random initial centers. For k-means clustering, to minimize the random effect, the averaged performance over 10 repetitions is reported here.

### 3.1 Dataset

The dataset that was used for our experiments here is the gene expression data regarding yeast cell cycles [7], in which the expression levels of more than 6000 genes over two cell cycles, i.e., at 17 time points, are included. Out of approximately 6000 genes, Cho et al. [7] identified 420 genes which show the peak at different time points and categorized them into five phases of cell cycle. However, among those, some genes show the peak at more than one phase of cell cycle. Thus, by removing such genes out of 420 genes, 384 genes were identified [8] and used in our experiments, whose expression levels clearly show the peak at one of the five phases of yeast cell cycle.

### 3.2 Analysis Results on Yeast Cell Cycle Data

For the dataset shown in Fig. 1 (a), we first generated $k=5$ clusters with the control parameter $\sigma$ in the range of $\sigma = (0.25:0.25:2.0)$, since we know that the true number of clusters for this dataset is five. Recall that the value of $\sigma$ is heuristically determined in the range of $0 < \sigma < \sqrt{d}/2$, where $d$ is the number of conditions. Since the number of conditions in the synthetic data is 17, the range of $\sigma$ is chosen as $0 < \sigma < \sqrt{17}/2$, that is 2.0616.

With $k=5$, our procedure of seed extraction produced five seeds as shown in Fig. 1 (b), which generated the clustering results of (c) and (d) in Fig. 1 by seed-kmeans method and seed-nearest method, respectively. Depending on the value of $\sigma$, the chosen seeds were slightly different, which also led to some changes in final clustering results. The sensitivity of clustering results to $\sigma$ is presented in Fig. 2. Here, as can be seen, seed-nearest method is a little more sensitive to $\sigma$ than seed-kmeans method. For comparative analysis, we used $\sigma = 2$ for our clustering solutions, and compared them with the results from other clustering methods. Further, the sensitivity of clustering results to $k$ was also studied for $k=3$ to 10.

The results from our seed-based methods and other four methods are shown graphically in Fig. 3. Note that the higher quantity of adjusted rand index implies better agreement between clustering results and the true clusters. Surprisingly, both of our seed-based methods clearly showed the peak point of the adjusted rand index at $k=5$, which is the true number of clusters, while most other methods except complete-linkage do not capture it. Further, it is clearly seen that our seed-kmeans method,
Fig. 1. (a) Gene expression profiles in yeast cell cycle data which were used for our experiments. (b) Automatically extracted 5 seeds by our proposed algorithm. (c) Clustering results obtained by seed-kmeans method with our 5 chosen seeds. ($\sigma = 2.0$) The value of adjusted rand index is 0.5138. (d) Clustering result obtained by seed-nearest method with our 5 chosen seeds. The value of adjusted rand index is 0.5078. ($\sigma = 2.0$)

Fig. 2. Sensitivity analysis of our two seed-based methods to the choice of control parameter $\sigma$ in the yeast cell cycle data ($k=5$)
Fig. 3. Sensitivity analysis of our two seed-based methods and other four methods to the number of clusters k in the yeast cell cycle data ($\sigma$=2.0) amongst all the algorithms and all the k values, produces the highest value of adjusted rand index for k=5, the true number of clusters.

4 Conclusions

In this paper we introduced a new clustering methodology consisting of three stages: seed extraction, cluster generation and its evaluation. Specifically, we proposed two seed-based clustering methods, seed-nearest method and seed-kmeans method, for gene expression data analysis. For the resulting clusters, their quality is assessed with external biological knowledge if available as a validation measure. Compared with other approaches, our seed-based clustering methods have a distinctive feature that, for cluster generation, they employ the seeds automatically chosen by our systematic procedure presented earlier. The method of seed extraction has strong mathematical foundations, so that $k$ seeds for a given dataset are uniquely and systematically identified.

In addition, through experiments with a real dataset, we showed the applicability and usefulness of the seed-based clustering methods in identifying biologically relevant groups of genes based on microarray gene expression data. Since these methods are based on algebraic principles, they are systematic, reproducible, and yield consistent results. Further, our empirical studies show that our seed-based clustering methods are very favorable for gene expression data analysis, as alternatives to current
clustering methods. We believe that, due to their very attractive features, the seed-based clustering methods should be very appealing to practitioners in terms of their ease of use and their performance.

References

1. Eisen M. B, Spellman P. T, Brown P. O and Botstein D: Cluster Analysis and Display of Genome-Wide Expression Patterns, Proc. Natl. Acad. Sci. (1998), Vol. 95, 14863-14868.
2. Tavazoie S, Hughes J. D, Campbell M. J, Cho R. J and Church G. M: Systematic Determination of Genetic Network Architecture, Nature Genetics (1999), Vol. 22, 281-285.
3. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E. S and Golub T. R: Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation, Proc. Natl. Acad. Sci. (1999), Vol. 96, 2907-2912.
4. Yeung K. Y and Ruzzo W. L: Principle Component Analysis for Clustering Gene Expression Data, Bioinformatics (2001), Vol. 17(9), 763-774.
5. Golub, G.H. and Van Loan, C.F.: Matrix Computation (3rd edition), The Johns Hopkins University Press (1996).
6. Yeung K. Y, Haynor D. R and Ruzzo W. L: Validating Clustering for Gene Expression Data, Bioinformatics (2001), Vol. 17(4), 309-318.
7. Cho R. J, Campbell M. J, Winzeler E. A, Steinmetz L, Conway A, Wodicka L, Wolfsberg T. G, Gabrielian A. E, Landsman D, Lockhart D. J, and Davis R. W, A genome-wide transcriptional analysis of the mitotic cell cycle, Molecular Cell, 2:65-73, 1998.
8. The dataset is available at http://staff.washington.edu/kayee/cluster/.
Prediction of Plasma Membrane Spanning Region and Topology Using Hidden Markov Model and Neural Network

Min Kyung Kim¹, Hyun Seok Park², and Seon Hee Park ³

¹ Engineering Research Center, Ewha University,
11-1 Daehyun-dong,
Seodaemun-gu, Seoul, 120-750, Korea
² Department of Computer Science and Engineering,
Seodaemun-gu, Seoul, 120-750, Korea
{minkykim, neo}@ewha.ac.kr
³ Bioinformatics Team, Electronics and Telecommunication Research Institute,
161 Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea
shp@etri.re.kr

Abstract. Unlike bacteria, which generally consist of a single intracellular compartment surrounded by a plasma membrane, a eukaryotic cell is elaborately subdivided into functionally distinct, membrane-enclosed intracellular compartments that are composed of the nucleus, mitochondria, and chloroplast. Although transmembrane spanning region and topology prediction tools are available, such software cannot distinguish plasma membrane from intracellular membrane. Moreover, the presence of signal peptide, which has information of intracellular targeting, complicates the transmembrane topology prediction because the hydrophobic composite of signal peptide is considered to be a putative transmembrane region. By immediately detecting a signal peptide and transmembrane topology in a query sequence, we can discriminate plasma membrane spanning proteins from intracellular membrane spanning proteins. Moreover, the prediction performance significantly increases when signal peptide is contained in queries. Transmembrane region prediction algorithm based on the Hidden Markov Model and ER signal peptide detection architecture for neural networks has been used for the actual implementation of the software.

1 Introduction

The genome-sequencing project has generated and will continue to generate enormous amounts of sequence data. With sequencing data accumulating so rapidly, many researchers are entering the next phase of genome projects and are functionally analyzing raw genomic sequences. They are targeting plasma membrane proteins as areas of primary interest since many of them are surface receptors or enzymes. These are involved in intracellular signaling pathways which are triggered by extracellular hormones and ligands, and are also more easily accessible for drugging than intracellular
proteins. Regardless of the importance of membrane proteins, there are few known 3D structures since membrane proteins are not easily crystallized and are hardly tractable by NMR microscopy. The membrane region of the protein is surrounded by lipid bilayer, thus encoding most transmembrane alpha helices by an unusually long stretch of hydrophobic residues. These residues simplify the prediction process of the transmembrane region. Therefore, the state-of-the-art in this field performs relatively higher than 3D structure prediction, like globular alpha helix detection, for instance. Modeling biological systems is not a trivial matter when only a few experimental data have been provided. Still, the relationship between the input and output can be predicted by machine learning approaches to a certain extent.

Plasma membranes separate the cell from the outside world. Proteins are synthesized in the cytosol, and are transported to targeting organelles like chloroplast (in plants), mitochondria, peroxisome and ER by their signal peptides. ER transported proteins are reached at the plasma membrane through the vesicle transport. Cells and intracellular organelles are surrounded by a lipid bilayer, the membrane architecture.

Intracellular organelles such as mitochondria, chloroplast and nucleus are also surrounded by lipid bilayer. For this reason, finding the transmembrane region is not enough when looking for plasma membrane protein, which initiate signaling cascade and are candidates for pharmaceutical targets. Proteins are synthesized on ribosomes in the cytosol, and their subsequent fate depends on their N-terminal amino acid sequence. These can contain signal peptide that directs their transport from the cytosol into the nucleus, the ER, mitochondria, chloroplast, etc. Plasma membrane proteins and secreted proteins are delivered to the cell surface through the vesicle transport from the ER.

Proteins embedded in the bilayer are called transmembrane protein, which are transverse hydrophobic region. Therefore, transmembrane region finding programs cannot discriminate receptors or enzymes that are located in plasma membrane since intracellular membranes have similar membrane structures.

Protein targeting prediction tools are available on the web. They are implemented by diverse approaches, for example, based on N-terminal signal sequences and support vector machines using full sequence. However, the targeting information does not combine with transmembrane prediction tools.

A variety of tools for transmembrane prediction have been implemented by diverse approaches such as hydrophathy value, biological rule and machine learning approaches. Among these, TMHMM [1] and HMMTOP [2] based on the Hidden Markov Model have been reported to have the best performance [3]. These models predict the location of membrane-spanning region and topology.

Plasma transmembrane proteins have a long stretch of hydrophobic transmembrane region and ER signal peptide in their N-terminal sequences. A major advancement in our system – called PASS (Prediction of Alpha-Helix Transmembrane by Separating ER Signal Sequence) in predicting plasma membrane spanning region and topology prediction – integrates two different kinds of modules: the CSP module (signal peptide and cleavage site prediction) and the TM module (transmembrane spanning region and topology prediction).
2 Related Works

2.1 Transmembrane Spanning Region and Topology Prediction

The use of hydrophobicity and positively charged amino acids compositions are used in early models in predicting membrane-spanning regions. Tools in Table 1 are based on neural network, matrix, or multiple alignments. According to Table 1, TMHMM, based on a hidden markov model approach, is currently the best performing transmembrane prediction program.

Table 1. This table presents currently developed prediction tools, methods and their performance. The performance represented prediction accuracy (%), which defines the correctness of all membrane spanning regions found (before slash) and transmembrane topology (after slash) “-” means they are not defined because they don’t predict the topology

| TM Prediction Tools | Methods                        | Performance (%) |
|---------------------|--------------------------------|-----------------|
| Tmpred              | Dense alignment surface method  | 37/17           |
| DAS                 | Special scoring matrix         | 37/-            |
| TMAP                | Multiple sequence alignment     | 43/26           |
| MEMSAT1.5           | Statistical tables             | 53/77           |
| SOSUI               | Scoring matrix                 | 36/-            |
| PRED-TMR2           | Hydrophobicity calculations    |                 |
| TMHMM 2.0           | Hidden Markov Model            | 68/70           |
| HMMTOP              | Hidden Markov Model            | 44/82           |

2.2 Subcellular Location Prediction

Most subcellular location prediction methods are grouped into two approaches: one is based on the detection of the N-terminal signal peptide, and the other is based on full length amino acid composition. SigalP [4] and TargetP [5] are prediction servers based on N-terminal sequence approaches. These programs also supply the information of mature protein by cleavage of signal peptides. However, N-terminal signal peptides are limited in some intracellular organelles, therefore preventing predictions from taking on a general approach in covering all kinds of parts in a cell.

Machine learning methods such as Neural Network [6] and Support Vector Machines [7] were also used to predict subcellular location of proteins. They are motivated by the variation of amino acid compositions of full-length protein according to their location in the cell [8]. They show improved performance and more subdivided categorization than N-terminal sequences based approach.

2.3 Problems with Previous Systems

Although researchers’ primary interest lies in the plasma membrane proteins, currently available software do not discriminate plasma membrane spanning from intracellular membranes. Moreover, the existence of the N-terminal peptide sequence complicates the prediction of the transmembrane regions. N-terminal signal peptide
can be regarded as a transmembrane region due to the hydrophobic nature of signal peptide [9]. Once resigned as a transmembrane region, it affects other topology predictions by frame shift.

We have constructed prediction system called **PASS** that is based on the Hidden Markov Model for transmembrane region finding and neural network trained modules for ER signal peptide and cleavage site detection. For the recognition of cleavage site, we used the approach based on N-terminal signal peptide detection. Preprocessing of ER signal peptide can improve the prediction rate by blocking the frame shift occurrence, caused by misinterpreting signal peptide as a transmembrane region. ER-targeted proteins will be transported to the plasma membrane through the vesicle transport by membrane fusion. For this reason, it makes it possible to discriminate plasma membrane proteins from other intracellular transmembrane spanning proteins.

### 3 Implementation and Evaluation

**PASS** has two different modules as shown in Figure 1: **CSP** (Cleavage Site Prediction) and **TM** (TransMembrane region and topology prediction). If ER signal peptide and cleavage site exist in the **PASS** query sequences, the signal peptide is removed before the entrance of the **TM** module. The output of **PASS** will be sequences labeled as i (inner), M (membrane spanning region), and o (outer region of the membrane).

![Fig. 1. System architecture of PASS (Prediction of Alpha-Helix Transmembrane by Separating ER Signal Sequence)](image)

#### 3.1 Cleavage Site Prediction (CSP)

For the recognition of ER signal peptide and cleavage site, we use neural network architectures (especially back propagation) through sliding windows. Input protein sequences are inverted to 21 bit values, for example, M (amino acids for methionine) represented as “000000000010000000000”. From the fixed size window of 21 bits, **CSP** calculates the C and S-scores. Then S-score and C-score designed for cleavage site is used to detect the signal peptide. Training data is available from the SignalP
training set [4]. For input data model for C-scores, the value is labeled “0” except the following cleavage, which is labeled “1” in protein sequences. In case of S-score data model, the signal peptide region is marked as “1”. Weight values are adjusted by target values according to their training. After the training of neural nets, CSP calculated the Y-score derived from the C and S-score. CSP concludes the result by considering the values of S, C and Y-scores.

![Diagram of neural network](image.png)

**Fig. 2.** Neural Network models used in CSP modules (Right). It consists of three different layers. Unit number and window sizes are selected after the test of performance. Cleavage site is decided by a combined use of C and S-score (Left)

| Methods | Cleavage Site Location (% Correct) | Signal Peptide Discrimination (Correlation) |
|---------|----------------------------------|---------------------------------------------|
| SignalP | 68.0 (67.9)                      | 0.96 (0.97)                                 |
| CSP     | 81.0 (81.1)                      | 0.77 (0.77)                                 |

### 3.2 Transmembrane Spanning Region and Topology Prediction (TM)

TM architecture is similar to TMHMM, another kinds of transmembrane prediction tools based on hidden markov model. TM discriminates transmembrane and non-transmembrane (inner/outer) region of proteins by labeled sequences such as “i”, “M” and “o”. Basic idea is that the state set of HMM for transmembrane protein is designed as their biological meaning. And each state has the probability value of the appearance of 20 amino acids, which is trained by Baum-Welch methods. Training data derived from topology defined positive data that was used in TMHMM. Prediction is carried out by Viterbi calculations that approximate the total probability by the probability of the most likely path among each states.

Although, TMHMM shows the best performance in this field, it disregards the information of signal peptide. Training data that were used in TMHMM include sequences containing signal peptide. Therefore, we test the effect of signal peptide in training data. The exclusion of signal peptide, not containing signal peptide (Method
2) improves the prediction performance in TM modules in aspect of topology prediction and TM region prediction than data sets containing signal peptide (Method 1). It means that signal peptide affects the performance of membrane spanning region predictions significantly (Table 3). Therefore, signal peptide should be considered not only in designing the prediction module but also in selecting training data.

Table 3. Signal peptide contained sequences in training set affect the prediction result

| Methods | Training Set | Topology Prediction | TM region Prediction | Single TM Sensitivity | Single TM Specificity |
|---------|--------------|---------------------|----------------------|-----------------------|----------------------|
| 1       | 160          | 65.0                | 74.4                 | 96.4                  | 96.2                 |
| 2       | 150          | 68.1                | 76.3                 | 96.4                  | 96.7                 |

3.3 Prediction of Alpha-Helix Transmembrane by Separating ER Signal Sequence (PASS)

PASS is the integrated system of CSP and TM for plasma membrane spanning region and topology prediction. For the discrimination of plasma membrane from intracellular membranes, we preprocessed the N-terminal signal peptide to improve the performance especially when signal peptide is contained in query sequences.

To test the performance of PASS in this case, we selected previously reported 24 protein sequences with experimentally determined transmembrane topology [9]. Test data were selected from the Moller’s TM data set [10], SWISS-PROT [11] and TMPDB [12]. They were filtered by multiple alignment tools (clustalW) down to 30% homology.

Accuracy of each method is shown in Table 4. The accuracy of TM topology was predicted as same or below than the accuracy of segments and position. Therefore, the correct decision of TM segments is a base of TM topology prediction. Confusing signal peptide as a transmembrane region affects not only the number of TM segments but also TM topologies. PASS has an advantage for the detection of transmembrane segment. For this reason, PASS performs better than any other prediction tool in case of signal peptide contained queries.

4 Conclusion and Future Works

PASS integrated the location prediction module and transmembrane region and topology prediction module (http://dblab.sejong.ac.kr:8080/pass/home.jsp). The main virtues are that PASS discriminates the integral proteins into the plasma membrane from intracellular membranes and eliminates the possibility of misrecognition of signal peptide as a transmembrane region.

Many subcellular location prediction servers are now available. However, for fine resolution such as mitochondrial inner space or mitochondrial membrane, some comprehensive remedial steps are required, e.g., combined use of transmembrane...
prediction with location prediction. Moreover, for the improvement of transmembrane prediction accuracy, a consensus prediction method based on combining the results from five currently used prediction servers were suggested [13].

However, SVM based approaches for location prediction tools show the best performances in subcellular location prediction fields. Still, it has a disadvantage in the recognition of cleavage sites. For this reason, CSP modules in PASS were designed by N-terminal sequences detection instead of amino acid composition of full-length protein by neural network. Combined use of SVM and Neural Net in CSP module should be considered for the coverage of intracellular compartments.

Transmembrane region of protein has two kinds of secondary structures: alpha-helix and beta-barrel. TM modules in PASS predict only the alpha-helical transmembrane region. Most programs in this category are focused on alpha-helical transmembrane proteins, which are majority. Recently, beta-barrel transmembrane region and prediction have been attempted by HMM models [14]. Therefore, we would want to cover all kinds of membrane and secondary structures in the near future.

References

1. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. J. Mol. Biol. Vol. 305 (2001) 567-580
2. Tusnady, G.E., Simon, I.: The HMMTOP transmembrane topology prediction server. Bioinformatics. Vol. 17 (2001) 849-850
3. Moller, S., Croning, M.D., Apweiler, R.: Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics. Vol. 17 (2001) 646-653
4. Nielsen, H., Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. Vol. 10 (1997) 1-6
5. Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G.: Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. Vol. 300 (2000) 1005-1016
6. Reinhardt A, Hubbard T.: Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. Vol. 26 (1998) 2230-2236
7. Park, K.J., Kanehisa, M.: Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics. Vol. 19(2003) 1656-1663
8. Nakashima H, Nishikawa K.: Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol. Vol. 238 (1994) 54-61
9. Lao DM, Arai M, Ikeda M, Shimizu T.: The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics. Vol. 18 (2002) 1562-1566
10. Moller, S., Kriventseva, E.V., Apweiler, R. A collection of well characterised integral membrane proteins. Bioinformatics. Vol. 16 (2000) 1159-1160
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. Vol. 31 (2003) 365-370.
12. Ikeda, M., Arai, M., Okuno, T., Shimizu, T., TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Res. Vol. 31(2003) 406-409
13. Nilsson, J., Persson, B., von Heijne, G.: Consensus predictions of membrane protein topology. FEBS Lett. Vol. 486 (2000) 267-269
14. Martelli, P.L., Fariselli, P., Krogh, A., Casadio, R.: A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics. Suppl 1 (2002) S46-53.
Speed Control and Torque Ripple Minimization in Switch Reluctance Motors Using Context Based Brain Emotional Learning

Mehran Rashidi, Farzan Rashidi, Mohammad Hossein Aghdaei, and Hamid Monavar

Hormozgan Regional Electric Co., Bandar-Abbas, Iran
PO. BOX: 791451879, Postal Code: 79167-95599, Bandar-Abbas, Iran
mrashidi@mehr.sharif.edu, f.rashidi@ece.ut.ac.ir, hamid_monavar@yahoo.com

Abstract. Switched Reluctance (SR) drive technology is a serious contender for replacing the existing technologies, because of its technical and economic advantages. If some remaining problems like excessive torque ripple could be resolved through intelligent control, it would enjoy enormous comparative advantages for grabbing significant market share. To torque ripple minimization and also speed control of SR motors, in this paper, we apply a modified version of context based brain emotional learning (CBBEL) to Switched Reluctance Motor. Our proposed solution, which is biologically motivated, can achieve very robust and satisfactory performance. The results show superior control characteristics especially very fast response, simple implementation and robustness with respect to disturbances and manufacturing imperfections. The proposed method is very flexible and can easily be enforced via defining proper emotional cues.

1 Introduction

Electric motors play an important role in consumer and manufacturing industries. Among all different kinds of electric motors, Switched Reluctance Motors (SRMs) have a special place. That is because of their simple construction, high speeds and their very low cost comparing with the other Motors [1,2]. In these motors, the rotor does not have any permanent magnet or windings, in contrast with other motors [3]. This reduces the rotor weight and consequently, with a precise commutation, makes it possible to reach much higher speeds. Variable reluctance in the air gap between rotor and stator poles provides torque in SRMs. In other words, the motion system tends to retain its energy in the minimum state, so the rotor poles, rotate in such a way to face the stator poles [4]. However, the different angles between the rotor and stator poles cause some of them to be out of phase and with a proper commutation a smooth motion can be achieved. Another advantage of SRMs is their reliability, which is the result of a high independency in the physical and electrical phases. On the other hand, in SRMs, not only do we need a commutator, but also a complex control system is required, Because of the nonlinearity of the magnetic characteristics. Robustness requirements become more demanding if designing production line for adjustable
speed drives rather than designing a drive for a given electromotor is intended. Here
unmodelled dynamics due to manufacturing variations should be considered in
addition to aging and similar effects. Moreover, the need to have a shaft position
sensor (or utilize sensor less control systems with very high robustness/adaptability),
high torque ripples, vibrations and acoustic noise are some of the other problems that
should be dealt with in SRMs. This paper focuses on solving these complex control
problems via an innovative approach: the use of context based brain emotional
learning technique [5]. The structure of this paper is as follows:

Section 2 describes the SRM and its mathematical model. In section 3, the whole
structure of the proposed CBBEL for speed control and torque ripple minimization of
SRM is formulated. Section 4 describes the structure of proposed controller and its
application in speed control and torque ripple minimization of SR motor. The
performance of the proposed method under the whole range of operation conditions of
SRM is then shown in Section 5. Finally some conclusion and remarks is discussed in
section 6.

2 Mathematical Model of SR Motors

The most important properties of the SRMs are their nonlinear angular positioning
parameters and nonlinear magnetic characteristics. The first group contains: winding
inductance, produced torque and Back EMF, which depend on the rotor angle. On the
other hand magnetic saturation causes the nonlinear magnetic characteristics. The
main principle for SRMs’ modeling is based on the Magnetic-position curve, which
shows the linking flux versus current in different rotor angles. The appendix contains
the proper model parameters for a three phase SRM with six poles on the stator and
four poles on the rotor, which is used in this paper. The complete mathematical model
for the SRM including the magnetic and electrical equations would be achieved with a
whole considering of the magnetic saturation. The mathematical model of the SRM is
as follows [6]:

\[
\omega = \frac{d\theta}{dt}
\]

\[
\frac{d\omega}{dt} = \frac{1}{J} (T(i, \theta) - T_L - B \omega)
\]

In which, \( \omega \) stands for angular velocity, \( T_L \) for load torque, \( B \) for friction
coefficient, and \( J \) for the moment of inertia. Here we adopt the above equation for
speed control and torque ripple minimization in SR motors.

3 Context Based Brain Emotional Learning

Motivated by the success in functional modeling of emotions in control engineering
applications [7-9], the main purpose of this research is to use a structural model based
on the limbic system of mammalian brain, for decision making and control
engineering applications. We have adopted a network model developed by Moren and
Balkenius [10, 11], as a computational model that mimics amygdala, orbitofrontal
cortex (OFC), thalamus, hippocampus, sensory input cortex and generally, those parts
of the brain thought responsible for processing emotions. There are two approaches to intelligent and cognitive control. In the indirect approach, the intelligent system is utilized for tuning the parameters of the controller. We have adopted the second, so called direct approach, where the intelligent system, in our case the computational model termed CBBEL, is used as the controller block. There are several inputs to the amygdala. Unlocalized stimuli are coming straight from the sensory areas. Localized, unexpected stimuli bound to a given place enter from the hippocampus. These stimuli are treated the same way in the amygdala. There is also a scalar reinforcer, used for the actual conditioning. Outputs from the amygdala are the conditioned signals to the OFC, and the emotional conditioning. The OFC uses the conditioned signals to determine whether to inhibit the amygdaloid output, and sends an inhibitory signal to the amygdala. The stimulus input to the amygdala consists of the stimuli CS and the bind node outputs from the hippocampus. These are concatenated into one stimulus vector $S$. The bind node outputs are thus treated as just another kind of stimuli. Due to the nature of this model, the actual input is not the bind signals themselves, as the model is not designed to handle continuous signals. Instead it is the change in bind input strength – in practice, it signals bind node activation with a one-step stimulus spike. An additional input signal is the scalar $R$ representing the reinforcer. The output $E$ of the amygdala is calculated by:

$$E = \left[ \sum_{i=0}^{n} A_i - E_0 \right]^+$$

$$A_i = S_i V_i$$

Where $V_i$ is the connection weight vector, and. $E_0$ is the inhibitory signal from the OFC (Orbitofrontal Cortex). The connection weights are updated as:

$$\delta V_i = \alpha S_{i,t-1} \left[ R - \sum_{j} A_{j,t-1} \right]^+$$

Where the $\alpha$ parameter is a standard learning rate parameter, settable between 0 (no learning) and 1 (instant adaptation). The orbitofrontal cortex's role is to inhibit other areas in response to changing situations. In contrast to the amygdala, these learned inhibitions can be learned as well as extinguished. Also, these inhibitions are context dependent. The OFC receives two new inputs and a new subsystem to integrate contextual and stimulus information. The OFC receives the same inputs the amygdala system does – CS, Bind nodes and $R$ – and also a context representation $CON$ from the hippocampus and the intended output $A$ from the amygdala. As for the amygdala model, the CS and Bind node inputs are concatenated into one vector $S$. This vector is multiplied with the context vector $CON$ to create an input matrix $T$:

$$T = \sum_{i=0}^{S} \sum_{j=0}^{CON} S_i CON_j$$

Thus there will be one $T$ node for every combination of $S$ and $CON$ nodes. The output $E_0$ of the OFC is calculated the same as for the amygdala:
\[ E_0 = \sum_{i=0}^{T} O_i \]  
\[ O_i = T W_i \]

with \( W \) being the connection weights for \( T \). The learning mechanism is almost the same as for the amygdala, the difference being that it is not constrained to only increase:

\[ \partial W_i = \beta T_{i,j} R_0 \]  

The reward function is different however:

\[ R_0 = \begin{cases} 
\sum \left[ A_i - R \right] - \sum_i O_i & R \neq 0 \\
\left[ \sum A_i - \sum_i O_i \right]^+ & \text{Otherwise} 
\end{cases} \]

\[ \text{Fig. 1. The abstract structure of CBBEL} \]

The function looks the way it does simply because when there is to lower the inhibition for the stimulus even if the inhibition is stronger than it needs to be – it is quite probably already set at the proper level for inhibition in an already encountered situation that may well occur again. As seen in fig. 1, the model is composed of three heavily interconnected components. While each component does something minimally useful by itself, it is the interconnections that make it work like an integrated whole. The hippocampus creates two kinds of outputs: Bound stimuli that fire whenever a given stimuli first shows up in a given place, and a context, used by the OFC for context-dependent inhibition. The amygdala uses stimuli and the primary reinforcer to persistently learn emotional associations. It creates an emotional value that may be partially or totally inhibited by the OFC before exiting the model. It also sends a vector of the current emotional outputs to OFC. The OFC binds together the stimuli (both bound and unbound) with the current context. It then compares the
suggested output of the amygdale with the current reinforcer and inhibits as needed, using the context-dependent stimuli.

4 The Control System

Controllers based on emotional learning have shown very good robustness and uncertainty handling properties [7-9], while being simple and easily implementable. To utilize our version of the Moren-Balkenius model as a controller, we note that it essentially converts two sets of inputs into the decision signal as its output. We have implemented a closed loop configuration using this block (termed CBBEL) in the feed forward loop of the total system in an appropriate manner so that the input signals have the proper interpretations. The block implicitly implemented the critic; the learning algorithm and the action selection mechanism used in functional implementations of emotionally based (or generally reinforcement learning based) controllers, all at the same time [9, 10, 11]. The control method for the SRMs is chosen based on some parameters including: usage, performance and speed range. Figure 2 shows the block diagram of the control system. In this structure, the speed controller produces the demanded motor torque in order to tracking the speed set point, in spite of the existing torque disturbances. This controller is based on CBBEL method. The commutator’s duty is to translate the demanded torque into the needed phase current, which enters the current controller. On the other hand, the current controller’s duty is to track the needed current for each phase.

5 Simulation Results

The main desirable characteristics achieved through the method proposed in this paper is the ability to pursue different objectives simultaneously through choosing suitable emotional cues, which are correlated to corresponding objectives. This is justified through three Simulink Simulations of a complete model of SRM. In the first simulation, our purpose is to reduce the steady state error of the speed response. The results are represented in figures 3.a, 3.b, and 3.c. It could be seen that the control effort and torque ripple are very high. In order to reduce the control effort, in the second step, the weighting coefficients was changed in such a way that the control effort weight becomes greater which limits the control effort. As could be seen in
figures 3.d, 3.e and 3.f although the steady state error is increased to about 1%, which is reasonable, the torque ripple is reduced. Then to minimize the energy of control effort we change the reward function (9) as a proper form. As could be seen in figures 3.g, 3.h and 3.i the steady state error is reduced to 0.5% and the control effort retained limited. Finally we could reach an interesting result in our 3rd simulation, which is depicted in figures 3.j, 3.k and 3.l. It could be seen that in comparison with the previous condition, the steady state error is reduced to 0.3%, the control effort is confined and the torque ripple is reduced. This interesting result is achieved by adding a redundant factor to the reward function equality (9). It seems our future works would be mainly about using this controller in smoothing control effort, reducing noise effects and increasing the system efficiency. Moreover, the controller design does not have much dependency on the plant model. In other words, a precise plant description is not essential for designing the controller, which renders the design process very simple. The necessary computations are not too complex either, so a simple microcontroller or a DSP chip can easily implement the controller. In comparison to the previous work [12], not only the system has retained its robustness, but also the responses have become more fast. Moreover, beyond the one in [12], this controller has the adaptation property due learning capability of the neural network implementation of CBBEL.

Fig. 3. Simulation results, in each row, speed response from 1000 to 3000 rpm (nominal speed), control effort and torque are represented, respectively

6 Conclusion

In this paper, we used an innovative intelligent controller, CBBEL, based on mammalian limbic emotional learning algorithms. We demonstrated the performance of CBBEL by applying it in speed regulation of a SRM. To emphasize the advantages of the proposed approach, we have shown how, through a proper choice of the emotional cue, we can consider secondary objectives. In the benchmark presented in the paper, for example, we have considered limiting the control effort as a secondary objective, in addition to tracking the set point, and have achieved, indirectly, control of torque ripple as well, which is also an objective of great importance in SRM
applications. We think that an algorithm for optimum tuning of CBBEL could make it even more efficient. It has also been shown that the proposed algorithm is very flexible; and it is possible to pursue other different goals simultaneously or to consider alternative goal fusion methodologies, e.g. fuzzy aggregation of goals, so as to form a suitable emotion signal.

References

[1] G.J. Donovari, P.J.Roche, R.C.Kavanagh, M.G.Egan and J.M.D.Murphi, “Neural network based torque ripple minimization in a switched reluctance motor”, 20th international conference, on Ind. Electronics, Control and Instrumentation, vol.2, pp. 1226-1231, 1994
[2] J.W.Finch and H.M.B.Metwally, “Control aspects of brushless drives using switched reluctance motor”, 4th International Conference of Power Electronics and Variable-SpeedDrives, IEE Conf. Publication, No.324, London, UK, pp. 237-242, 1990
[3] P.J. Lawrenson, J.M. Stephenson, P.T. Blenkinsop, J. Corda and N.N. Fulton, “Variable-speed switched reluctance motor”, IEE Proceedings Inst. Elect. Eng., Vol. 127, no. 4, pp. 253-265, June 1980.
[4] T.J.E Miller, Switched reluctance motors and their control, Oxford, Magma Physics Publishing and Clarendon Press-Oxford, 1993.
[5] Rashidi, F., “Design of context based emotional learning and reinforcement learning based intelligent controller for multiobjective systems”, M.sc. thesis, Tehran University, 2001.
[6] F.Gentili and S.Nicosia, Observer based control system design for switched reluctance motors” Proc of the 34th Conf. on Decision and Control, vol 1., pp115-120, 1995
[7] X. Xu, H-G He, D. Hu “Efficient Reinforcement Learning Using Recursive Least-Squares Methods” Journal of Artificial Intelligence Research 16, PP: 259-292, 2002
[8] M. Fatourechi, C. Lucas, A. Khaki Sedigh “Reducing Control Effort by means of Emotional Learning” Proceedings of 9th Iranian Conference on Electrical Engineering, (ICEE2001), Tehran, PP: 41-1 to 41-8, 2001
[9] Rashidi, F., Rashidi, M., “Emotional temporal difference learning based MLP neural network”, To appear in LNAI, 2004
[10] Moren, C. Balkenius “A Computational Model of Emotional Learning in The Amygdala” From animals to animals 6: Proceedings of the 6th International conference on the simulation of adaptive behavior, Cambridge, Mass., 2000. The MIT Press
[11] C. Balkenius, J. Moren “A Computational Model of Emotional conditioning in the Brain” Proceeding of the workshop on Grounding Emotions in Adaptive Systems, Zurich, 1998
[12] Rashidi, F., Lucas, C., “Speed control of switch reluctance motors using genetic algorithm based PID controller”, WSEAS Tran. On systems, 2002
1 Problems in Practical Situations

An automatic reasoning system usually consists of the following major components: (1) a formal language that represents knowledge, (2) a semantics that defines meaning and truth value in the language, (3) a set of inference rules that derives new knowledge from existing knowledge, (4) a memory that stores knowledge, and (5) a control mechanism that chooses premises and rules in each inference step. The first three components are usually referred to as a logic, or the logical part of the reasoning system, and the last two as an implementation of, or the control part of the system.

The most influential theory for the logic part of reasoning systems is the modern symbolic logic, especially, first-order predicate logic. For the control part, it is the theory of computability and computational complexity. Though these theories have been very successful in many domains, their application to reasoning in practical situations shows fundamental differences from human reasoning in these situations.

Traditional theories of reasoning are certain in several aspects, whereas actual human reasoning in practical situations is often uncertain in these aspects.

- The meaning of a term in traditional logic is determined according to an interpretation, therefore it does not change as the system runs. On the contrary, the meaning of a term in human mind often changes according to experience and context. Example: What is “game”?
- In traditional logic, the meaning of a compound term is completely determined by its “definition”, which reduces its meaning into the meaning of its components and the operator (connector) that joins the components. On the contrary, the meaning of a term in human mind often cannot be fully reduced to that of its components, though is still related to them. Example: Is a “blackboard” exactly a black board?
- In traditional logic, a statement is either true or false, but people often take truth value of certain statements as between true and false. Example: Is “Tomorrow will be cloudy” true or false?
- In traditional logic, the truth value of a statement does not change over time. However, people often revise their beliefs after getting new information. Example: After learning that Tweety is a penguin, you may change some of your beliefs formed when you only know that it is a bird.
In traditional logic, a contradiction leads to the “proof” of any arbitrary conclusion. However, the existence of a contradiction in a human mind will not make the person to do so. Example: Have you ever had a contradiction in your mind? Do you believe $1 + 1 = 3$ at that time?

In traditional reasoning systems, inference processes follow algorithms, therefore are predictable. On the other hand, human reasoning processes are often unpredictable, in the sense that sometimes a inference process “jumps” in an unanticipated direction. Example: Have you ever waited for “inspiration” for your writing assignment?

In traditional reasoning systems, how a conclusion is derived can be accurately explained. On the contrary, human mind often generate conclusions whose source cannot be backtracked. Example: Have you ever said “I don’t know why I believe that. It’s just my intuition.”?

In traditional reasoning systems, every inference process has a pre-specified goal, and the process stops whenever its goal is achieved. However, though human reasoning processes are also guided by various goals, they often cannot be completely achieved. Example: Have you ever tried to find the goal of your life? When can you stop thinking about it?

Furthermore, all the inference rules of traditional logic are deduction rules, where the truth of the premises is supposed to guarantee the truth of the conclusion. In a sense, in deduction the information in a conclusion is already in the premises, and the inference rule just reveals what is previously implicit. For example, from “Robins are birds” and “Birds have feather,” it is valid to derive “Robins have feather.”

In everyday reasoning, however, there are other inference rules, where the conclusions seem to contain information not available in the premises:

**Induction** produces generalizations from special cases. Example: from “Robins are birds” and “Robins have feather” to derive “Birds have feather.”

**Abduction** produces explanations for given cases. Example: from “Birds have feather” and “Robins have feather” to derive “Robins are birds.”

**Analogy** produces similarity-based results. Example: from “Swallows are similar to robins” and “Robins have feather” to derive “Swallows have feather.”

None of the above inference rules guarantee the truth of the conclusion even when the truth of the premises can be supposed. Therefore, they are not valid rules in traditional logic. On the other hand, these kinds of inference seem to play important roles in learning and creative thinking. If they are not valid according to traditional theories, then in what sense they are better than arbitrary guesses?

Finally, traditional logic often generates conclusions that are different from what people usually do.

**Sorites Paradox:** No one grain of wheat can be identified as making the difference between being a heap and not being a heap. Given then that one grain of wheat does not make a heap, it would seem to follow that two do not, thus three do not, and so on. In the end it would appear that no amount of wheat can make a heap.
Implication Paradox: Traditional logic uses “$P \rightarrow Q$” to represent “If $P$, then $Q$”. By definition the implication proposition is true if $P$ is false or if $Q$ is true, but “If $1+1 = 3$, then the Moon is made of cheese” and “If life exists on Mars, then robins have feather” don’t sound right.

Confirmation Paradox: Black ravens are usually taken as positive evidence for “Ravens are black.” For the same reason, non-black non-ravens should be taken as positive evidence for “Non-black things are not ravens.” Since the two statements are equivalent in traditional logic, white sacks are also positive evidence for “Ravens are black,” which is counter-intuitive.

Wason’s Selection Task: Suppose that I show you four cards, showing A, B, 4, and 7, respectively, and give you the following rule to test: “If a card has a vowel on one side, then it has an even number on the other side.” Which cards should you turn over in order to decide the truth value of the rule? According to logic, the answer is A and 7, but people often pick A and 4.

2 Assumptions of Reasoning Systems

None of the problems listed in the previous section is new. Actually, each of them have obtained many proposed solutions, in the form of various non-classical logics and reasoning systems. However, few of these solutions try to treat the problems altogether, but see them as separate issues.

A careful analysis reveals a common nature of these problems: they exist in “practical situations”, not in mathematics, which is about idealized situations.

At the time of Aristotle, the goal of logic was to find abstract patterns in valid inference that apply to all domains of human thinking. It remained to be the case until the time of Frege, Russell, and Whitehead, whose major interest was to set up a solid logic foundation for mathematics. For this reason, they developed a new logic to model valid inference in mathematics, typically the binary deduction processes that derives theorems from axioms and postulations.

What is the deference between “practical situations” as in everyday life and “idealized situations” as in mathematics? A key difference is their assumption on whether their knowledge and resources are sufficient to solve the problems they face. On this aspect, we can distinguish three types of reasoning systems:

Pure-Axiomatic Systems. These systems are designed under the assumption that both knowledge and resources are sufficient. A typical example is the notion of “formal system” suggested by Hilbert (and many others), in which all answers are deduced from a set of axioms by a deterministic algorithm, and which is applied to some domain using model-theoretical semantics. Such a system is built on the idea of sufficient knowledge and resources, because all relevant knowledge is assumed to be fully embedded in the axioms, and because questions have no time constraints, as long as they are answered in finite time. If a question requires information beyond the scope of the axioms, it is not the system’s fault but the questioner’s, so no attempt is made to allow the system to improve its capacities and to adapt to its environment.
Semi-axiomatic Systems. These systems are designed under the assumption that knowledge and resources are insufficient in some, but not all, aspects. Consequently, adaptation is necessary. Most current AI approaches fall into this category. For example, non-monotonic logics draw tentative conclusions (such as “Tweety can fly”) from defaults (such as “Birds normally can fly”) and facts (such as “Tweety is a bird”), and revise such conclusions when new facts (such as “Tweety is a penguin”) arrive. However, in these systems, defaults and facts are usually unchangeable, and time pressure is not taken into account [Reiter, 1987]. Many learning systems attempt to improve their behavior, but still work solely with binary logic where everything is black-and-white, and persist in always seeking optimal solutions of problems [Michalski, 1993]. Although some heuristic-search systems look for less-than-optimal solutions when working within time limits, they usually do not attempt to learn from experience, and do not consider possible variations of time pressure.

Non-axiomatic Systems. In this kind of system, the insufficiency of knowledge and resources is built in as the ground floor (as explained in the following).

Pure-axiomatic systems are very useful in mathematics, where the aim of study is to idealize knowledge and questions to such an extent that the revision of knowledge and the deadlines of questions can be ignored. In such situations, questions can be answered in a manner so accurate and reliable that the procedure can be reproduced by an algorithm.

We need intelligence only when no such pure-axiomatic method can be used, due to the insufficiency of knowledge and resources. Many arguments against logicist AI [Birnbaum, 1991] [McDermott, 1987], symbolic AI [Dreyfus, 1992], or AI as a whole [Searle, 1980] [Penrose, 1994], are actually arguments against a more restricted target: pure-axiomatic systems. These arguments are valid when they reveal many aspects of intelligence that cannot be produced by a pure-axiomatic system (though these authors do not use this term), but some of the arguments seriously mislead by taking the limitations of these systems as restricting all possible AI systems.

Unlike in mathematics, in practical situations a system has to work with insufficient knowledge and resources. By that, I mean the system works under the following restrictions:

Finite: The system has a constant information-processing capacity.
Real-Time: All tasks have time constraints attached to them.
Open: No constraint is put on the content of the experience that the system may have, as long as they are representable in the interface language.

3 NARS Overview

NARS (Non-Axiomatic Reasoning System) is an intelligent reasoning system designed to be adaptive and works under the assumption of insufficient knowledge and resources [Wang, 1995]. Here the major components of the system are briefly introduced. For detailed technical discussion, please visit the author’s website.
When a system has to work with insufficient knowledge and resources, what is the criteria of validity or rationality? This issue needs to be addressed, because the aim of NARS is to provide a normative model for intelligence in general, not a descriptive model of human intelligence. It means that what the system does should be “the right thing to do,” that is, can be justified against certain simple and intuitively attractive principles of validity or rationality.

In traditional logic, a “valid” or “sound” inference rule is one that never derives a false conclusion (that is, it will by contradicted by the future experience of the system) from true premises. However, such a standard cannot be used in NARS, which has no way to guarantee the infallibility of its conclusions. However, this does not mean that every conclusion is equally valid.

Since NARS is an adaptive system whose behavior is determined by the assumption that future situations are similar to past situations, in NARS a “valid inference rule” is one whose conclusions are supported by evidence provided by the premises used to derive them.

Furthermore, restricted by insufficient resources, NARS cannot exhaustively check every possible conclusion to find the best conclusion for every given task. Instead, it has to settle down with the best it can find with available resources.

Model-theoretic semantics is the dominant theory in the semantics of formal languages. For a language \( L \), a model \( M \) consists of the relevant part of some domain described in another language \( ML \), and an interpretation \( I \) that maps the items in \( L \) onto the objects in the domain (labeled by words in \( ML \)). \( ML \) is referred to as a “meta-language,” which can be either a natural language, like English, or another formal language. The meaning of a term in \( L \) is defined as its image in \( M \) under \( I \), and whether a sentence in \( L \) is true is determined by whether it is mapped by \( I \) onto a “state of affairs” that holds in \( M \).

With insufficient knowledge and resources, what relates the language \( L \), used by a system \( R \), to the environment is not a model, but the system’s experience. For a reasoning system like NARS, the experience of the system is a stream of sentences in \( L \), provided by a human user or another computer. In such a situation, the basic semantic notions of “meaning” and “truth” still make sense. The system may treat terms and sentences in \( L \), not solely according to their syntax (shape), but in addition taking into account their relations to the environment. Therefore, What we need is an experience-grounded semantics.

NARS does not (and cannot) use “true” and “false” as the only truth values of sentences. To handle conflicts in experience properly, the system needs to determine what counts as positive evidence in support of a sentence, and what counts as negative evidence against it, and in addition we need some way to measure the amount of evidence in terms of some fixed unit. In this way, a truth value will simply be a numerical summary of available evidence. Similarly, the meaning of a term (or word) in \( L \) is defined by the role it plays in the experience of the system, that is, by its relations with other terms, according to the experience of the system.

The “experience” in NARS is represented in \( L \), too. Therefore, in \( L \) the truth value of a sentence, or the meaning of a word, is defined by a set of sentences,
also in \( \mathbf{L} \), with their own truth values and meanings — which seems to have led us into a circular definition or an infinite regress. The way out of this seeming circularity in NARS is “bootstrapping.” That is, a very simple subset of \( \mathbf{L} \) is defined first, with its semantics. Then, it is used to define the semantics of the whole \( \mathbf{L} \).

As a result, the truth value of statements in NAL uniformly represents various types of uncertainty, such as randomness, fuzziness, and ignorance. The semantics specifies how to understand sentences in \( \mathbf{L} \), and provides justifications for the various inference rules.

As said above, NARS needs a formal language in which the meaning of a term is represented by its relationship with other terms, and the truth value of a sentence is determined by available evidence. For these purposes, the concept of (positive or negative) evidence should be naturally introduced into the language. Unfortunately, the most popular formal language used in first-order predicate logic does not satisfy the requirement, as revealed by the “Confirmation Paradox” \cite{Hempel, 1943}. A traditional rival to predicate logic is known as term logic. Such logics, exemplified by Aristotle’s Syllogistic, have the following features: \cite{Bocheński, 1970, Englebretsen, 1981}

1. Each sentence is \textit{categorical}, in the sense that it consists of a \textit{subject term} and a \textit{predicate term}, related by a \textit{copula} intuitively interpreted as “to be.”
2. Each inference rule is \textit{syllogistic}, in the sense that it takes two sentences that share a common term as premises, and from them derives a conclusion in which the other two (unshared) terms are related by a copula.

Traditional term logic has been criticized for its poor expressive power. In NARS, this problem is solved by introducing various types of \textit{compound terms} into the language, to represent \textit{set}, \textit{intersection} and \textit{difference}, \textit{product} and \textit{image}, \textit{statement}, and so on. The inference rules in term logic correspond to inheritance-based inference. Basically, each of them indicates how to use one item as another one, according to the experience of the system. Different rules correspond to different combinations of premises, and use different truth-value functions to calculate the truth value from those of the premises, justified according to the semantics of the system. The inference rules in NAL uniformly carry out \textit{choice}, \textit{revision}, \textit{deduction}, \textit{abduction}, \textit{induction}, \textit{exemplification}, \textit{comparison}, \textit{analogy}, \textit{compound term formation} and \textit{transformation}, and so on.

NARS cannot guarantee to process every task optimally — with insufficient knowledge, the best way to carry out a task is unknown; with insufficient resources, the system cannot exhaustively try all possibilities. Since NARS still needs to try its best in this situation, the solution used in NARS is to let the items and activities in the system compete for the limited resources. Again, the validity of the resource allocation policy is justified according to the past experience of the system (rather than its future experience), and the aim is to satisfy the goals of the system as much as possible.

In the system, different data items (tasks, beliefs, and concepts) have different priority values attached, according to which the system’s resources are
distributed. These values are determined according to the past experience of the system, and are adjusted according to the change of situation. A special data structure is developed to implement a probabilistic priority queue with a limited storage. Using it, each access to an item takes roughly a constant time, and the accessibility of an item depends on its priority value. When no space is left, items with low priority will be removed. The memory of the system contains a collection of concepts, each of which is identified by a term in the formal language. Within the concept, all the tasks and beliefs that have the term as subject or predicate are collected together.

The running of NARS consists of individual inference steps. In each step, a concept is selected probabilistically (according to its priority), then a task and a belief are selected (also probabilistically), and some inference rules take the task and the belief as premises to derive new tasks and beliefs, which are added into the memory. The system runs continuously, and interacts with its environment all the time, without stopping at the beginning and ending of each task. The processing of a task is interwoven with the processing of other existing tasks, so as to give the system a dynamic and context-sensitive nature.

4 Conclusion

To work in practical situations, a reasoning system should adapt to its environment, and works with insufficient knowledge and resources. Since traditional theories do not make such an assumption, new theories are needed. The practice of NARS shows that it is possible to build an intelligent reasoning system according to the above requirement, and such a system provides unified solutions to many problems.

References

[Birnbaum, 1991] Birnbaum, L. (1991). Rigor mortis: a response to Nilsson’s “Logic and artificial intelligence”. Artificial Intelligence, 47:57-77.
[Bocheński, 1970] Bocheński, I. (1970). A History of Formal Logic. Chelsea Publishing Company, New York. Translated and edited by I. Thomas.
[Dreyfus, 1992] Dreyfus, H. (1992). What Computers Still Can’t Do. MIT Press, Cambridge, Massachusetts.
[Englebretsen, 1981] Englebretsen, G. (1981). Three Logicians. Van Gorcum, Assen, The Netherlands.
[Hempel, 1943] Hempel, C. (1943). A purely syntactical definition of confirmation. Journal of Symbolic Logic, 8:122-143.
[McDermott, 1987] McDermott, D. (1987). A critique of pure reason. Computational Intelligence, 3:151-160.
[Michalski, 1993] Michalski, R. (1993). Inference theory of learning as a conceptual basis for multistrategy learning. Machine Learning, 11:111-151.
[Penrose, 1994] Penrose, R. (1994). Shadows of the Mind. Oxford University Press.
[Reiter, 1987] Reiter, R. (1987). Nonmonotonic reasoning. *Annual Review of Computer Science*, 2:147-186.

[Searle, 1980] Searle, J. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417-424.

[Wang, 1995] Wang, P. (1995). *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*. PhD thesis, Indiana University.
Commonsense Reasoning in and Over Natural Language

Hugo Liu and Push Singh

Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{hugo,push}@media.mit.edu

Abstract. ConceptNet is a very large semantic network of commonsense knowledge suitable for making various kinds of practical inferences over text. ConceptNet captures a wide range of commonsense concepts and relations like those in Cyc, while its simple semantic network structure lends it an ease-of-use comparable to WordNet. To meet the dual challenge of having to encode complex higher-order concepts, and maintaining ease-of-use, we introduce a novel use of semi-structured natural language fragments as the knowledge representation of commonsense concepts. In this paper, we present a methodology for reasoning flexibly about these semi-structured natural language fragments. We also examine the tradeoffs associated with representing commonsense knowledge in formal logic versus in natural language. We conclude that the flexibility of natural language makes it a highly suitable representation for achieving practical inferences over text, such as context finding, inference chaining, and conceptual analogy.

1 What Is ConceptNet?

ConceptNet (www.conceptnet.org) is the largest freely available, machine-useable commonsense resource. Structured as a network of semi-structured natural language fragments, ConceptNet presently consists of over 250,000 elements of commonsense knowledge. We were inspired dually by the range of commonsense concepts and relations in Cyc (Lenat, 1995), and by the ease-of-use of WordNet (Fellbaum, 1998), and hoped to combine the best of both worlds. As a result, we adopted the semantic network representation of WordNet, but extended the representation in several key ways.

First, we extended WordNet’s lexical notion of nodes to a conceptual notion of nodes, but we kept the nodes in natural language, because one of the primary strengths of WordNet in the textual domain is that its knowledge representation is itself textual. ConceptNet’s nodes are thus natural language fragments which are semi-structured according to an ontology of allowable syntactic patterns, and accommodate both first-order concepts given as noun phrases (e.g. “potato chips”), and second-order concepts given as verb phrases (e.g. “buy potato chips”).

Second, we extended WordNet’s small ontology of semantic relations, which are primarily taxonomic in nature, to include a richer set of relations appropriate to concept-level nodes. At present there are 19 semantic relations used in ConceptNet, representing categories of, inter alia, temporal, spatial, causal, and functional knowl-
edge. By combining higher order nodes with this relational ontology, it is possible to represent richer kinds of knowledge in ConceptNet beyond what can be represented in WordNet (Fig 1.). For example, we can represent a layman’s common sense observation that “you may be hurt if you get into an accident” in ConceptNet as EffectOff(“get into accident”, “be hurt”). Note that because the knowledge representation is semi-structured natural language, there are often various ways to represent the same knowledge. This is a source of ambiguity, but as we will argue in this paper, maintaining some ambiguity lends us greater flexibility for reasoning.

*Fig. 1.* An excerpt from ConceptNet’s semantic network of commonsense knowledge. Relation names are expanded here for clarity

Third, we supplement the ConceptNet semantic network with some methodology for reasoning over semi-structured natural language fragments. This methodology prescribes techniques for managing the ambiguity of natural language fragments, and for determining the context-specific similarity of nodes. For example, sometimes we want the nodes “buy food” and “purchase groceries” to be synonymous in an inference chain, and other times, not.

Fourth, we supplement the ConceptNet semantic network with a toolkit and API which supports making practical commonsense inferences about text, such as context finding, inference chaining, and conceptual analogy.

In a related paper (Liu & Singh, 2004a), we describe how ConceptNet was created, how it is structured, how it compares with other commonsense knowledge bases, and how it has been used and evaluated. This paper begins with a few pertinent details of the system but largely focuses on the knowledge representation aspect of ConceptNet,
that is, how to reason over concepts expressed as semi-structured natural language fragments. We also speak more generally about the suitability of natural language as a knowledge representation for commonsense reasoning.

This paper is structured as follows. First, we give a brief history of the origin and structure of ConceptNet. Second, we discuss the affordances and limitations of representing commonsense knowledge in natural language, particularly in regards to its suitability for making practical commonsense inferences about text. Third, we present some methodology for reasoning about semi-structured natural language concepts in ConceptNet. We conclude with a summary of contribution.

2 Origin and Structure of ConceptNet

We explain briefly the history of ConceptNet with respect to the Open Mind Commonsense (OMCS) Project, how ConceptNet was mined from the OMCS repository of sentences, and how ConceptNet’s nodes and edges are structured.

2.1 Origin

ConceptNet is a machine usable resource mined out of the Open Mind Commonsense (OMCS) corpus (Singh et al. 2002), a collection of nearly 700,000 English sentences of commonsense facts collected from over 14,000 contributors from around the world, over the past three years.

CRIS (Commonsense Robust Inference System), ConceptNet’s earliest predecessor, mined predicate argument structures from OMCS, where arguments were semi-structured natural language fragments, and used this for conceptual expansion (Liu & Lieberman, 2002). Since then, we finalized the ontology of semantic relations and added an inference toolkit that is distributed with the semantic network to support some practically useful textual inferences tasks. Recently, we also added methods for automatically comparing and reconciling natural language nodes to make inference more flexible and robust.

ConceptNet is produced by an automatic process which applies a set of ‘commonsense extraction rules’ to the semi-structured English sentences of the OMCS corpus. The key to being able to do this is that the OMCS website already elicits knowledge in a semi-structured way by prompting users with fill-in-the-blank templates (e.g. “The effect of [falling off a bike] is [you get hurt]”). A pattern matching parser uses roughly 40 mapping rules to easily parse semi-structured sentences into an ontology of predicate relations, and arguments which are short fragments of English. These arguments are then normalized to conform to preferred syntactic patterns. Certain stop-words and stop-parts-of-speech are filtered out, and the verb and nouns are reduced to their canonical base forms. A small part-of-speech-driven grammar filters out non-compliant text fragments (thus only a subset of the OMCS knowledge is used in ConceptNet) to ensure all arguments conform to these syntactic constraints.
2.2 Structure

ConceptNet nodes are natural language fragments semi-structured to conform to preferred syntactic patterns which fall into three general classes: Noun Phrases (things, places, people), Attributes (modifiers), and Activity Phrases (actions and actions compounded with a noun phrase or prepositional phrase, e.g.: “turn on water,” “wash hair”). In the normalization process, verbs are stripped to their base form, the count of nouns is stripped, and parts-of-speech which have lesser semantic value like determiners (e.g. “a”, “the”, “two”) and modals (e.g. “might”, “could”, “may”) are stripped. A portion of the concept grammar is given as part-of-speech patterns in Table 1.

| Node class | A portion of the grammar | Examples of valid nodes |
|------------|--------------------------|-------------------------|
| Noun Phrases | NOUN; NOUN NOUN; ADJ NOUN; NOUN PREP NOUN | “apple”; “San Francisco”; “fast car”; “life of party” |
| Attributes | ADJ; ADV ADJ | “red”; “very red” |
| Activity Phrases | VERB; VERB NOUN; ADV VERB; VERB PREP NOUN; VERB NOUN PREP NOUN | “eat”; “eat cookie”; “quickly eat”; “get into accident”; “eat food with fork” |

ConceptNet edges are described by an ontology of at present 19 binary relations shown below in Table 2. These relations were chosen because the original OMCS corpus was built largely through its users filling in the blanks of templates like ‘a hammer is for _____’. Thus the relations we chose to extract largely reflect the original choice of templates used on the OMCS web site.

| Category | Semantic Relations – (and an explanation) |
|----------|------------------------------------------|
| Things | IsA – (corresponds loosely to hypernym in WordNet) |
|         | PropertyOf – (e.g. (PropertyOf “apple” “healthy”)) |
|         | PartOf – (corresponds loosely to holonym in WordNet) |
|         | MadeOf – (e.g. (MadeOf “bottle” “plastic”)) |
| Events | FirstSubeventOf, LastSubeventOf – (e.g. (FirstSubeventOf “act in play” “learn script”)) |
|         | EventForGoalEvent – (e.g. (EventForGoalEvent “drive to grocery store” “buy food”)) |
|         | EventForGoalState – (e.g. (EventForGoalState “meditate” “enlightenment”)) |
|         | EventRequiresObject – (e.g. (EventRequiresObject “apply for job” “resume”)) |
| Actions | EffectOf – (e.g. (EffectOf “commit perjury” “go to jail”)) |
|         | EffectOfIsState – (e.g. (EffectOfIsState “commit perjury” “criminal prosecution”)) |
|         | CapableOf – (e.g. (CapableOf “police officer” “make arrest”)) |
| Spatial | OftenNear – (e.g. (OftenNear “sailboat” “marina”)) |
|         | LocationOf – (e.g. (LocationOf “money” “in bank account”)) |
| Goals | DesiresEvent, DesiresNotEvent – (e.g. (DesiresEvent “child” “be loved”)) |
| Functions | UsedFor – (e.g. (UsedFor “whistle” “attract attention”)) |
| Generic | CanDo – (e.g. (CanDo “ball” “bounce”)) |
|         | ConceptuallyRelatedTo – (e.g. (ConceptuallyRelatedTo “wedding” “bride”; “groom”) |
As illustrated by the examples in Table 1, the semantics of the predicate relations in ConceptNet are quite informal. Even for a particular semantic relation, the syntactic and/or semantic type of the arguments are not formally constrained, though some predicate names imply some typing (e.g. EventForGoalEvent, EventForGoalState, EventRequiresObject). In general, the usage and scope of each semantic relation can be best and most intuitively ascertained by looking at the original choice of templates used on the OMCS web site (At: http://openmind.media.mit.edu)

3 Commonsense Reasoning in Natural Language?

In this section, we discuss some of the strengths and weaknesses of representing and reasoning with commonsense knowledge in natural language.

3.1 Where Logic Excels

There is an important representational tradeoff between logic and natural language. Logic is precise – its symbols are unambiguous, and inference amounts to deductive theorem proving. Its strength lies in its stable and systematic way of evaluating and maintaining the truth of expressions. In technical domains with little ambiguity where precision is important, logic is a superb framework. But what about the vague notion of a “common sense” domain? Cyc, for example, represents its commonsense knowledge in a language called CycL, which is essentially a second-order logical language with second-order features such as quantification over predicates. John McCarthy first outlined the basic approach of representing commonsense knowledge with predicate logic in his classic paper, “Programs with Common Sense” (1958). Examples given in the paper seem quite appealing for their elegance. For example, you can represent the statement, “if you are at your car and you drive from home to the airport, then you are at the airport,” using the following logical statement: canachult(at(I,car), go(home,airport,driving), at(I,airport)). Yet the development and application of commonsense reasoning systems using the logical approach has turned out to be not so straightforward as this example might suggest. To make this example really work requires a substantial amount of additional logical scaffolding, to precisely define the terms used in the statement and their interrelationships, which has turned out to be a task of daunting complexity.

3.2 How do People Represent and Reason?

There seems to be a divergence between the logical approach to reasoning and what is known about how people reason. Human commonsense reasoning has a number of properties that distinguish it from traditional logical reasoning, and that have inspired various extensions to the logical approach. Most notably, commonsense knowledge is largely defeasible and context-sensitive. People have no problem believing that “birds can fly,” even though they know that “penguins are birds who cannot fly”, and that “birds with broken wings cannot fly.”
One explanation for why human reasoning is defeasible and context-sensitive is Lakoff and Johnson’s prototype theory of human categorization (1980), which argues that people define categories based on its most central and characteristic (and often caricaturalistic) example or prototype. Whereas ontologists would like to define all members in a category as being equal in membership, define a category as having sharp boundaries, and define membership by a set of common features, empirical work on human categorization shows this to not be true. People confer membership in a category to varying extents, often draw fuzzy boundaries for a category, and group members into a category based on how members resemble each other. Fehr & Russell’s empirical study of common sense emotions reveals the extent that prototypes play in our categorization of emotions (1984). Logical notation is rigorous, and is completely amenable to ontologies, strict boundaries, and clean definitions, but has trouble with the inexactness and highly nuanced nature of human prototype categorization.

In addition, whereas logical reasoning is deductive, human reasoning is largely inductive, abductive, and empirical, where (over-)generalizations from known experiences plays a prominent role. One reason why humans are such successful inductive reasoners is that we collect a very large number of features in observing objects and events, thus providing a wealth of information from which patterns can emerge. Whereas the logical tradition excels at deductive reasoning, it has had much difficulty formalizing induction, and attempts to do so have generally involved great complexity, cf. (Flach, 1995). While logic is a highly precise and expressive language, it has difficulty modeling the imprecise way that human categorize and compare things based on prototypes, and also difficulty emulating human reasoning which is largely inductive and highly associative.

3.3 Natural Language as a Lingua for Commonsense

Such concerns led us to consider using natural language expressions as more central components of a commonsense knowledge representation. In our work with ConceptNet, we are exploring natural language as a lingua for representing common sense. In many ways, natural language fragments are a much more flexible and accessible representation for common sense. Whereas logical symbols have no a priori meaning outside of their immediate definition, words and expressions automatically inherit meaning from the way they used in our culture. Because we don’t have to first read off the axioms in which they play part to interpret word symbols, the representation becomes much simpler to author and inspect. Consider that Cyc employs logicians to author its knowledge, while ConceptNet was automatically mined from a knowledge base of English sentences contributed by 14,000 people over the web, many of whom have no computer science background. Cyc, posed in logical notation, does indeed contain very high quality knowledge. Still, accessibility and lowered complexity of authoring is a great boon, because it enables new kinds of knowledge acquisition, like the OMCS web site. It may be appropriate to give the caveat at this point in the paper that unlike logic, natural language and thus, ConceptNet’s reasoning, will not have a complete formal semantics, evading exactness and absolute truths.
By posing common sense in natural language, we can benefit from the implicit conceptual framework of human language. For example, WordNet is a lexical-conceptual framework which gives us the subsumption relationships between words. This lexical hierarchy allows us to heuristically infer the related between two nodes. For example, in ConceptNet, “buy food” and “purchase groceries” are two intrinsically similar nodes. Computational resources like WordNet, Longman’s Dictionary of Contemporary English (LDOCE), Beth Levin’s English Verb Classes (1993), and FrameNet (Fillmore & Baker, 2001) reveal the various synonym and subsumption relationships between “buy” and “purchase” and between “food” and “groceries.” Allowing us to quantify the semantic similarity between two nodes or symbols affords us the ability to reason inductively over concepts, and almost by the very nature of representing knowledge in natural language, we are categorizing objects and events like people do. Whereas logic is a synthetic representation, creating and manipulating symbols in a closed world, natural language is an empirical representation, juggling concepts that are already defined and related in a human language.

Ambiguity is a particular aspect that needs to be dealt with when reasoning in natural language. In logic, it is approached as something negative, to be eradicated. But human language is full of ambiguity, and perhaps ambiguity is one of the greatest strengths to human language. Remember that the reason why there is ambiguity in words is because we have such a wealth of knowledge backing various interpretations of that word, and having this background knowledge for free is hardly a bad situation. What is necessary is a way to bound and manage the ambiguity of natural language fragments so that unambiguous computer algorithms can manipulate them in a systematic way. In the next section, we will present a methodology which prescribes ways of managing the ambiguity of concepts represented in natural language.

Natural language has its weaknesses as a representation for common sense knowledge. Whereas a logical symbol is concise, there may exist many different natural language fragments which mean essentially the same thing, and may seem equally suitable to include, and so in this sense logic can be seen as more economical. It is also sometimes more difficult to be precise in natural language. For example, what is the precise color of a “red apple?” In logic, we might be able to formally represent the range in the color spectrum corresponding to a “red apple,” but in natural language, the word “red” is imprecise and has various interpretations. Consider the differing colors which map to “red apple” versus “red wine” versus “red hair.” WordNet has tried to address this issue of semantic leakage by imposing boundaries on word called word senses. In many cases, such boundaries are very clear, as in the case of homonyms (e.g. river bank versus financial bank), but in the case of more systematic polysemies (e.g. WordNet has different senses for a short sleep versus a long sleep), it is clear that such boundaries are artificial.

Thus far, we have argued that representing common sense in natural language is a good idea because the implicit conceptual framework of human language makes nodes and symbols meaningful by default, gives us a way to quantify the similarity of nodes and symbols, and is thus more amenable to inductive reasoning. We have argued that natural language is a more accessible representation for authoring. In
addition to these points, we would like to add that natural language as a representation is highly desirable when the goal of the application is to reason over text. One of the primary applications of commonsense knowledge bases is to draw inferences about text. Using a logical framework of commonsense like Cyc to reason about text is quite complex. Text, which is inherently ambiguous, must first be mapped into Cyc’s unambiguous logic, which is often problematic. There must be rules to map every variety of textual expression into an ontology of logical concepts, and this generally requires a very deep (and very hard to build) semantic parser. By maintaining a natural language knowledge representation, we can more readily reason about text, requiring only a surface parse of the text to extract all the contained concepts. Now all the diverse ways of expressing the same concept come in handy for concept recognition.

We now move on to the next section, where we present a methodology for reasoning over semi-structured natural language fragments. This will flesh out some of the discussion in this section.

4 Methodology for Reasoning Over Natural Language Concepts

In this section, we present some methodology for reasoning over semi-structured natural language fragments used in ConceptNet. All natural language concepts in ConceptNet possess a rich *intrinsic semantics*, based on the meaning they inherit from human language. For example, creating the concept “fast car,” and accepting the caveat that we consider chiefly the most common interpretation, we instantly know (tempered with an uncertainly model) that this is a type of “fast vehicle,” a fast form of transportation, a car with a speed, a concept that bears family resemblance to “slow car” and to “fast bullet,” and a myriad of other inferences could be made from the very certain to wildly speculative. If we can tolerate the interpretational ambiguity by taking the most common interpretation, we can situate the concept within the conceptual framework of human language and see how the concept bears similarities to other concepts in numerous ways – similarities that can be computed using computational resources. The following subsections present methodology for computing pair-wise similarities between concepts, and flexible inferencing.

4.1 Computing Conceptual Similarity

The basic premise of concepts as natural language fragments is that their interpretation is situated within the conceptual framework of language. For each concept, there is a way to decompose that concept into first-order atomic concepts by applying a surface parse. For the structure of concepts in ConceptNet, please refer to the syntactic grammar of concepts given in Table 1. The first-order atomic concepts will consist of simple noun phrases (note that the grammatical class “Noun Phrases” from Table 1 contains both simple and compound members), simple prepositional phrases, simple attributes, and simple verbs.
To compute the meaning of a concept, the concept is first decomposed into first-order atomic concepts, while preserving the dependence relationships between the concepts. For example: “buy good cheese” decomposes into “buy,” “good,” and “cheese” where “good” is an attribute of “cheese” and “cheese” plays the thematic patient role to “buy.” Note that this is only a surface parse. Liu’s Bubble Lexicon (2003) would support deeper parses and more highly nuanced interpretations (e.g. “fast car” can be interpreted variously with “fast” describing the car’s top speed, the car’s speed of motion, the speeding rating of the car’s tires, et cetera), but this is left for future work. After a concept is decomposed, each atom is situated within the conceptual frameworks of WordNet, Longman’s Dictionary of Contemporary English (LDOCE), Beth Levin’s English Verb Classes, and FrameNet. We chose to maintain these multiple representations because we are concerned that the inferential distance within any single resource will be overly biased. For example, the atoms “good” and “cheese” and “buy” are mapped onto the lexical entry for cheese in WordNet and LDOCE, and the verb “buy” is mapped into the lexicons of Levin’s Verb Classes, and FrameNet. The meaning of a concept is represented as a collection of pointers from the decomposed concept’s atoms into each of the semantic resources of WordNet, LDOCE, Levin Verb Classes, and FrameNet.

To compute the similarity of two concepts, we produce a heuristic score by comparing corresponding atoms (verb matching verb, noun modifier matching noun modifier, etc.) of the two decomposed concepts using each of the semantic resources. First, within each resource, a similarity score is produced for each pair of corresponding atoms. WordNet, LDOCE, and FrameNet’s inheritance structure for verbs can be structured as semantic networks in which inferential distance is given to be proportional to the number of hops away (we heuristically weight isA and synonym links differently). For example, in WordNet’s subsumption hierarchy, the inferential distance between “apple” and “food” is proportional to 3, because “apple” isA “edible fruit” isA “produce” isA “food”. LDOCE also gives morphological relationships, which helps with action/state variations such as “relax” versus “feel relaxation.” A similarity score should be inversely proportional to inference distance. In Levin’s Verb Classes, each verb belongs to multiple alternation classes. Inferential distance here is proportional to the percentage of alternation classes shared. The weighted sum of the similarity scores is produced for each atom using each of the resources is taken. Weights on each semantic resource should be proportional to the predictive accuracy of that resource. Weights on atoms should be proportional to the relative importance of the different atom types. For example, a noun modifier is generally not as important as the noun it modifies.

The particular coefficients used in heuristic similarity vary with different semantic resources, and change depending on the context of the reasoning task. Some similarity percentages of concepts are computed in ConceptNet as given in Table 3. (Assuming default importance weights on verbs, modifiers, noun phrases, and prepositional phrases.)
Table 3. Some pairwise similarities in ConceptNet

| Similarity 1                  | Similarity 2                  | Percentage |
|------------------------------|------------------------------|------------|
| "apple" ~ "red apple" (76%)  | "buy food" ~ "purchase groceries" (69%) |            |
| "big dog" ~ "animal" (53%)   | "relax" ~ "feel relaxation" (72%) |            |
| "red" ~ "red apple" (36%)    | "have accident" ~ "get into accident" (64%) |            |

Of course computing conceptual similarity using lexical inferential distance is very difficult, as demonstrated in Table 3. Without additional insight into how a concept is generally interpreted by default (which would require a difficult, deep parse), we can only make heuristic approximations as to the relative contributions of the verb, noun phrase, attribute, and prepositional phrase to the meaning of a concept. In future work, we hope to further exploit knowledge in WordNet glosses and FrameNet frames to further nuance similarity scoring. However, our knowledge-based scoring of concepts based on inferential distance already goes beyond some previous work in reconciliation of natural language fragments, such as notably, William Cohen’s WHIRL system (2000), which uses TF-IDF, a statistical vector similarity metric.

4.2 Flexible Inference

One of the strengths of representing concepts in natural language is the ability to add flexibility and fuzziness to improve inference. We again give the caveat that inferences in semantic networks are not logical deductions like in Cyc, but rather are based on graph reasoning methods like spreading activation (Collins & Loftus, 1975), structure mapping (Gentner, 1983), and network traversal. Graph-based reasoning is associative and thus not as expressive, exact, or certain as logical inferences, but it is much more straightforward to perform, and useful for reasoning practically over text. In this section, we demonstrate three kinds of flexible inference in ConceptNet: context finding, inference chaining, and conceptual analogy.

Context Finding. One task useful across many textual reasoning applications is determining the context around a concept, or around the intersection of several concepts. The GetContext() feature in the API makes this easy. For example, computing the top ten concepts in the contextual neighborhood of “go to bed” yields “take off clothes,” “go to sleep,” “sleep,” “lie down,” “lay down,” “close eye,” “turn off light,” “dream,” “brush tooth,” and “snore.”

The contextual neighborhood around a node is found by performing spreading activation from that source node, radiating outwardly to include other concepts. The relatedness of any particular node is not just a function of the number of links away it is, but also considers how many paths there are from that node to the source node, and the directionality of the edge. In addition, pairwise similarity of nodes indicates the mutual information between the two nodes, allowing similar nodes to be aggregated, leading to a more accurate estimation of contextual neighborhood. For example, in the above example, the co-presence of “sleep” with “go to sleep” and “lay down” with “lie down” mutually promote each other higher up the list of relevant concepts.
How has GetContext() been applied for practical commonsense reasoning? Musa et al.’s GloBuddy system (2003) is a dynamic Berlitz phrase book that uses ConceptNet’s GetContext() feature to generate a topical collection of phrases paired with their translations. For example, entering “restaurant” would return phrases like “order food” and “waiter” and “menu,” and their translations into the target language. Now suppose we feed in all the extracted concepts in a particular passage of text into GetContext() and take their intersection. GetContext() used in this way serves as a “topic spotter” of sorts. Eagle et al. (2003) used ConceptNet and this method to gist conversation topics from heard conversations.

**Inference Chaining.** Another basic type of inference that can be done on a graph is building inference chains: Traversing the graph from one node to another node via some path of connectedness. This is not logical inference per se but a simplification of modus ponens transitive reasoning. The FindPathsBetween() feature in the ConceptNet Practical Reasoning API supports building inference chains. Temporal and spatial chains are particularly good examples. For example, ConceptNet can generate all the temporal chains between “buy food” and “fall asleep.” One chain may be: “buy food” $\rightarrow$ “have food” $\rightarrow$ “eat food” $\rightarrow$ “feel full” $\rightarrow$ “feel sleepy” $\rightarrow$ “fall asleep.” Each of these chains can be seen as being akin to a “script.” Being able to compute the pairwise conceptual similarity is particularly crucial to the robustness of inference chaining, because it makes these chains “fuzzy.” Suppose that we started with “buy steak” instead of “buy food,” and suppose there is no temporal knowledge about what happens after “buy steak.” By knowing that “buy steak” is a special case of “buy food,” since “food” subsumes “steak,” we can now make the inference “fall asleep.”

Liu et al.’s Emotus Ponens (2003) system performs affective text classification using a slightly different representation than ConceptNet. It uses essentially inference chaining for assessing the affect of a concept. Consider that a small subset of the concepts in ConceptNet are first affectively classified into one of six affect categories (happy, sad, angry, fearful, disgusted, surprised). The affect of any unclassified concept can be assessed by finding all the paths which lead to each of these six affectively known categories, and then judging the strength and frequency of each set of paths. This is the graph-based equivalent of a $k$-nearest-neighbor classifier.

**Conceptual Analogy.** A third practical textual inference task is finding concepts which are structurally analogous. In the ConceptNet Practical Reasoning API, there is a GetAnalogousConcepts() feature that returns a list of structurally analogous concepts given a source concept. Structural analogy is not just a measure of semantic distance. For example, “wedding” and “bride” are semantically close but structurally unlike. Structurally, a “funeral” is much more like a “wedding.” Here is another example. Typing “couch” into the GetAnalogousConcepts(), examples of top results returned include “sofa,” “chair,” “bed,” “seat” because they share similar properties and have similar functions. We are employing structure-mapping methods (Gentner, 1983) over the ConceptNet graph to generate these simple conceptual analogies. Just like we’ve done with the GetContext() feature, it is also easy to contextually bias the GetAnalogousConcepts() feature. We can prefer to see analogous concepts which fall within a particular domain (defined by another GetContext()), or by biasing the
numerical weights of particular semantic relations, we can emphasize functional similarity versus object attribute similarity versus temporal similarity. As with context finding and inference chaining, conceptual analogy is made flexible by using computed node similarity as glue to prevent missed structural similarities. For example, if functionOf(“massage”, “feel relaxation”) and functionOf(“meditation”, “unwind”), knowing that “feel relaxation” and “unwind” are very similar prevents GetAnalogousConcepts() from overlooking this shared property of “massage” and “meditation.”

Liu et al. are using this same idea of finding what concepts have in common to augment the aforementioned Emotus Ponens system. The basic idea behind this augmentation is that certain kinds of structural analogy, such as concepts sharing PropertyOf’s, IsA’s, and UsedFor’s, can be predictive of affective similarity. They hope that expanding concepts with analogous concepts can expand the coverage of the system and thus improve the performance of the affective classification.

**A Word on Evaluation.** Traditionally, it is quite difficult to produce useful standalone objective evaluations of knowledgebase quality. While we have performed some evaluative analysis over ConceptNet and written about it in (Liu & Singh, 2004a), it is often equally insightful to see evaluations of ConceptNet in the context of how they improve intelligent applications. We and others at our lab have developed a host of applications using early versions of ConceptNet. We survey these applications in (Lieberman et al., 2004). These seem to be entirely new kinds of applications, in that it is difficult to imagine how they could possibly be built without making use of commonsense inferences over natural language text. Many of these projects are evaluated and we invite the reader to follow the literature if he/she is interested in these in-context evaluations of ConceptNet.

## 5 Conclusion

Presently the largest freely available commonsense resource, ConceptNet comes with a knowledge browser, and a preliminary set of tools to support several kinds of practical inferences over text. ConceptNet follows the easy-to-use semantic network structure of WordNet, but incorporates a greater diversity of relations and concepts inspired by Cyc.

To maintain an easy-to-use knowledge representation, while at the same time incorporating more complex higher-order commonsense concepts and relations, we chose to represent concepts as semi-structured natural language fragments. This novel use of language as knowledge representation can very elegantly represent both first-order (e.g. “apple pie”) and second-order concepts (e.g. “bake apple pie”), and unlike logical symbols, the a priori meaning of the words make it possible to quantify the implicit similarity of two concepts.

In this paper we presented some novel methodology for computing the pairwise similarity of concepts using a variety of lexical resources such as WordNet, LDOCE, FrameNet, and Levin Verb Classes. We showed how computing the similarities between concepts enables more flexible and robust inferences. We also looked more
broadly at the knowledge representational tradeoffs between formal logic and semi-structured natural language, and concluded that the flexibility afforded by natural language made it a highly suitable representation for a system whose goal is to make practical inferences over text.

That ConceptNet is already being widely used in a number of research projects such as those surveyed in (Lieberman et al., 2004) is testament to the resource’s practicality and usefulness to researchers with no background in linguistics or commonsense reasoning. We hope that this paper has encouraged the reader to consider using ConceptNet within their own projects, and that it will spur further thinking about semi-structured natural language as a serious representation for reasoning.

Acknowledgements

We extend our thanks to the many people at the Media Lab who have used ConceptNet in their projects, especially Barbara Barry, Nathan Eagle, Henry Lieberman, and Austin Wang, and especially to the 14,000 people across the web who contributed some of their common sense to the Open Mind Common Sense web site. This work was supported by the many sponsors of the Media Lab.

References

1. Cohen, W. (2000). WHIRL: A word-based information representation language. Journal of Artificial Intelligence, 118 (163-196).
2. Collins, A. and Loftus, E. (1975). A Spreading-Activation Theory of Semantic Processing. Psychological Review, 82(6):407-428.
3. Eagle, N., Singh, P., and Pentland, A. (2003). Common sense conversations: understanding casual conversation using a common sense database. Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003).
4. Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. Journal of Experimental Psychology: General, 113, 464-486.
5. Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. MIT Press.
6. Fillmore, C., & Baker, C.F. (2001). Frame semantics for text understanding. Proceedings of WordNet and Other Lexical Resources Workshop, NAACL.
7. Flach, P.A. (1995). Conjectures. An inquiry concerning the logic of induction. PhD thesis, Katholieke Universiteit Brabant.
8. Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, pp 155-170.
9. Lakoff, G. & Johnson, M. (1980). Metaphors We Live by. University of Chicago Press.
10. Lenat, D.B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11)
11. Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation. MIT Press.
12. Lieberman, H., Liu, H., Singh, P., Barry, B. (2004). Beating Some Common Sense Into Interactive Applications. To Appear in AI Magazine.
13. Liu, H. (2003). Unpacking meaning from words: A context-centered approach to computational lexicon design. In Modeling and Using Context, 4th International and Interdisciplinary Conference, Proceedings, CONTEXT 2003. pp. 218-232. LNCS. Springer.
14. Liu, H. and Lieberman, H. (2002). Robust photo retrieval using world semantics. Proceedings of LREC2002 Workshop: Using Semantics for IR, Canary Islands, 15-20
15. Liu, H., Lieberman, H., Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. In Proceedings of IUI 2003. Miami, Florida.
16. Liu, H., Singh, P. (2004a). ConceptNet: A Practical Commonsense Reasoning Toolkit. To appear in BT Technology Journal. At: http://web.media.mit.edu/~push/ConceptNet.pdf
17. McCarthy, J. (1958). Programs with Common Sense. Proceedings of the Teddington Conference on the Mechanization of Thought Processes
18. Musa, R., Scheidegger, M., Kulas, A., Anguilet, Y. (2003) GloBuddy, a Dynamic Broad Context Phrase Book. In Proceedings of CONTEXT 2003. pp. 467-474. LNCS. Springer.
19. Singh, P. et al. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In Proceedings of ODBASE’02. LNCS. Heidelberg: Springer-Verlag.
A Library of Behaviors: Implementing Commonsense Reasoning About Mental World

Boris Galitsky

School of Computer Science and Information Systems,
Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK
galitsky@dcs.bbk.ac.uk

Abstract. We introduce the simulation approach to reasoning about mental world that is based on exhaustive search through the space of available behaviors. This approach to reasoning is implemented as a logic program in a natural language multiagent mental simulator NL_MAMS, which is capable of yielding the totality of possible mental states few steps in advance, given the current mental state of participating agents in natural language. NL_MAMS is intended to contain a domain-independent library of behaviors and serve as a commonsense reasoning component of a large-scale decision-making, control or educational software system. We perform the analysis of how the behaviors obtained in one domain cover another domain in terms of decision-making of participating agents.

1 Introduction: Practical Reasoning About Mental States

Reasoning about mental attributes and behavior patterns is an important component of human intellectual activity. Quite a few formalisms have been suggested to reproduce the peculiarities of human reasoning in the way of logical calculi. In these calculi the laws of “mental world” are encoded via axioms, and derived theorems are expected to describe the states and actions of agents in the mental world. Particularly, the modal logics are quite successful means to represent the notion of knowledge, belief and intention in connection to the other (“physical”) properties of the real world.

Building the practical systems which model the mental world, we have come to the conclusions that pure axiomatic reasoning cannot deliver sufficiently rich number of theorems to adequately describe the mental states of agents. It has been understood a few decades ago that staying within the bounds of classical logic, it is impossible to represent the certain phenomenology of human reasoning. Non-classical logics have enabled artificial intelligence to model reasoning of agents in time and space, in the conditions of uncertainty and inconsistency, and reasoning about the behaviors of each other. However, nowadays there is still a lack of complex real-world examples, requiring software implementation of non-classical calculi.

In the series of studies we intend to build a framework to simulate the human reasoning within the bounds of mental world in as detailed way as possible. We believe
that rather simulation-based approach than the deductive reasoning-based one is suitable to express the laws of mental world and to apply them to produce as realistic scenarios as possible for practical applications. The main goal of the desired system is formulated as obtaining a set of consecutive mental states, which are expected to follow the initial mental state that is given. We consider a solution to this problem valid if it is as close to the natural behavior (from the experts’ viewpoint) as possible. The task of the simulator to be presented, NL_MAMS, is to generate such scenarios for the totality of initial mental states, extending solely the library of behaviors. NL_MAMS stands for Natural Language Multiagent Mental Simulator ([3], www.dcs.bbk.ac.uk/~galitsky/Nl_Mams). We have already verified [4] that the simulation approach is applicable in a variety of domain of various natures. In this paper we analyze how the library of behaviors affects functionality of the simulator.

We start with traditional considerations concerning mental world and comment on the limitations of reasoning about it using modal logics. Mental world is the subset of the reality described by the entities belief, desire and intention (BDI model [1]). The mental world is traditionally described using classical first order logic, extended by respective modal operators for mental entities involved. Lexical units above express multiple meanings (modalities), therefore we can increase the number of modalities to express respective meanings [2]. Taking into account the fact that mental processes develop in time, temporal logics with branching time comes into play [8]. Resultant formalisms are applicable in partial cases of multiagent systems, but there is no generic implementation due to the following:

• Rather weak subset of commonsense laws of mental world is expressible via assertions between modalities;
• Too few theorems are deducible from the axioms for modalities as laws of mental world to describe its phenomena in detail;
• Attempts to build sound and complete (in logical sense) formalizations of mental world are associated with the drop of the expressiveness of resultant language: only a subset of observed mental states can be reproduced;
• Representing mental entities as independent modalities moves the modal logic-based approach away from the natural language, which is capable of merging the multiple cohesive meanings in a single lexical unit for mental entity;
• Implementation of reasoning as a first-order theorem proving is inefficient; also, it seems to be hard to directly take advantage of the practical limitation on the complexity of mental formulas.
• First-order logic (particularly, modal and lambda calculi) is oriented to handle certain phenomena of natural language such as quantification and especially language syntax-semantic connections (e.g. Montague grammars). At the same time, these calculi (furthermore, their model theories) are not well adjusted to the peculiarities of ambiguity in mental natural language expressions.

Analyzing these limitations, one may come to conclusions that the mental world is quite different from physical world in terms of how the reasoning is organized.

We desire applications in every field, where human intellectual activity is subject to modeling. Education and mental treatment are worth special mentioning [4]. The next
steps to overcome the limitations above towards the practical implementation of reasoning about mental world are:

1) Using simulation of decision making rather than representing it as a pure deduction [3];

2) **Describing the multiagent interaction, ascend from the level of atomic actions of agents to the level of behaviors;**

3) Limiting the complexity of mental formulas;

4) Following closer the natural language in describing the mental world, using a wide range of entities;

5) Taking advantage of approximation machinery. We express an arbitrary mental entity through the basis knowledge-belief-intention (informing, deceiving, pretending, reconciling etc., [4]);

6) Using emotions as characterizations (fluents) assigned to mental states [5].

Here we focus on the second step and introduce the library of behaviors. Below we evaluate how the formalized forms of behavior cover the set of scenarios of multiagent interaction in a real-world domain. We will build a domain-independent set of behaviors, which can be applied to an arbitrary problem involving multiagent interaction.

## 2 Choosing a Most Favorable Behavior

NL_MAMS inputs NL descriptions of initial and current states of contacting agents (particularly, in the form of messages from them) and outputs its own strategy and predicted strategy of the other conflicting or collaborating agents. The strategies are the deterministic scenarios of intelligent behavior of agents, capable of analyzing and predicting the consequences of mental and physical actions of themselves and others [4].

To choose the best action, each agent considers each action he can possibly perform at the current step. For each such action, the agent assumes it has committed it and analyzes the consequences. They include the actions of other agents and resultant states, some of which may be unfavorable for this agent. The agent either decides to perform the action delivering the least unwanted state or action by another agent, or to do nothing. If there are multiple possible actions which do not lead, in the agent belief, to unwanted consequences, this agent either chooses the preferred action, if there is an explicit preference relation, or the action, whose conjectures using the beliefs concerning the other agents in the least degree. If our agent wants a particular action of the other agent, it can either do nothing or perform an action that is neither explicitly wanted nor unwanted, but is followed by the desired action of that other agent, in accordance to our agent’s belief. In particular, it can ask that agent directly or suggest a third agent to encourage the second one to perform that action.

The high-level algorithm for the action choice is presented below as a logic program clause. Comments to the code (currier font) start with ‘%’.

The predicate availableAction (bold above) will be the focus of our considerations of behavior forms in the following section.
chooseAction(Agent, ChosenActions, History):-
% generates the set of available actions and chose those leading to acceptable states
findall( PossibleAction, ( % finds all objects satisfying conditions below
availableAction(Agent, PossibleAction, History),
% choosing (forming) a behavior
assume(PossibleAction), % assume that the selected action is performed
involveKnowledgeOfOthers(Agent),
% substitutes own knowledge by own knowledge of others’ knowledge
assumeOtherAgents(Agent, OthersActions),
% Similar assumption concerning others’ actions. They are obtained based on the
% acceptable states of the others from the viewpoint of the given agent.
% Here the agent thinks for its opponents what would they do to achieve their goals
acceptableState(Agent),
% To verify that the state to be achieved by is acceptable (not worse than the current state)
clean_assume(PossibleAction), % cleans the assumptions
), AccumulatedPossibleActions),
chooseBestActions(AccumulatedPossibleActions, ChosenActions).
% choosing the best action in accordance to the preference relation on the set of accessible states

3 The Library of Available Behaviors

We have discovered that the totality of mental entities can be expressed in the basis want-know-believe [3]. The clauses for pre-conditions of behaviors (as aggregated mental actions) we define in this section indeed contain these predicates (enumerated in Table 1). The head of each clause is the predicate generateAction(Agent, GeneratedAction, History) which returns the second argument. The complete set of clauses for the library of behaviors is available at www.dcs.bbk.ac.uk/~galitsky/NL_Mams/Behlibrary.htm.

We present the clauses for behaviours in details to introduce a flavour of how to define mental entities in the basis of want-know-believe in a procedural manner, based on the current mental state and the history of mental actions History. Note that we take a strong advantage of metaprogramming to express a wider set of meanings and to achieve a higher level of abstraction. For brevity we merge know and believe in the clauses below most of times.

We start with the clause to generate a physical action that is included in agent's intention. It may be a potential action of another agent, which is selected by a given agent. The clause finds a subformula of intention so that its argument ranges over physical objects.
generateAction(Agent, ActionFull, _):-want(Agent, StateORAction),
expand(StateORAction, SOAs) !], % getting a list of all subterms of a term
member(PhysFull, SOAs), % finding a term for physical action
argrep(PhysFull, 1, Agent, ActionFull).
% substitution of itself instead of another agent into selected action
% The clause forms an own action for an agent that causes desired state of another agent
\texttt{generateAction(Agent, MyAction, _):-}
\texttt{want(Agent, State), \% search for a clause which is a subject of knowledge}
\texttt{(clause(State, MyAction); know(Agent, (State:- MyAction)), State=..[_], OthAg\=Agent, MyAction=..[_], Agent=..[_]),}
\texttt{not want(Agent, not MyAction), \% it is not an unwanted action}
\texttt{not know(Agent, not MyAction). \% this action is not known as impossible}

We proceed to the generic clause for \texttt{inform}
\texttt{generateAction(Agent, inform(Agent, Addressee, Smth), _):-}
\texttt{know(Agent, want(Addressee, know(Addressee, Smth)));}
\texttt{want(Agent, know(Addressee, Smth)).}

If an agent being informed, it may add a belief (reaction to being informed)
\texttt{generateAction(Agent, assert(believe(Agent, Smth)), History):-}
\texttt{prevStep(inform(AgentInform, Agent, Fact), History), not believe(Agent, not Smth), not know(Agent, not Smth), not believe(Agent, not trust(Agent, Smth)).}

The following clause specifies how an agent forms mistrust when it discovers that it is being
informed a lie
\texttt{generateAction(Agent, believe(Agent, not trust(Agent, Smth)), History):-prevStep(inform(AgentInform, Agent, Fact), History), member(FactOp, History), opposite(Fact, FactOp).}

The clause clarifying when to ask with intention to gain knowledge (possibly believing that someone knows) looks like
\texttt{generateAction(Agent, ask(Agent, InfAgent, Smth), _):-}
\texttt{((want(Agent, know(Agent, Smth)), \% Smth is instantiated knowledge}
\texttt{believe(Agent,know(InfAgent,Smth)), nonvar(Smth)) ;}
\texttt{(want(Agent, know(Agent, Smth)), nonvar(Smth)) ,}
\texttt{ifthen(var(InfAgent),(agents(Ags),member(InfAgent,Ags)))).}

% substitutes an agent if currently unknown

The clause introduces the conditions of when to answer: history includes asking, an agent
answers if it knows and/or wants addressee to know; believe/know options are considered
\texttt{generateAction(Agent, ActionFull, History):-prevStep(ask(AgentAsk, Agent, Smth), History), (}
\texttt{(believe(Agent, Smth), want(Agent, know(AgentAsk, Smth))},
\texttt{ActionFull= answer(Agent, AgentAsk, believe(Agent,Smth))) ;}
\texttt{(believe(Agent, not Smth), want(Agent, know(AgentAsk, Smth))},
\texttt{ActionFull=answer(Agent, AgentAsk, believe(Agent,not Smth)) ;}
\texttt{(know(Agent, SmthRelevant), expand(SmthRelevant, SmthRE), mem-
\texttt{ber(Smth, SmthRE), \% sharing a part (subterm) of knowledge}
\texttt{ActionFull= answer(Agent, AgentAsk, SmthRelevant) )}).

We proceed to the clause for generation of a suggestion. If an agent wants someone’s ac-
tion and does not have a belief that this agent does not want to perform that action then that
action is suggested.
\texttt{generateAction(Agent, suggest(Agent, OtherAg, OtherAgAction), History):-}
\texttt{want(Agent, OtherAgAction),OtherAgAction=..[Action,OtherAg|_], not believe(Agent, not want(OtherAg, OtherAgAction))}.

% substitutes an agent if currently unknown

If an agent is being suggested something, the following clause specify the conditions to follow these suggestions
\texttt{generateAction(Agent, Smth, History):-}
\texttt{prevStep(suggest(AgentAsk, Agent, Smth), History),}
\texttt{((Smth=..[Action, Agent|_]);
\texttt{((Smth=(not NSmth)), NSmth=..[Action, Agent|_])).}
We evaluate the tool trying to reproduce a number of scenarios of multiagent interaction. If for a particular target scenario we cannot adjust the initial mental state such that NL_MAMS generated the scenario which is similar to the target one. For the training, we collected the scenarios from various domains over duration of NL_MAMS project (over 7 years). If a given scenario required adding a new form of behavior, respective clause has been added. Sometimes other units of NL_MAMS were extended to accommodate the training scenarios.

The evaluation of the accumulated behavior library and overall system performance is conducted in the domains of customer complaints. Complaints are the scenarios of interaction between a complainant and company representatives; these conflicting scenarios are mostly occurring in a mental space. We have observed that the trained behaviors adequately cover the evaluation domain (Table 1). All clauses for behavior forms trained in the domain of randomly accumulated scenarios were deployed in a complaint domain. Conversely, to explain the rational multiagent behavior of complaint agents, it is sufficient to use accumulated clauses for behaviors.

Clearly, formal descriptions of the behavior of complainants and their opponents in more detail would benefit from additional complaint-specific behavior patterns. However, we revealed that increasing the complexity of the formal descriptions of textual scenarios does not make them more consistent, because the majority of intermediate mental states are not explicitly mentioned. We conclude that the formed library of behaviors is sufficient to provide an adequate (most consistent) description of multiagent interactions.

Table 1. Evaluation of the behavioral coverage of scenarios. One scenario contains on average 3.2 forms of behavior in the training dataset and 4.3 forms of behavior in the evaluation dataset.
Note that our evaluation by no means intended to predict the scenarios; instead, we try to include all necessary information in the initial mental state so that the scenario is generated as a respective sequence. The problem of prediction the consecutive mental states in the conditions of lack of information is posed differently [5] and requires machine learning and reasoning about actions [7] components in addition to NL_MAMS.

4 Conclusions

In this paper we demonstrated that reasoning about mental world can be implemented via exhaustive search through the possible behaviors, evaluating achieved mental states. Generic representation of reasoning about mental world may be viewed as augmentation of logical axioms to perform reasoning about a particular domain (represented by means of applied axioms). Therefore we follow along the line of classical axiomatic method stating that the same set of logical axioms is sufficient to perform reasoning in an arbitrary domain [6]. This basic principle of axiomatic method has been verified formalizing, in particular, algebraic and topological structures. In this study we have verified that the set of behaviors observed in one domain can be applied in an intact form to another domain with different physical axioms to produce adequate multiagent scenarios.

In this paper we have discussed the limitations of using modal logic for reasoning about mental world. Clearly, a lot of observations about the multiagent behavior can be deduced from the axioms; however the set of theorems does not constitute a basis to enumerate a set of consecutive mental states. We conclude for the generic implementation of reasoning simulation is required, which is implemented as en exhaustive search in the space of possible behaviors. It has been observed in this study that the simulation for realistic mental states for a few agents is not computationally intensive.

References

1. Bratman, M.E.: Intention, plans and practical reason. Harvard University Press: Cambridge MA (1987).
2. Fagin,R., Halpern,J.Y., Moses,Y., Vardi,M.Y. Reasoning about knowledge MIT Press, Cambridge, MA, London, England (1995).
3. Galitsky, B.: On the training of mental reasoning: searching the works of literature. FLAIRS (2002), Pensacola Beach, FL.
4. Galitsky, B.: Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Australia (2003).
5. Galitsky, B. and Tumarkina, I.: Justification of Customer Complaints using Emotional States and Mental Actions FLAIRS (2004), Miami, FL.
6. Peirce, C. Deduction, Induction, Hypothesis. in Collected Papers of Charles Sanders C. Peirce. Volume II. Cambridge, MA The Belknap press of Harvard University press (1965).
7. Shanahan, M. Solving the frame problem. MIT Press (1997).
8. Wooldridge, M. Reasoning about Rational Agents. The MIT Press Cambridge MA (2000).
Handling Default Rules by Autistic Reasoning

Don Peterson¹ and Boris Galitsky²

¹ Institute of Education, University of London,
20 Bedford Way London, WC1H 0AL, UK
² School of Computer Science and Information Systems,
Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK
d.peterson@ioe.ac.uk, galitsky@dcs.bbk.ac.uk

Abstract. We employ the formalism of default logic to model the phenomena of autistic reasoning. Our main finding is that while people with autism may be able to process single default rules, they have a characteristic difficulty in cases where two default rules conflict. Even though default reasoning was intended to simulate the reasoning of typical human subjects, it turns out that following the operational semantics of default reasoning in a literal way leads to the peculiarities of autistic behavior observed in the literature.

1 Introduction

The syndrome of autism was first identified in the 1940’s and exhibits a variety of phenomena: some of an interpersonal and some of a pragmatic character. One problem confronting the understanding of the syndrome is that of conceptualization: although the practitioner becomes accustomed to recognizing and responding to the various tendencies exhibited in the syndrome, it can nevertheless be difficult adequately to describe them. Various theories attempt to provide conceptualizations of the syndrome: the best known being the ‘theory of mind’ account ([2], the ‘central coherence’ account [8], and the ‘executive function’ account [10]. These theories all, however, have well-known difficulties, and there is a need for further contribution to the conceptualization of the syndrome or parts of it.

In this paper, we draw on a branch of logic in order to articulate the character of some major subsets of the phenomena belonging to the syndrome. This branch is the logic of default practical reasoning, that is, of reasoning which is practical in the sense that its conclusions specify actions, and default in the sense that additional context can cause a conclusion to be modified or withdrawn. This allows us to characterize some phenomena of autism in a fresh and precise way and suggests new lines of empirical experimentation. An advantage of this approach is that it allows us to benefit from the rich vocabulary of concepts, notations and distinctions which has been developed during the history of logic. We describe the peculiarities of autistic reasoning in terms of posing the problems a logician needs to solve while applying particular formalisms to implement the decision-making.

Default reasoning is intended as a model of real-world commonsense reasoning in cases which include typical and non-typical features. A default rule states that a
situation should be considered as typical and an action should be chosen accordingly unless the typicality assumption is inconsistent. We observe that autistic intelligence is capable of operating with stand-alone default rules in a correct manner most of times.

When there is a system of conflicting default rules, the formal treatment (operational semantics) has been developed so that multiple valid actions can be chosen in a given situation, depending on the order in which the default rules are applied. All such actions are formally accepted in such a situation, and the default logic approach does not provide means for preference of some of these actions over the other ones. Analyzing the behaviour of people with autism, we will observe that unlike the controls, children with autism lack the capability to choose the more appropriate action instead of a less appropriate. In this respect we will see that the model of default reasoning suits autistic subjects better than controls.

This study branches out of our earlier studies of counterfactual reasoning [9] and reasoning about mental states of autistic patients [7], as well as extending the lines of our existing rehabilitation strategies [6].

2 Characterizing Autistic Reasoning

In this study we argue that the inability to use default rules properly leads to certain phenomena of autistic reasoning identified in existing experimental studies:

1. Non-toleration of novelty of any sort;
2. Incapability to change plan online when necessary;
3. Easy deviation from a reasoning context, caused by an insignificant detail;
4. Lack of capability to distinguish more important from less important features for a given situation;
5. Inability to properly perceive the level of generality of a feature appropriate for a given situation.

Note that these peculiarities of reasoning can be distinguished from reasoning about mental attitudes, which are usually corrupted in a higher degree in case of autism [2].

Our approaches considers the mechanisms of how typical reasoning is performed from the computational prospective, and then compares these mechanisms with the limitations of experimentally observed autistic reasoning. We take advantage of significant achievements of logical artificial intelligence in modelling human reasoning and understanding the mechanisms of solving the problems suggested to autistic and controls during the experiments. This computational approach therefore complements the findings of psychological experimentation in the study of autism.

Default reasoning is a particular machinery intended to simulate how human reasoning handles typical and atypical features and situations. Apart from reasoning about mental attitudes which is essential in presenting autism, we apply default reasoning to conceptualise a wide range of phenomena of autistic reasoning, taking advantage of the experience of computer implementation of default reasoning.
Peculiarities of autistic reasoning can then be matched against the known possibilities of malfunctioning of artificial default reasoning systems.

In the context of artificial intelligence, the phenomena of autistic reasoning are of particular interest, since they help us to locate the actual significance of formal models of default reasoning. At the same time, we expect this study to shed light on how autistic reasoning may be improved by default reasoning-based rehabilitation techniques.

3 Handling a Single Default Rule by Autistic Reasoning

An abstract default logic distinguishes between two kinds of knowledge: the usual formulas of predicate logic (axioms, facts) and “rules of thumb” (defaults) [1]. Corrupted reasoning may handle improperly either kind of knowledge, and we pose the question which kind may function improperly in autistic reasoning. Moreover, we consider the possibility that an improper interaction between the facts and rules of thumb may be a cause for corrupted reasoning.

Default theory [4,5] includes a set of facts which represent certain, but usually incomplete, information about the world; and a set of defaults which cause plausible but not necessarily true conclusions (for example, because of the lack of a world knowledge or a particular situation-specific knowledge). In the course of routine thinking of human and automatic agents some of these conclusions have to be revised when additional context information becomes available.

Let us consider the traditional example quoted in the literature on nonmonotonic reasoning:

\[
\begin{align*}
\text{bird}(X) & : \text{fly}(X) \\
\hline
\text{fly}(X)
\end{align*}
\]

One reads it as *If X is a bird and it is consistent to assume that X flies, then conclude that X flies.* In the real life, if one sees a bird, she assumes that it flies as long as no exceptions can be observed.

\[
\begin{align*}
\text{fly}(X)& : \text{not penguin}(X). \quad \text{fly}(X) & : \text{not sick}(X). \quad \text{fly}(X)& : \text{not just_born}(X). \ldots \\
\end{align*}
\]

Exceptions are the potentially extensive list of clauses implying that X does *not fly.* It would be inefficient to start reasoning based on exceptions; it should be first assumed that there are no exceptions, then verified that this is true and then proceed to the consequent of a default rule.

A penguin (the bird which does not fly) is a *novelty* (it is atypical). Conventional reasoning first assumes that there are no novelties (there is no exception) and then performs the reasoning step, concluding that X flies. If this assumption is wrong (e.g. X-novelty is taking place) then the rule is inapplicable for penguins and it cannot be deduced that X flies. It is quite hard for autistic reasoning to update this kind of belief because it handles typical and atypical situations in the same manner, unlike the default rule machinery suggests. It is quite computationally expensive to handle typical and atypical situations similarly, because a typical situation is compact and most likely to occur, and an atypical situation comprises an extensive set of cases (clauses) each of which is unlikely to occur.
Let us now view this example from the perspectives of five phenomena mentioned above:

Unlike normal subjects, and similar to software systems, autistic subjects can hardly tolerate the `Additional_features_of_environment_do_not_change_routine` when they have a `Usual_intention` to `Follow_usual_routine`:

\[
\text{Usual_intention} : \text{Additional_features_of_environment_do_not_change_routine}
\]

\[
\text{Follow_usual_routine}
\]

This default rule schema is read as follows: when there is a `Usual_intention`, and the assumption that `Additional_features_of_environment_do_not_change_routine` is consistent, then it is OK to `Follow_usual_routine`. There should be clauses specifying the situations where this assumption fails:

\[
\text{Additional_features_of_environment_do_not_change_routine:- not ( alarm(fire) } \lor \text{ desire(DoSomethingElse)} \lor \ldots ).
\]

This clause (assumption) fails because of either external reasons or internal ones, and the list of potential reasons is rather long.

| A child knows that birds fly. The child sees observes that penguins do not fly | Child updates the list of exceptions for not property flies | Child adds new rule that penguins do not fly |
|------------------------------------------------------------------------------|----------------------------------------------------------|--------------------------------------------|
| The flying default rules stays intact.                                      | It is necessary to update the existing rule of flying and all the rest of affected rules |
| The process of accepting new exceptions is not computationally expensive     | This process takes substantial computational efforts and, therefore, is quite undesirable and overloading. |
| Observing a novelty and remembering exceptions is a routine activity         | Observing a novelty is stressful                          |

A good example here is that the autistic child runs into tremendous problems under deviation in an external environment which typical cognition would consider to be insignificant.

We proceed to the phenomenon of Incapability to change a plan online when necessary. A characteristic example is that of an autistic child who does not walk around a puddle which is blocking her customary route to school, but rather walks through it and gets wet as a result. This happens not because the autistic child does not know that she would get wet stepping through a puddle, but because the underlying reasoning for puddle avoidance is not integrated into the process of reasoning. Let us consider the reasoning steps a default system needs to come through.

Initial plan to follow a certain path is subject to application (verification) by the following default rule:

\[
\text{need(Child, cross(Child, Area))} : \text{normal(Area)}
\]
Here we consider a general case of an arbitrary area to pass by, Area = puddle in our example above. The rule sounds as follows: “If it is necessary to go across an area, and it is consistent to assume that it is normal (there is nothing abnormal there, including water, mud, danger etc.) then go ahead and do it). A control individual would apply the default rule and associated clause above to choose her action, if the Area is normal. Otherwise, the companion default rule below is to be applied and alternative AreaNearBy is chosen.

need(Child, cross(Child, Area)), abnormal(Area) :- normal(AreaNearBy)  
cross(Child, AreaNearBy)  

Note that formally one needs a similar default rule for the case something is wrong with AreaNearBy: abnormal(AreaNearBy). A control individual ignores it to make a decision with reasonable time and efforts; on the contrary, autistic child keeps applying the default rules, finds herself in a loop, gives up and goes across the puddle.

In other words, autistic reasoning literally propagates through the totality of relevant default rules and run into the memory/operations overflow whereas a normal human reasoning stops after the first or second rule is applied.

What are the peculiarities of how autistic children apply a newly acquired rule? First of all, they do their best in applying it, however, they follow it literally. Let us consider the following example:

An autistic girl was advised by her parents not to speak with strangers in the street. On one occasion a policeman approached the girl and started asking questions, but was ignored by her. In spite of his multiple attempts to encourage the girl to communicate, they failed and he became upset.

After the parents were told about the incident they suggested that the girl should not have treated policemen as a stranger. They also confirmed that the girl knew who policemen were. The girl required that she needed the new explicit rule overwriting the initial one that a policeman was not a typical stranger and should have been treated differently.

On the basis of the analysis presented here, this anecdote could be given the following interpretation.

1. The subject is doing her best to follow the rule, and readily accepts new rules.
2. The girl did know that the approaching man was a policeman, but she did not know him as a person, therefore she categorised him as a stranger in the context of the behavioural rule.
3. In this situation the girl was familiar with who policemen are, as she knew that policemen should not be ignored.
4. However, she was not able to handle a policeman as an exception in the rule for stranger.
5. If she had had the explicit rule for how to respond to strangers who are policemen then she would have followed it.

We conjecture that the girl had sufficient knowledge of the subject and was capable of applying the rules, taken separately. What she was not able of doing was to resolve a conflict between considering the same individual as a stranger and as a policemen in the context of decision whether to communicate or to ignore.

\[
\text{in\_street}(me) :- \text{stranger}(Person) \\
\text{not talk}(me, Person)
\]

Usually, strangers do not fall into a special category; however, exceptions are possible:

\[
\text{stranger}(Person):-\text{not (policeman(Person) \lor rescue(Person) \lor military(Person \lor \ldots)}.
\]

Indeed, the girl is likely capable of identifying the categories of persons above, but not in the context of a stranger rule, which is indeed an opposing rule to the one for handling exceptions:

\[
\text{talk}(me, Person):-\text{not (Person)}.
\]

If the parent would incorporate the rule above into the default rule explicitly, then it is likely that the girl would treat the policemen properly.

4 Conclusion

This paper has drawn on a branch of logic in order to provide a framework for the understanding of the elusive phenomena of autism. Our thesis is that difficulty arises in autism specifically in those situations where two default rules conflict, and this provides a relatively precise tool for understanding some of the phenomena of autism. This will be a basis for further work which will investigate the following conjectures:

1. Non-toleration of novelty of any sort, because it requires update of the whole commonsense knowledge, since it is not adequately divided into typical and atypical cases, norms and exceptions;
2. Incapability to change plan online when necessary, because it requires substantial computational efforts to exhaustively search the space of all possibilities;
3. Easy deviation from a reasoning context, caused by an insignificant detail, because there is a high number of issues to address at each reasoning step; each such issue is seemed to be plausible;
4. Lack of capability to distinguish more important from less important features for given situation, because feature importance is mainly measured in the context of being a justification of default rule.
5. Inability to properly perceive the level of generality of features appropriate for a given situation is due to the problem of estimating which generality of a given feature is most typical, and which is less typical to be applied as a justification of a default rule.
We observed that loss of reasoning efficiency due to improper use of default rules leads to a wide range of reasoning problems reflected in behaviours characteristic of autistic subjects.

Finally, we mention the methodology for experimental testing of our hypothesis that inability of applying default rules leads to a series of significant deviations of reasoning capabilities in autism. A typical situation where a default rule is naturally applied arises while understanding an ambiguous sentence (command), where one meaning is typical and another is atypical. Conducting a conversation with an autistic individual, an experimenter may ask ambiguous questions or give ambiguous commands, and track the reactions of the patient. Five phenomena of this study can be addressed in such a scenario, and observed in terms of how handling ambiguity via default rules influences these phenomena. We have conducted preliminary experiments along this line, and more detailed experimental observations of this sort is the subject of our further study.

References

1. Antoniou, G. Nonmonotonic reasoning. MIT Press Cambridge, MA London England (1997).
2. Baron-Cohen, S. Mindblindness: An Essay on Autism and Theory of Mind. Cambridge, Massachusetts: MIT Press (1995).
3. Bratman, M.E.: Intention, plans and practical reason. Harvard University Press: Cambridge MA (1987).
4. Brewka, G., Dix, J., Konolige, K. (1995). Nonmonotonic reasoning: an overview. CSLI Lecture Notes 73.
5. Bochman, A. (2001). A logical theory of Nonmonotonic Inference and Belief Change. Springer Verlag.
6. Galitsky, B.: Using mental simulator for emotional rehabilitation of autistic patients. FLAIRS (2003), St. Augustine, FL.
7. Galitsky, B.: Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Australia (2003).
8. Happe F.G. Studying weak central coherence at low levels: children with autism do not succumb to visual illusions. A research note. J. Child Psychol Psychiatry. Oct;37(7):873-7 (1996).
9. Peterson, D.M and Bowler, D.M. Counterfactual reasoning and false belief understanding in children with autism. Autism: The International Journal of Research and Practice 4 (4): 391-405. (2000).
10. Russel, J., Ed. Autism as an Executive Disorder Oxford Univ. Press (1997).
An Ontology-Driven Approach
to Metadata Design in the Mining
of Software Process Events

Gabriele Gianini and Ernesto Damiani
University of Milan, Department of Information Technologies,
via Bramante 65, Crema (CR) 26013, Italy
{gianini,damiani}@dti.unimi.it

Abstract. In this paper we report about an ongoing experience of meta-
data design in the context of the quantitative modeling of the software
development process. We use an ontology driven metadata design pre-
scription, where the metadata structure is determined by two strictly
interleaved ontologies: a software development process ontology and a
knowledge discovery process ontology.

1 Introduction

Experience has shown that quantifying the software process operation can greatly
improve process insight and bring long term benefits, through the identification
of interesting process patterns and dependencies, best practice examples or even
through process model discovery [Coo97]. However, the quantitative empirical
study of the software process needs often to follow a complex Knowledge Dis-
covery (KD) path, involving heterogeneous data taking, data preprocessing and
data mining tools; for the KD process to be effective, all those tools must be
integrated in a framework which has to be as flexible as possible, so as to ac-
commodate the continuously evolving needs of data analysis, where each finding
or each result assessment might trigger new questions.

Metadata can play a cornerstone role in such a KD framework, allowing the
management of the high volume of data generated by automatic event collection
tools under heterogeneous contexts, and supporting the tracking of the manipu-
lations undergone by the data sets during the preprocessing and analysis path.

Here we report about an ontology-driven metadata design paradigm, where
the metadata structure is determined by two strictly interleaved ontologies: a
software development process ontology and a KD process ontology. Of the latter
we adopt, and informally outline here, a formulation based upon the measure-
ment process paradigm and on the inferential KD paradigm; as of the former,
the choice of a particular process model will have to vary from case to case. We
will point here to an example employed within the study of Agile Methodologies,
however, since the usage of a similar layout is, in principle, not restricted to a
specific case, we will outline motivations, problem structure and approach from
the broader perspective of the KD from data of a general Software Development Process (SDP).

The paper is structured as follows: after reviewing some general motivations for the employment of a unifying metadata design, we start our exposition by looking at the SDP as at the object under investigation and outline a standard model for a corresponding KD process. The informal description of the KD process and of its rationale aims to account for its intrinsically complex structure. In the meanwhile we point at the role that metadata have to play at each step, in order to review the motivations supporting the metadata key role in the process. Next the prescription for the design of the metadata structure, which uses both the software process model and the KD process model, is given, along with some illustrative examples.

2 Metadata Roles in the Knowledge Discovery Process

Metadata are data about data. They shape data into useful information, from which knowledge can then be extracted by induction (generating intension from extension), and facilitate the integration over heterogeneously generated and manipulated information. Within the process of KD from SDP data, they represent the gears which transfer the information from one phase to the other, preserving its meaning and keeping track of the previous elaboration steps.

Traditionally, process logs and metrics were collected manually. However, manual collection can represent a heavy burden for software developers and process managers and is prone to error, in a context where data quality is the key for obtaining added value. Automated data collection tools can help solving this problem, meanwhile providing large data samples.

Furthermore, automated data collection is the optimal way for reaching the high statistics needed to support the findings of the KD analysis [Bri95]. In fact, in order for the empirical measures to be meaningful, so as to lead to reliable conclusions, one must collect a data sample with sufficient statistical power, namely by gathering data on a high number of cases, where all the features under study are conveniently represented and all the main factors influencing the measure carefully recorded [Mil97].

For the above reasons, in our setting, the metadata are mainly data about automatically collected events, and about the elaboration phases they have undergone; their main purpose is to inform about the context in which a given data set was produced and manipulated, allowing different and heterogeneous tools to take over the KD process where the previous tools have left, without loosing the memory of the previous manipulations and choices.

3 The Knowledge Discovery Process

There are several ways of structuring a KD process (see for instance [Mor95]). The structure for a KD process we use is made of four main phases: a preliminary
phase, a data taking and reconstruction phase, a data mining phase and a result assessment phase, after which the knowledge discovery process can be iterated. Each phase can be broken down into several sub phases.

3.1 The Preliminary Phase

The preliminary phase includes an inception and a planning sub phase.

In the inception sub phase one has to identify the questions to which the knowledge discovery process is going to try to find an answer. The most common objective of a KD from SDP data is the quest for interesting process patterns and dependencies and best practice examples, another possible goal is that of measuring the conformance of a real process with a given process model [Coo99], a more long reaching goal could be that of process model discovery [Coo98].

In the inception phase one has also to to survey the status of the art in the area, relatively to the investigated issues, so as to to acquire, on one hand, SDP domain knowledge, and on the other hand, KD process knowledge; after this, one can try to adopt or produce a more or less formal model of the investigated domain and of the knowledge discovery process, which are going to be the bases for the subsequent activities. In our setting the choice of the two formal models is a key activity, since the metadata design will rely upon both the software development process ontology and the KD process ontology. Even in the case where the goal of the overall process is that of SDP model discovery, a preliminary minimal formal model will have to be adopted as a starting point.

All the previous activities will help in formally stating the goals of the overall process, and in better defining the data mining objectives, both in terms of required statistics for the analysis and in terms control samples’ features.

In the next sub phase, the planning sub phase, one has to make a survey of the available resources such as the data collection tools and the data analysis tools, then one has to take decisions over which of the existing tools have to be adopted and what tools have instead to be developed, this in turn will involve deciding the algorithms most suitable for the proposed analysis. The outcome of this sub phase is a detailed action plan stating where, how and by whom the empirical data have to be collected, possibly transported, stored, prepared and analyzed. The same holds for the production, storage and preparation of the control data, to be used later for the assessment.

3.2 The Data Taking and Reconstruction Phase

The data taking and reconstruction phase can be seen and modeled as a measure process affected by some inefficiency and distortion, due to the intrinsic incompleteness and imperfection of the data collection tools.

In a typical measure process of this kind you have some interesting phenomenon mingled with some uninteresting phenomenon; you cannot help measuring them together and you do it through some data taking device and process. The output of the data taking will therefore be given by an interesting phenomenon mixed with some uninteresting phenomena (part of which will re-
semble to the interesting one). From the recordings of both phenomena, some un-
observed part, which cannot be seen by the measuring apparatus, will be missing.
Finally what does not go lost will undergo, before going into the records, some
sort of distortion due to the intrinsic imperfection of the measuring apparatus.

The first goal of the knowledge discovery process is that of trying to reverse
the effects of inefficiencies and distortions, so as to recover and restore as much
information as possible over the target phenomenon. This is done by combining
the data with the knowledge of the data collection process to attack the problems
of limited acceptance, distortion and contamination. Only at the end the data
will be available for the second phase of the knowledge discovery process, the
actual mining phase.

Consequential to this paradigm is the break down of the data taking and
reconstruction phase into the following sub phases: metadata design (we come
back to this activity in the next section), collection tool preparation, data taking,
data quality control and selection, data exploration and reconstruction, with
feature extraction.

The data collection tool preparation might include the tool development and
tuning. The actual data collection sub phase is one of the most critical parts
of the process, here the utmost care must be devoted to data quality and to
careful documentation of the gathered samples. The metadata here must allow
the recording of all the context information related to the project where the data
have been gathered.

The next sub phase, data quality control, is aimed to assess the general
quality of the collected data in terms of completeness and conformity to the
expected standards. Knowledge of the data taking process can help identifying
incoherences present in the data and in selecting the data which can actually be
used by the next phases.

Data exploration, at this point of the process, serves not only a documen-
tation purpose, but can also be used to check for inconsistencies against the
domain knowledge, which can result in the decision to discard or to amend the
inconsistent data. Amending the data and enriching them with information in-
ferred using domain knowledge (feature extraction) is the responsibility of the
reconstruction sub phase, whose output will be data usable for the analysis. Ev-
ery choice about the data transformation and enrichment must be captured here
by suitable metadata structures.

3.3 The Data Mining Phase

The rationale behind data mining will depend on the mining objectives, however
the most commonly adopted paradigm is that of an inferential statistics learning
process. It is not by chance that, although represented as separated phases, the
data mining phase and the result assessment phase are strictly intermixed and
are often repeated one after the other until satisfying results are attained. This
reflects the iterative and adaptive character of the learning process.

The overall mining phase can be broken down into the following sub phases:
data subsample creation, data reduction, data format conversion and, finally,
data mining. In this phase the reconstructed data first undergo some preparation, aimed to create distinct samples for the distinct questions the analysis will attempt to answer: this selection may be done together with some data reduction, in order to isolate only the features that are relevant to each stream of analysis (feature selection), resulting in the actual analysis samples. Then the data will have to be cast into the formats required by each analysis application. After these operations the actual analysis can finally take place.

The analysis could be performed along several possible lines such as frequent pattern discovery, association rule discovery, process model discovery, or conformance assessment with respect to some preexisting process model.

3.4 The Result Assessment Phase

The assessment phase consists in the preparation of the control samples and into the actual assessment of the analysis results. The control data can be synthetically produced by some simulation models or can be extracted from the collected data. In the latter case, the control samples are usually prepared along with the analysis samples. The assessment usually consists in rerunning the analysis procedures over the control samples and comparing the results. This in turn can result in some hints for a reiteration of the analysis, with some changes in the parameters or in the procedures, or for a reiteration of the whole process starting from the data taking phase.

Metadata, here must allow the backward tracking from the results to the reconstructed data samples they come from and to the test samples used for the assessment.

4 A Prescription for Metadata Design

In this section we address the issue of how to design metadata for the benefit of our knowledge discovery process. Due to the high complexity of the overall process and to the risk of loss of relevant information which could be needed in subsequent process iterations, one has to strictly keep all the information about the process data and about the processing they undergo within the KD path, although this might result in a high volume of data. A key prescription is that of providing each data instance with the pointers linking it with the surrounding nodes in the domain ontology and, as the datum passes through the different elaboration phases, progressively adding the processing outcome and new pointers linking the datum with the undergone processing phase, the agent and tool performing the operation and possibly the set of underlying assumptions and rules used.

4.1 Some Examples

For instance one can use at both levels (investigated domain SDP and KD process) a generic process ontology. In this generic ontology an agent, in a given role,
performs at some given time or within some time interval (and for some purposes, if known and relevant) an action over some object (obtaining the intended result or failing, if relevant).

```xml
<event-record>
  <data-collection-phase>
    <event-data>
      <id>124567890</id>
      <time>20040102134512</time>
      <identity>AA</identity>
      <role>developer</role>
      <project>PP</project>
      <dev-tool>IDE-XY</dev-tool>
      <file>FF</file>
      <event-type>Compile</event-type>
      <result>true</result>
    </event-data>
    <environment-data>
      <source>PROM_01.2</source>
      <time>20040102134513</time>
    </environment-data>
  </data-collection-phase>
  <data-check-phase>
    <rule>RKD-54</rule>
    <result>true</result>
  </data-check-phase>
</event-record>
```

Fig. 1. A portion of an example event record

At the SDP level the data instance could be that of the developer AA successfully performing a compilation action at time within a given IDE over the file FF within the project PP; this event could be tagged with an unique ID. At the KD process level the event would be complemented with the information about the data collection tool (name and version) and the time of the recording, which, for instance for network latency reasons, might take place only after some time.

In a subsequent data quality assurance phase, the same event could be checked for internal consistency against the (KD domain knowledge) rule that no recording time can be earlier than the event occurrence time (this is of little interest, in practice, and is reported here only for sake of clarity): events with such an anomaly would be corrected by setting the recording time stamp at the value of the event occurrence time stamp.
The new event record instance would report all the information present in the original record, plus the indication of the undergone check (through a rule ID) and the positive or negative outcome of the test (if for any reason the original recording time stamp has to be recovered the unique event ID can lead to the original record, in the original data sample). The check against other KD domain knowledge rules and SDP domain knowledge rules could be analogously reported. At this point a sketch of relevant portion of the event record could be similar to the one shown in Fig. 1.

5 Conclusions

We synthetically reported on an ongoing experience over ontology driven metadata design, based on the formal model of a software development process and on the formal model of a corresponding knowledge discovery process. Our approach relies over a simple prescription which, through the relative completeness and consistence of the underlying ontologies, can guarantee data flexibility, tool interoperability and can supports the iterative and adaptive character of the knowledge discovery process.

We use this approach for metadata design in a concrete case of KD from SDP data, within the study of Agile Methodologies, the project MAPS (*Agile Methodologies for Software Production*) [Jan03]; in our case the base ontology for the SDP is described in [Cer03].

References

[Bri95] Briand L., El Emam K. and Morasca S.: On the Application of Measurement Theory in Software Engineering. Empirical Software Engineering: An International Journal, 1 (1), (1996)

[Cer03] Ceravolo P., Damiani E., Marchesi M., Pinna S., Zavattarelli F.: An Ontology-based Process Modeling for XP. Proc.of APSEC 2003, Chiangmai, Thailand

[Coo98] Cook J.E. and Wolf A.L.: Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology, vol. 7, no. 3 (1998) 215-249

[Coo99] Cook J.E. and Wolf A.L.: Software Process Validation: Quantitatively Measuring the Correspondence of a Process to a Model. ACM Transactions on Software Engineering and Methodology, vol. 8, no. 2 (1999) 147-176

[Jan03] Janes A.: Measuring the Effectiveness of Agile Methodologies Using Data Mining, Knowledge Discovery and Information Visualization. Proc.of XP 2003, Genova, Italy, Lecture Notes in Comp.Sci. 2675 (2003) 445-446

[Mil97] Miller J., Daly J., Wood M., Roper M. and Brooks A.: Statistical power and its subcomponents – missing and misunderstood concepts in software engineering empirical research. J.of Information and Software Technology, no.39 (1997) 285-295

[Mor95] Morisio, M.: A methodology to measure the software process. Proc.of the 7th Annual Oregon Workshop on Software Metrics, Silver Falls, OR (1995)
Knowledge Extraction from Semi-structured Data Based on Fuzzy Techniques

Paolo Ceravolo, Maria Cristina Nocerino, and Marco Viviani

Dipartimento di Tecnologie dell’Informazione di Crema, Università degli Studi di Milano
Via Bramante, 65 – 26013 Crema (CR) – Italy
{pceravolo, mnocerino, mviviani}@crema.unimi.it

Abstract. In this work we propose a fuzzy technique to compare XML documents belonging to a semi-structured flow and sharing a common vocabulary of tags. Our approach is based on the idea of representing documents as fuzzy bags and, using a measure of comparison, evaluating structural similarities between them. Then we suggest how to organize the extracted knowledge in a class hierarchy, choosing a technique related to the domain of interest, later to be converted into a user ontology.

1 Introduction

Nowadays much of the existing electronic data lies outside of database management system: data where structure is “non-relational”, sometimes irregular, as in HTML or SGML documents. In particular the subset of the Web formed by XML documents is growing into a large XML data repository. Our knowledge extraction approach is particularly useful in those cases where XML information is available that was built over a wide vocabulary of optional tags; even more so, if the tags are allowed to occur multiple times within documents. In these cases, there may well be a schema, but the fact that every tag is optional make it quite useless.

These situations are frequent on the Web when users are requested to insert many information on a form, many of which are not mandatory. For example the fact that a user fills in certain fields rather than others makes it possible to recognize different types of users, depending on their properties.

In this paper we will show a method that, computing similarity between docs extending to bags a similarity measure defined for fuzzy objects, allow to build a taxonomy as a result of a sort of cluster analysis.

2 Modelling XML Documents

Similarity plays a fundamental role in theories of knowledge and behavior. The theoretical analysis of similarity relations has been dominated by geometric models. These models represent objects as points in some coordinate space such that the observed similarities between objects correspond to the metric distances between the respective points. The assessment of similarity between objects may be better described as a
comparison of features rather than as the computation of metric distance between points [1].

XML documents are composed of a sequence of nested elements (possibly repeated) called tags ($\chi$), each containing another tag, a value, or both of them. In our approach we only take into consideration tag names and not their value (following an approach value-blind). We suppose the presence of a schema, whose elements (the tag names) are taken from a vocabulary ($V$, the universe of tags). The presence of the elements of the vocabulary in the schema is totally facultative. A domain expert associates a membership degree (between 0 and 1) to each element in the vocabulary, depending on the importance of the tag name (in the domain).

We model XML documents as fuzzy bags due to the fact that the same tag can be repeated at different levels and/or with different cardinalities at the same level. We provide a closure process using the following aggregator:

$$\sum_{i=a}^{L} \mu_{a_i} / L!$$

where $a_i$ is the root node, $L$ is the nesting level and $\mu_{a_i}$ the membership value of the element $a_i$ as read from the vocabulary. This function is not a T-norm [2], but it has the advantage that nesting level and parenthood relations are considered in the process of transformation (closure). So after this process the new membership degree expresses XML tags’ presence and nesting level.

![Closure process](image)

Fig. 1. Closure process. Giving a vocabulary $V = \{R/1, a/0.9, b/0.8, c/0.8, d/0.4, e/0.4\}$, applying (1) to the tree representation of a generic XML document $A$.xml or $B$.xml, we obtain the fuzzy bag $A = B = \{R/1, a/0.53, a/0.36, b/0.33, c/0.33, d/0.7, e/0.65\}$

As shown in Fig. 1 a part of the topological information is lost: applying (1) to $A$.xml or $B$.xml we obtain the same result. This could seems a disadvantage but this loss of information makes less computationally expansive the next phase of comparison of documents, leading to meaningful results.

## 3 Comparison of Documents

In order to compare the XML documents modelled as fuzzy bags we have chosen between measures of comparison studied in [3][4]. We privilege measures giving
higher similarity weight to the bags where elements (tags) belonging to the intersection are less nested.

This is motivated by the fact that, if a tag is nearest the root respect to another, it seems reasonable to assume that it has a higher semantic value.

![Diagram of three sets A, B, and C with elements R, a, b, c, l, h, x, w showing similarity based on distance to root.]

**Fig. 2.** A.xml and B.xml are more similar due to the fact that same tags are nearest to the root

Having a data flow, no reference documents are available. So we need to compare all documents each other (using a symmetric measure). Measure that better satisfies these characteristics is a *measure of resemblance*:

\[
S(A, B) = \frac{M(A \cap B)}{M(A \cup B) + M(A^c \cap B) + M(B^c \cap A)}. \quad (2)
\]

Where

\[
M = \sum_{\mu \in \mathcal{X}} \mu(A \cap B) = \sum_{x \in \mathcal{X}} \mu_x(A \cap B),
\]

\[
\mu_{A^c \cap B}(x) = \max(0, \mu_A(x) - \mu_B(x)),
\]

\[
f_{A^c \cap B}(x) = \mu_{A^c \cap B}(x) = \min(\mu_A(x), \mu_B(x)).
\]

This measure has been studied in [3] on fuzzy sets. Our aim is to apply it on fuzzy bags, so as explained in [5] and thanks to the extension principle, union and intersection on fuzzy bags are defined as:

\[
\Omega_{A \cup B}(x) = \min(\Omega_A(x), \Omega_B(x)),
\]

\[
\Omega_{A \cap B}(x) = \max(\Omega_A(x), \Omega_B(x)). \quad (3)
\]

Where \( \Omega_A(x) \) is a fuzzy integer representing the occurrences of an element \( x \) in a fuzzy bag \( A \).

\[
\mu_{\Omega_{A \cup B}}(x) = \sup_{z = \min(t, v)} \min(\mu_{\Omega_{A}(t)}, \mu_{\Omega_{B}(v)}),
\]

\[
\mu_{\Omega_{A \cap B}}(x) = \sup_{z = \min(t, v)} \min(\mu_{\Omega_{A}(t)}, \mu_{\Omega_{B}(v)}). \quad (4)
\]

For all \( z \in \mathbb{N} \).

---

1 See [5] for more details about operations defined on fuzzy bags.
3.1 Example: Sensitivity to Cardinality

|   | student   | 1 | d | university_data | 0.4 | g | number | 0.6 |
|---|-----------|---|---|-----------------|-----|---|--------|-----|
| b | teacher   | 1 | e | course_information | 0.8 | h | name   | 0.4 |
| c | personal_data | 0.8 | f | course | 0.6 | i | surname | 0.4 |

**Fig. 3.** The vocabulary of tags associated to the university domain and their membership degree

![Tree representation of university data](image)

**Fig. 4.** The tree representation of three XML documents (S1.xml, S2.xml, S3.xml): three different students following a different number of courses

The fuzzy bag representation of the documents illustrated in Fig. 4 is shown below:

- **S1** = \{1/0, 1/1}*a, \{1/0, 0.9/1}*c, \{1/0, 0.9/1}*d, \{1/0, 0.43/1}*h, \{1/0, 0.43/1}*i, \{1/0, 0.5/1}*g, \{1/0, 0.57/1}*e, \{1/0, 0.3/1}*f\}
- **S2** = \{1/0, 1/1}*a, \{1/0, 0.9/1}*c, \{1/0, 0.9/1}*d, \{1/0, 0.43/1}*h, \{1/0, 0.43/1}*i, \{1/0, 0.5/1}*g, \{1/0, 0.57/3}*e, \{1/0, 0.3/3}*f\}
- **S3** = \{1/0, 1/1}*a, \{1/0, 0.9/1}*c, \{1/0, 0.9/1}*d, \{1/0, 0.43/1}*h, \{1/0, 0.43/1}*i, \{1/0, 0.5/1}*g, \{1/0, 0.57/4}*e, \{1/0, 0.3/4}*f\}

\[S1 \cap S2 = \{1/1, 0.9/3, 0.57/4, 0.5/5, 0.43/7, 0.3/8\} \equiv 5.03\]
\[S1 \cup S2 = \{1/1, 0.9/3, 0.57/6, 0.5/7, 0.43/9, 0.3/12\} \equiv 6.77\]
\[(S1 \cap S2) / (S1 \cup S2) = 5.03/6.77 = 0.74\]
S1 ∩ S3 = {1/1, 0.9/3, 0.57/7, 0.5/8, 0.43/7, 0.3/8} ≡ 5.03
S1 ∪ S3 = {1/1, 0.9/3, 0.57/7, 0.5/8, 0.43/10, 0.3/1} ≡ 7.64
(S1 ∩ S3) / (S1 ∪ S3) = 5.03/7.64 = 0.66
S2 ∩ S3 = {1/1, 0.9/3, 0.57/6, 0.5/7, 0.43/9, 0.3/12} ≡ 6.77
S2 ∪ S3 = {1/1, 0.9/3, 0.57/7, 0.5/8, 0.43/10, 0.3/1} ≡ 7.64
(S2 ∩ S3) / (S2 ∪ S3) = 6.77/7.64 = 0.89

As expected S2 and S3 seem more similar than S1 in respect to S2 or S3. This show how a model based on fuzzy bags is sensitive to differences in cardinality between tags of the same name and position. Our technique is also sensitive in differences in tags name and position, that is in differences in document structure, as shown in the next example.

3.2 Example: Sensitivity to Differences in Document Structure

Fig. 5. The XML documents D1.xml, S2.xml and D2.xml represents two different classes of users: students and teachers. They are built using the vocabulary in Fig. 3

D1 = {{1/0, 1/1}*b, {1/0, 0.9/1}*c, {1/0, 0.43/1}*h, {1/0, 0.43/1}*i, {1/0, 0.57/1}*e, {1/0, 0.3/3}*f},
S2 = {{1/0, 1/1}*a, {1/0, 0.9/1}*c, {1/0, 0.9/1}*d, {1/0, 0.43/1}*h, {1/0, 0.43/1}*i, {1/0, 0.5/1}*g, {1/0, 0.57/3}*e, {1/0, 0.3/3}*f},
D2 = {{1/0, 1/1}*b, {1/0, 0.9/1}*c, {1/0, 0.43/1}*h, {1/0, 0.43/1}*i, {1/0, 0.57/1}*e, {1/0, 0.3/2}*f}.

D1 ∩ S2 = {0.9/1, 0.57/2, 0.43/4, 0.3/7} ≡ 3.23
D1 ∪ S2 = {1/2, 0.9/4, 0.57/7, 0.43/10, 0.3/13} ≡ 7.70
(D1 ∩ S2) / (D1 ∪ S2) = 3.23/7.70 = 0.42
D2 ∩ S2 = {0.9/1, 0.57/2, 0.43/4, 0.3/6} ≅ 2.93
D2 ∪ S2 = {1/2, 0.9/4, 0.57/7, 0.43/10, 0.3/13} ≅ 7.70
(D2 ∩ S2) / (D2 ∪ S2) = 2.93/7.70 = 0.38
D1 ∩ D2 = {1/1, 0.9/2, 0.57/3, 0.43/5, 0.3/7} ≅ 3.93
D1 ∪ D2 = {1/1, 0.9/2, 0.57/3, 0.43/5, 0.3/8} ≅ 4.23
(D1 ∩ D2) / (D1 ∪ D2) = 3.93/4.23 = 0.93

4 Building a Taxonomy

Once obtained similarity values between XML documents, it is not difficult creating a taxonomy as the end result of a sort of cluster analysis. This is the organization of a collection of patterns into clusters, based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. Having a measure of comparison instead of geometric distances to compare similarity between objects, it is not possible the use of a pure cluster technique.

According to the chosen measure and using a threshold is possible grouping documents with high similarity value in the same cluster. Now is necessary extract a simple description of each cluster named cluster head. Each cluster head represents a domain class. Once obtained classes the last task of our work should be discover relations between them, choosing an appropriate technique (for example using an inheritance graph).

The achieved taxonomy could be converted into an ontology format allowing to express restrictions like OWL [6].

5 Conclusions and Future Developments

The model illustrated in this work allow the comparison between XML documents formalized as fuzzy bags. Using measure of comparison we obtain structure similarity values between them.

To avoid the loss of information introduced by the use of fuzzy bag for modelling XML documents, it could be useful formalize documents as level k fuzzy bags. Referring to Fig. 1, using this method it is possible to obtain two different representation of the two documents, preserving all the topological information. Referring to Fig. 1, the level k fuzzy bag representation of A.xml and B.xml is the following:

A = { μ(R) | R ∈ { μ(d) | d }, { μ(e) | e }, { μ(a) | a, μ(b) | b } }),
B = { μ(R) | R ∈ { μ(d) | d }, { μ(a) | a, μ(b) | b } }, { μ(e) | e } },

where the membership degrees μ(xi) could be obtained using a T-norm or other aggregators, depending on the domain we deal with.
References

1. Tversky, A.: Features of Similarity. Psychological review, volume 84, no. 4 (1977) 327–352
2. Damiani, E., Tanca, L., Arcelli Fontana, F.: Fuzzy XML queries via context-based choice of aggregations. Kybernetika, volume 36, no. 6 (2000) 635–655
3. Bouchon-Meunier, B., Rifqi, M., Bothorel S.: Towards general measures of comparison of objects. Fuzzy Sets and Systems, volume 84 (1996) 143–153
4. Bosc, P., Damiani E.: Fuzzy Service Selection in a Distributed Object-Oriented Environment. IEEE Transactions on Fussy Systems, volume 9, no. 5 (2001) 682–698
5. Rocacher, D.: On fuzzy bags and their application to flexible querying. Fuzzy Sets and Systems, volume 140, no. 1 (2003) 93–110
6. OWL Web Ontology Language. http://www.w3.org/2001/sw/WebOnt/
Managing Ontology Evolution Via Relational Constraints

Paolo Ceravolo$^1$, Angelo Corallo$^2$, Gianluca Elia$^2$, and Antonio Zilli$^2$

$^1$ Department of Information Technology
Via Bramante, 65 - 26013 Crema (Italy)
ceravolo@dti.unimi.it

$^2$ e-Business Management School - ISUFI
University of Lecce
Via per Monteroni, sn - 73100 Lecce (Italy)
gianluca.elia, angelo.corallo@isufi.unile.it
antonio.zilli@cclab.unile.it

Abstract. Ontology-based modelling is becoming increasingly important in the design of complex knowledge management applications. However, many problems related to large-scale ontology development, deployment and collaborative maintenance of related metadata still remain to be solved. Making online modifications to an ontology whose concepts are simultaneously being used for metadata generation may potentially disrupt metadata semantics and even introduce inconsistencies. In this paper we analyzed and classified operations on ontology according to their impact on metadata. The approach is aimed at environments (like our own Knowledge Hub) relying on a relational database for storing ontologies and metadata assertions and it is based on the use of database triggers for automating metadata maintenance.

Keywords: Ontology evolution, relational constraints, RDBMS

1 Introduction

Knowledge management applications enable organizations to create value by modelling their environment according to their knowledge strategy and business goals. Domain models can then be used to obtain the right resources at the right time\textsuperscript{1}. While domain modelling is traditionally slow, expensive and error-prone, the fast pace of change in today’s world means that all the elements of business applications, including domain and task models, are subject to continuous change. Nearly a decade of experience with e-business environments has shown that most artifacts used for representing business information and domain knowledge such as business ontologies, continuously undergo both light

\textsuperscript{1} This work was partially supported by the Italian Ministry of Research Basic Research Fund (FIRB) - Project KIWI.
modifications and heavy restructuring during system operation [10]. Ontology-based models, while increasingly important in the design of complex knowledge management systems, are critical to this effect. Indeed, making online modifications to an ontology whose concepts are simultaneously being used for large-scale metadata generation may potentially disrupt metadata semantics and even introduce inconsistencies [1], thus lowering the value of knowledge. Also, standard versioning techniques commonly used in the software industry for related artifacts like class diagrams and database schemata may prove less effective when applied to evolving ontologies [3]. Therefore, a strategy for controlling the evolution of domain knowledge is needed, preventing metadata obsolescence and knowledge loss.

In this paper we describe this problem referring to a scenario where a relational database is used for storing and updating ontologies and metadata in RDF/S format. Many well known implementation of this approach exist [5], including our own large scale knowledge base system (called Knowledge Hub [4]). Then, we propose an approach for preserving metadata semantics when maintaining and refining the ontological model.

2 Ontology Maintenance and Storage

Maintaining metadata such as database schemata is a time-honored and well-understood research problem, so one might wonder whether existing results [11] can be applied to ontology evolution as well. Indeed, database schemata and ontologies are both instruments for representing and describing 'a piece of the world', but there are some profound differences between them. Here, we briefly recall the ones that are most relevant to our knowledge maintenance problem [3]. The ontology evolution process can be defined as "a timely adaptation of the ontology to continuous changes in business environment, changing applications requirements and functionalities, and varying user needs and profiles" [8]. Noy and Klein [3], building on the classic definition of ontology given by Grüüber [6], have proposed three types of changes in ontology evolution: changes in domain, changes in conceptualization and changes in specification.

Changes in domain occur when some concepts or some relations between them appear or are modified according to the real world evolution. Changes in conceptualization result from a changing view of the world or from a change in usage perspective. Finally, changes in specification only take place when an ontology is translated from a knowledge representation language to another one, possibly with different semantics and expressive power. Domain and conceptualization changes are typical of everyday ontology maintenance. In Section 3 we shall see how the choice of storing ontologies and metadata in a relational database we adopted for our Knowledge Hub system enables managing some basic modifications involved in domain and conceptualization changes.

Many knowledge-based environments, including our own Knowledge Hub [4] rely on a relational database to store the RDF and RDFS used for represent-
iating respectively ontology-based assertions and the ontology structure itself\textsuperscript{2}. In our project we implemented a simple Entity-Relationship (ER) meta-model for RDF information. The ER model represents statements structured as triples using the predicate logic syntax: \((\text{arc-label} \ \text{object} \ \text{property-value})\). Triples are intended to model the concept of an object with a property value. Each triple consists of the arc label (a resource), the subject (again, a resource) and the property value (any data type from the ontology). The collection of triples forms a database relation storing the system metadata. For example, given a simple metadata like the following one.

\[
\begin{align*}
\text{<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22rdf-syntax-ns#"} \\
\text{xmlns:myonto="http://myorg.org/myonto.rdf">} \\
\text{<rdf:Description rdf:about="http://www.dti.unimi.it/ceravolo.htm">} \\
\text{<myonto:isacopyof rdf:resource="http://www.paoloceravolo.com">} \\
\text{</rdf:Description>} \\
\text{</rdf:RDF>}
\end{align*}
\]

The triple that is added to the database is:

\[
\text{(http://myorg.org/myonto.rdf#isacopyof http://www.dti.unimi.it/ceravolo.htm http://www.paoloceravolo.com)}
\]

An auxiliary table can be used to store all the triples forming the ontology schema (e.g., the assertion that \text{isacopyof} is property and its range and domain are \text{Resources}). Ontology triples are often stored separately from metadata for two main reasons: First of all, modifications to the ontology involve a separate table, making it easier to control the effects of ontology evolution on metadata, e.g., via integrity constraints (We shall develop this notion in Section\textsuperscript{3}). Secondly, RDF queries can be designed using the vocabulary defined by the ontology table and then straightforwardly executed against the metadata table\textsuperscript{12}. Another important issue is the materialization of implicit assertions generated by inheritance hierarchy. If an assertion exist stating that a resource belongs to a class, then several other assertions are implicitly stated about the membership of the resource in the corresponding superclasses (simply put, a \text{Lecturer} is also a \text{Person}). Of course, these implicit assertions can be materialized as relational tuples in order to enhance performance at the expense of storage space. Unless otherwise stated, we will assume a \text{verbose} policy where all implicit triples are materialized\textsuperscript{3}.

\textsuperscript{2} For the sake of conciseness, we do not attempt to describe RDF and RDFS here. Their primitives are defined in the W3C namespaces \url{http://www.w3.org/2000/01/rdf-schema} and \url{http://www.w3.org/1999/02/22-rdf-syntax-ns}.

\textsuperscript{3} Of course, this policy may not be sustainable for huge and deeply nested ontology trees. In our experience, however, this is seldom the case for a real world ontology. The inheritance tree of a business ontology about the Italian shoe industry used in the KH has at most 5 levels.
3 Supporting Ontology Evolution

We are now ready to explain our technique for supporting ontology evolution. Of course, the simplest ontology maintenance events, i.e. deletions, can be straightforwardly intercepted by the RDBMS and Data Manipulation Language (DML) instructions can be used to update the assertion base accordingly. On the other hand, some more sophisticated maintenance events can be managed using database triggers for automatically modifying property ranges or domains in the stored assertions. In this section, we group operations on classes and properties on the basis of their impact on metadata.

Let us use some examples to clarify our point. We will start with an operation of merging, (Fig 1). For instance suppose we designed an ontology where two sibling classes exist: Researcher and Lecturer. After some time, we realize this distinction is not useful in our domain and we decide to merge these two classes. The newly created class, for example Professor, will have as properties and subclasses the union of properties and subclasses belonging to the two original classes it originates from. Also, if we suppose to apply this operation to two classes which are one a subclass or superclass of the other, the results do not change. The trigger for this case is therefore very simple: we only need to find the statement whose properties range or domain are Researcher or Lecturer and change them into the new class Professor.

The previous example was easy, but in other cases maintenance triggers are much more complex because it is necessary to look at properties’ values. Let us consider a portion of an ontology designed as in Fig 2. The concept Person has properties name (type literal), has Role (type class), has Skill (type class) and work (type class) pointing to Work Group class. Further modelling effort may bring the designer to realize that the property has Skill should be moved from Person to Work Group, taking the property from a class to another class referenced by the origin via a property.

It is easy to see that moving a property is an operation depending on the context. For example, if we move a property from a class to its superclass, metadata do not need any maintenance because the moved property continues to belong

---

4 We say that a RDFS class A references class B when a property exist whose domain is A and range is B.
Managing Ontology Evolution Via Relational Constraints 339

Fig. 2. Moving a Property from a Class to another

to the original class due to inheritance. If we move the property to a subclass we must delete the assertions mentioning the original class, but not the auxiliary ones mentioning subclasses that could have been materialized by a verbose policy. On the other hand, if we move a property from a class to another unrelated one no fixing is possible: we have to delete all metadata assertions where the moved property occurs having the original class, or its subclasses, as its domain. In the case of our previous example, i.e. moving a property to a referenced class, some policies for maintaining metadata are possible, but it is not possible to fix the metadata without reading the corresponding data. For instance, suppose that we have a metadata asserting Person X has_Skill Skill Y and Person X works_in Work_Group Z. If we move the property has_Skill from class Person to Work_Group, we can fix the metadata as follows:\footnote{Metadata validation can be described by means of suitable inference rules; however such a description is outside the scope of this paper.} Person X works_in Work_Group Z and Work_Group Z has_Skill Skill Y.\footnote{In accord with \cite{2}, here we consider Move as a distinct operation than changing the domain of a property. When moving a property, the human designer has to be aware that metadata maintenance will be needed.}

3.1 A Classification of Operations and Triggers

We are now ready to present our operations’ classification, focusing on the notion of context-dependence. Some operations have the same effect on metadata independently from the context. Other operations produce different effects according to the relations among the classes involved. We call context-free the operations whose effects are not influenced by the type of relation occurring among classes involved. On the contrary, context-sensitive operations are the ones producing different effects depending on the relation type of involved classes. Another interesting distinction can be made among context-free operations, namely, the one between operations that affect the ontology graph and operations that only change the available set of datatypes. We call these operations respectively topology and type context-free operations. Tables in Annex summarize our basic ontology maintenance operations and the corresponding automatic effects implemented as triggers. Our list of operation usefully complements others like \cite{8} and \cite{3}.
3.2 Ontology Maintenance Events as Database Triggers

Database triggers are a well-known technique for defining a set of actions unleashed by a relational DML event such as a tuple insertion, deletion or update. Conceptually, the first step of a trigger (On event) recognizes a database event, while the second step (Do action) executes an ordered set of actions. The action may involve other tables beside the one where the event took place. For the sake of conciseness, here we will describe our ontology maintenance triggers using a simple pseudocode, where Table.field denotes a field in a relational table and Table.(field1=x).field2 projects on field2 the selected tuples whose value in field field1 equals x. In order to fill in the field field1 of a set of tuples with a value we will use assignment as follows: Table.field1←y (note that Table.(field1=x).field2←y puts the value y into field2 in all selected tuples where field1 = x). In order to store a set of field values in main memory we will assign it to a variable as follows: y←Table.field. We recall that we are using two tables: Schema for the triples of the ontology definition and Metadata for storing metadata assertions.

In order to clarify the syntax, let us consider a trigger dealing with the example shown in Fig. 2.

**Event**: Move Property from one existing Class to another referenced with the former. Change the domain of the property P from C1 to C2.

**On**

X ← Schema.(Property=P).IDSchema
Y ← Schema.(Domain=C1).IDSchema
Z = X ∩ Y
Schema.(IDSchema=Z).Domain ← C2

**Action**: If a Property P1 with C1 in the domain and C2 in the range, exist, a new assertion with C2 in the domain, P1 as property and P values as range is written.

**Do**

X ← Metadata.(Object=C1).IDMetadata
Y ← Metadata.(Arc-Label=P).IDMetadata
Z = X ∩ Y
K ← Metadata.(IDMetadata=Y).Property-Value
P ← Metadata.(Object=C1).IDMetadata
Q ← Metadata.(Property-Value=C2).IDMetadata
R = P ∩ Q
Delete Metadata.(IDMetadata=Y)
If ∃R
Then
Metadata.Object ← C2
Metadata.Arc-Label ← P
Metadata.Property-Value ← K
4 Conclusions and Further Developments

In this paper, we have presented some techniques enabling ontology maintenance aimed at a knowledge base system relying on a relational database like our Knowledge Hub. Specifically, we discussed two complementary approaches to ontology evolution operations. Currently, experimentation of these techniques on the Knowledge Hub is underway.

References

1. Parsons, J., Wand, Y.: Emancipating Instances from the Tyranny of Classes in Information Modelling. *ACM Transactions on Database Systems*, [2000].
2. Staudt Lerner, B., Habermann, A. N.: Beyond Schema Evolution to Database Reorganization. *Proceedings of OOPSLA/ECOOP* [1990].
3. Noy, N. F., Klein, M.: Ontology evolution: Not the Same as Schema Evolution. *Knowledge and Information Systems*, 5 [2003].
4. Corallo, A., Damiani, E., Elia, G.: A Knowledge Management System Enabling Regional Innovation. *Proceedings of the International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies* (KES 2002), IOS Press [2002].
5. Guha, R. V.: The rdfDB Query Language [http://guha.com/rdfdb/query.html](http://guha.com/rdfdb/query.html), [2000].
6. Gruber, T. R.: Toward principles for the design of Ontologies used for Knowledge Sharing. *Technical Report KSL 93-04*, Knowledge Systems Laboratory, Stanford University [1993].
7. Tiwana, A.: The Knowledge Management Tool Kit. *Prentice-Hall* Upper Side River [2000].
8. Stojanovic, L., Stojanovic, N., Handschuh, S.: Evolution of Metadata in Ontology-based Knowledge Management System. *Proceedings of ECIS 2002*, Gdansk [2002].
9. Carella, R., Corallo, A., Elia, G., Secundo, G.: Modelling a technological platform enabling organisational knowledge creation process. Proceedings XIII Riunione Scientifica AIIG 2002, Lecce (Italy) [2002].
10. Khosla, R., Damiani, E., Grosky, W.: Human-centered E-business, Kluwer Academic Publisher, [2003].
11. Batini, C., Ceri, S., Navathe, S.B.: Conceptual Database Design, Benjamin and Cummings, [1992].
12. Miller, L.: Tinkling, an small RDF API and Query Language implementation, [http://sw1.ilrt.org/rdfquery/](http://sw1.ilrt.org/rdfquery/), [2003].
Service Customization Supporting an Adaptive Information System

Antonio Caforio, Angelo Corallo, Gianluca Elia, and Gianluca Solazzo

e-Business Management School - ISUFI
University of Lecce
Via per Monteroni - 73100 Lecce (Italy)
{antonio.caforio, angelo.corallo, gianluca.elia, gianluca.solazzo}@ebms.it

Abstract. This work approaches the problem of discovering atomic web services that will realize complex business processes in an adaptive information system. It is proposed a model for semantic description of web services and user profile and the design of a semantic recommender engine based on this model. The recommender engine performs, during the web service discovery phase, a ”similarity evaluation” step in which it can be possible to estimate the similarity between what the service offers and what the user prefers. A semantic algorithm, that measures distance between concepts in an ontology, is used to rank the results of the semantic matching between the user profile and a list of web services, suggesting to the user the most suitable services.

1 Introduction

Web Services are modular, self-describing, self-contained applications that are accessible over the Internet and that can be composed to support business processes and value chains. Starting from this definition, an Information System can be viewed, at the application layer, as a composition of web services that can be also modified at runtime. This approach is used in the MAIS project, in order to develop a methodology and tools for the design of information systems capable of performing automatic and flexible composition of the available services, that satisfies user needs defined by his profile, the characteristics of the channel he uses to connect to the system and his geographical location. Development of such systems relies on the Service Oriented Architecture whose basic services are service descriptions and basic operations (publication, discovery, selection and binding) that produce or utilize such descriptions. Actually there are two major limits in the service description methodology: the first one is related to the lack of semantics in the standards for web service description (UDDI, WSDL) which causes

---

1 This work was partially supported by the Italian Ministry of Research - Basic Research Fund (FIRB), within MAIS project, Multi-channel Adaptive Information System. WebSite at http://black.elet.polimi.it/mais
minimal service interoperability and substituibility, weak retrieving functionalities; the second one is related to the lack of additional attributes that could be useful for end-users, such as product information that a specific service provides. Indeed, at this moment it's not possible to retrieve information about a service for hotel reservation that offers a parking service to its customers. In order to address these issues, this work focuses on the possibility to create a common layer for semantic description of both web service and user profile, using semantic annotation of attributes through a link of them with concepts in a domain ontology. Among standards for Semantic Web Service description, DAML-S enables to overcome the limits presented here. DAML-S is a service ontology, developed as part of the DARPA Agent Markup Language program, that should allow users and software agents to discover, invoke, compose and monitor Web resources offering particular services and having particular properties. Based on semantic description of the service and on the capability of extending such descriptions with additional information provided by DAML-S, our system is able to select among web services holding the same functional specification, that is the same input, output, pre-condition and post-condition parameters, those which have a greater degree of similarity with user preferences (service personalization), i.e. a user may prefer among different hotel reservation services those which advertise the payment by VISA credit card. Thus this system operates after the discovery phase in which an early selection of services that a user needs is made evaluating functional attributes. In order to address the service personalization issues, this work also focuses on the design of a Recommender Engine based on semantic algorithms used in order to calculate a similarity degree between a service and an user profile. To achieve this aim, we took inspiration from Recommender Systems (RS), which offer a solution to the information overload problem in the World Wide Web. Recommender Systems are able to learn time by time user preferences and, through their analysis, they are able to automatically identify and propose to the user interesting items. They dynamically track how single user interests change by building a user profile from his preferences. In particular we refer to two research activities which utilize user profiling techniques with an ontological representation of the application domain: OBIWAN and Quickstep. The aim of these projects is the realization of systems able to support users in the on-line search of documents both minimizing time spent for searching and rationalizing searching criteria using users’ interests. This paper is structured as follows: in section 2, we discuss how to add semantic description to web service referring also to web service publishing. In section 3, we give a brief description of user profile used by the Recommender Engine. In section 4, we introduce the Recommender Engine architecture exploiting the use of semantic matching algorithm.

2 Semantic Description of Services

A complete description of a web service should provide all the necessary information about "what" the service does, "how" it works and how to invoke its
functionalities. It also has to be consistent with real characteristics implemented by a service and contain enough information to allow a correct and more efficient execution of the discovery process. In this perspective, a service description can be composed mainly by three section:

- A human-readable description of generic characteristics of the service (name, short textual description of the functionalities, provider’s name, etc);
- A description of its interface by means of a list of functional attributes: input, output, pre and post-conditions;
- A set of non-functional attributes (i.e. quality of service and other additional attributes related to a particular instance of a service).

| General description | Name: CCII HotelReservation |
|---------------------|----------------------------|
| Description:        | This service provides the opportunity to book a room in the CCII Hotel ... |
| URI:                | http://...../CCIIHotel.html |
| ProvidedBy:         | <name>CCII Hotel</name>    |
|                     | <phone>0039 123 456 789</phone> |
|                     | <email>CCII@hotelCCII.com</email> |
|                     | <physicalAddress>Via per Monteroni, Lecce, Italy</physicalAddress> |

| Functional attributes | Input list: NumberOfPersons, DateOfArrival, DateOfDeparture, BookingReceipt, ... |
|                       | Pre conditions: CreditCardIsValid, ... |
|                       | Post conditions: ChargedAccount, ... |

| Non functional attributes | ServiceCategory: HotelReservation |
|                          | QoS parameters: Authentication, Accuracy, Cost, Stability, ... |
|                          | Additional Parameters: Five Star Category, Indoor Swimming Pool, Sauna, Air Conditioning, Conference Hall, Restaurant, ... |

The solution proposed in this work exploits semantic annotation to provide a richer and more formal service description in order to enable a more efficient discovery mechanism. Semantic annotation aims to create an agreement on a shared vocabulary and a semantic structure for information exchange about a specific domain. We can distinguish among different types of semantics. Each type depends on the part of the description which is associated to and on the purpose it is used for:

- Semantics of data/information: it consists in a formal definition of input and output data of a service. In this case it is used to improve the discovery process and to solve problems in the aggregation of services.
Operational semantics: it consists in a formal representation of capabilities of a service obtained by means of annotation of pre-conditions and post-conditions. It is used during the discovery process to support composition of services.

Semantics of non functional parameters: it consists in a formal description of a set of QoS parameters which are domain independent and a set of additional parameters which are specific to a domain and to a service. It is used to select the more appropriate service, given a user profile.

Service description and its semantic annotation are performed to support two phases of the lifecycle of a web service: Publishing and Discovery. In the publishing phase, a Service Provider has to provide information not only about service interface and quality of service but also about additional information which are specific to a service and could be more useful for customers to decide which service to invoke. In this phase, semantic annotation of parameters that describe a service will be performed providing a "link" between these parameters and concepts in a domain ontology. This "semantic markup" will be stored in a service description file that will be accessible over the Internet and retrieved through the use of some Registries. Considerations about which information have to be published in the Registries are not in the scope of this work. For example, a Provider who wants to publish a service to book rooms in his hotel, whose name is CCIIHotelReservation, will use, in the publishing phase, a Tourism or Accomodation ontology, available in an Ontology Repository, to semantically describe parameters of his service. Annotation is performed browsing the ontology (a small subset of the ontology is shown in (Fig 1)) and selecting concepts that have to be linked to Input, Output, Pre-conditions, Post-conditions, ServiceCategory, Quality of Service and Additional attributes of the service.

We assume that annotation of a parameter, in the CCIIHotelReservation service description, with Restaurant concept means that the Provider can provide all kinds of restaurant defined in the ontology that Restaurant concept subsumes, that is nationalMenuRestaurant, ethnicMenuRestaurant and internationalMenuRestaurant.

![Fig. 1. A subset of an Accomodation Ontology example](image-url)
3 User Profile

In the conception of the following model we have posed our attention on the information contained in the profile of a user that interacts with a platform offering web services. This information is expression of user preferences and needs related to the content of services and the way services are accessed. The proposed user profile model, includes all that information that describes user in terms of:

- User attributes:
  - Static information, i.e. date of birth, address, etc...;
  - Dynamic information (system dependent or independent), i.e. expertise in the use of the system, role played in the interaction with the system, etc);

- User preferences related to the configuration of devices used to connect to the system such as PDA, notebook, mobile phone, etc (settings, i.e. LocationEnable to enable geographical location of the user, etc) and personal preferences related to services a user is interested in (i.e. user prefers to stay in a 4star hotel);

- Behavioural information of the user (i.e. when a user travel by plane as a businessman he stays in a 5stars hotel).

For our work, the user profile section containing personal preferences in using services is the most interesting. Preferences are obtained through the analysis of transactional data and are subdivided in categories, these are the same categories which services belong to (classification systems for service categories are the NAICS or the UNSPSC). In particular, the proposed model highlights that a user can express preferences for each service category in terms of preferred QoS and additional parameters. It is worth noting that attributes of the user profile that references preferred service parameters, being derived from interaction between users and semantic services and analysis of transactional data, are semantically annotated.

4 Recommender Engine

There are two different kind of recommendation that can result from the description of the user contained in its profile:

- A content-based recommendation, which directly answers to a particular request from an user and analyzes exclusively both user and service description;
- A hybrid recommendation, processed from the description of a cluster of users;

Although both recommendation typologies recommend services, they present some differences in terms of amount of processed information and of processing context. It is clear that the first kind of recommendation will be processed after
the discovery phase will be over, the second one can be processed independently from any user request and it can be proposed in a proactive way to the users. Also the amount and the typology of processed information is different for the two kind of recommendation: for the first one we need the service and the user to be described in the same terms, in the second kind the recommendation will take into account all the information contained in the user profile. The latter recommendation[4] [5] can be obtained by building user clusters with item-based characteristics and subsequently, on those clusters, by processing recommendations. In order to achieve this goal we need first to create clusters of users; then to define a set of item (services) that can be associated to each cluster and finally to associate a weight to each item (service) in order to make a ranking. Concerning the content-based recommendation, our analysis will take into account service descriptions contained in a list that will be considered as the output of the discovery phase. In fact the discovery of a service produces as output a list of services with the equivalent functional specification (that is similar input, output, pre-condition and post-condition parameters) and compatible to the user request. An useful recommendation for the user is the ranking of services in the list, extracted after his request. In order to realize a ranking of services, we need first to analyze information contained in the user and service profile and then to preset an algorithm able to translate the analysis into a ranking. The user profile will be described in terms of a service category, operational metrics (Quality of Service) and in terms of additional attributes related to those services the user has made use of. This part of the user description will be compared to each service description contained in the list since user and service are described through the same attributes. Exploiting each attribute (both additional and QoS attributes) link to a concept in the ontology, we evaluate the service rank by measuring the semantic similarity of all the attributes in the service description with attributes in the user profile. The way the Recommender Engine will calculate a similarity, between a service and the user who requested it, takes into account the semantics associated to them. The algorithm developed in this work relies on the evaluation of the SemSimilarity() function defined as follow:

- SemSimilarity(UP,SP) is a function that returns a percentage value representing the similarity degree between its arguments
- The arguments are a User Profile(UP) and a service description profile(SP)
- The function measures the semantic similarity on all the non-functional parameters contained in the user profile and in the service profile

The semantic similarity is a consequence of the DegreeOfMatch calculus between two concepts. It depends on the relation between concepts contained in the user profile and in the service description. In particular, the DegreeOfMatch is given by the minimum distance between concepts in the ontology view and it’s possible to distinguish four different kind of match[6]:

1. **exact** can occur in two cases: in the simplest situation when two concepts coincide, and also when the concept specified in the user profile is a direct specialization (first level specialization) of the concept specified in the web service description and contained in the ontology;
2. plugIn occurs when the concept specified in the web service description is a direct specialization of the concept specified in the user profile. This kind of relation is less strong than the previous;
3. subsumes occurs when the concept specified in the user profile is a specialization of the concept specified in the web service description;
4. fail occurs when no transitive relation exists between two specific concepts.

The four cases that degreeOfMatch can assume will be associated to discrete values, considering that the preferable degree is the exact and the less preferable is the subsumes. For example, it is possible to assign discrete values to the four cases of the degreeOfMatch as follows:

| degreeOfMatch | Value assigned |
|---------------|---------------|
| Exact         | 1             |
| plugIn        | 0.6 - 0.7     |
| Subsumes      | 0.3 - 0.4     |
| Fail          | 0             |

It might be necessary to introduce some rules in order to be able to choose the best match and to reduce the computational complexity of the algorithm. If one of the additional attributes contained in User Profile match with more of one attributes of the Web Service Profile, we must choose the best combination. In general, we can say that having an user profile with n attributes, there will be n*n pairings: we must choose the n distinct pairings with the highest value. Finally, the value returned by the degreeOfMatch is normalized with the number of preferences contained in the user profile. More in particular:

SemSimilarity(UP,SP)=[\omega_1 QosSS(UP,SP) + \omega_2 AddSS(UP,SP)] \in [0,1],

and \( \omega_1 + \omega_2 = 1 \), in which QosSS is the semantic similarity function calculated on the quality of service parameters, whereas AddSS is the semantic similarity function measured on the additional attributes of a web service. Thus, given a user profile UP and an web service description SP, we define:

\[
QosSS(UP,SP) = \frac{\sum_i DegreeOfMatch(\alpha_i,\beta_i)}{n}
\]

and:

\[
AddSS(UP,SP) = \frac{\sum_i DegreeOfMatch(\delta_i,\lambda_i)}{m}
\]

The pairs \((\alpha_i,\beta_i)\) and \((\delta_i,\lambda_i)\) indicate pairs of concepts (to which a pair of non-functional parameters in the user profile and in the service description have been semantically linked) belonging to the ontology \(\Omega\) and \(\Omega'\). The number of QoS parameters is n while the number of additional attributes is m. DegreeOfMatch function calculates the distance between two concepts which belong to an ontology and return a value according to the four cases presented above. That is, the estimation of the semantic similarity is a direct consequence of the DegreeOfMatch applied to pairs of non-functional parameters.
5 Conclusion

This work contributes to open a new field for the application of user profiling techniques and information filtering systems: web services and adaptive information systems. Recommendation modifies traditional lifecycle of a web service, identifying a new value-added activity that supports discovery phase of services. We presented a semantic algorithm that exploits semantic description of both web services and user profile. This semantic layer allows to abstract from the specific meaning of a term, to operate on a vocabulary (a set of terms) and a semantic structure for information exchange about a specific domain. It also can support automation of the service discovery process enabling service customization mechanisms. We expect to enhance recommendation algorithms in order to evaluate complex business process, whose single web services have already been ranked by the Recommender Engine.

References

1. DAML-S Coalition, www.daml-s.org
2. A. Pretschner, S. Gauch, Ontology based Personalised Search, University of Kansas, 2000
3. S. Middleton, D. De Roure, N. Shadbolt, Capturing knowledge of user preferences: ontologies in recommender systems, Department of Electronics and Computer Science, University of Southampton, 2002
4. Alspector, J., Kolcz, A., Karunanithi, N.: Comparing Feature-Based and Clique-Based User Models for Movie Selection. In: Proceedings of the 3rd ACM Conference on Digital Libraries. Pittsburgh, PA (1998);
5. Sarwar, Karypis, Konstan, Riedl, Item-based Collaborative Filtering Recommendation Algorithms, GroupLens Research Group, University of Minnesota, USA, 2001
6. Paolucci et al., Semantic Matching of Web Services Capabilities, Carnegie Mellon University, Pittsburgh, USA, 2002;
7. North American Industry Classification System (NAICS 2002) http://www.census.gov/epcd/www/naics.html
8. United Nations Standard Products and Services Code (UNSPSC) http://www.unspsc.org
Using Design Information to Support Model-Based Fault Diagnosis Tasks

Katsuaki Tanaka¹, Yoshikiyo Kato², Shin’ichi Nakasuka³, and Koichi Hori¹

¹ AI Laboratory, RCAST, University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo, 153–8904, Japan
² Japan Aerospace Exploration Agency, 2-1-1 Sengen, Tsukuba-city, Ibaraki, 305–8505, Japan
³ Intelligent Space Systems Laboratory, Dept. of Aeronautics & Astronautics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan

Abstract. In today’s complex systems, it can be difficult to specify the cause of a fault. In general, backward reasoning such as Fault Tree Analysis, which derives a cause from a fault, is used for diagnosis. Knowledge that derives causes from faults is necessary for this inference, and it is required beforehand to make hypotheses about faults and their possible causes, and to validate them. However, especially in the field of space, each system is peculiar to each project, and knowledge that associates causes with faults tends to be insufficient. Model-based diagnosis is therefore used. It uses a model of a target system to recreate the fault, thus showing the condition of the target system. To construct such models, knowledge of the target system is required. We have developed a system that collects knowledge during the design process, and provides it to the model construction process.

1 Introduction

In today’s complex systems, it can be difficult to specify the cause of a fault. In general, backward reasoning such as Fault Tree Analysis, which derives a cause from a fault, is used for diagnosis. Knowledge that derives causes from faults is necessary for this inference, and it is required beforehand to make hypotheses about faults and their possible causes, and to validate them. However, especially in the field of space, target systems can be very remote and each system is peculiar to each project. Therefore, knowledge that associates faults with causes tends to be insufficient. Methods that simulate a fault in a real system on the ground are adopted to investigate possible causes. This method uses forward reasoning that leads to faults linked to particular causes. If a model is constructed and simulated in a computer, a fault can be recreated more easily, and less time is required to diagnose the real system. This is a model-based diagnostic system.

A great deal of knowledge concerning the target system is required to construct such a model. Knowledge that is made and used during the design of the target system is reusable because a model-based simulation replaces the examination of the real system. In this work, a system was constructed that collects design rationales during the design process and provides them to the model construction process with context.
Design rationales are applicable to the generation of backward diagnostic knowledge [1]. For this purpose, however, design rationales must be described formally for generation. We use design rationales to construct a model for the model-based diagnostic system, so design rationales do not require formal description. Applying design rationales for model construction is similar to applying design verification or redesign, as described in [1].

The system described in this paper consists of three subsystems: the model-based fault diagnostic system, Integrated Design Information Management System (IDIMS), and the complex problem management system. The model-based fault diagnostic system is a system that simulates a fault using a model, and specifies the cause of the fault. IDIMS manages the design rationales generated in the design, examination, and operation processes of artifacts. The complex problem management system manages the problem structure of a large-scale problem, and knowledge concerning it. It aims to support construction of a model that can be used by the model-based fault diagnostic system, by providing knowledge that IDIMS manages via the complex problem management system in an appropriate context.

2 Model-Based Fault Diagnostic System

The model-based fault diagnostic system simulates a target system with a model. This system judges that the model that gives simulation results nearest the appearance of the target system best describes the current condition of the target object [2].

The space satellite “XI–IV”, developed by the Intelligent Space Systems Laboratory at the University of Tokyo is used as an example. A model of the communication subsystem and the posture control subsystem was constructed based on the satellite, and the model was used for the diagnostic system.

We used values obtained from sensors built into the satellite (telemetry data) as input data. We used 43 series of telemetry data. Twelve of these are from the
posture control subsystem and 31 from the communication subsystem. In this paper, we diagnose the communication subsystem. Fig. 1 shows a model of it.

To simulate a communication subsystem is very difficult, because the I/O values of each component take continuous values, but it is assumed that the binary result of normal or abnormal is obtained as telemetry data.

2.1 The Diagnosis Process

The diagnosis is made by comparing generated data and telemetry data obtained from the satellite. First, the diagnostic system checks the values of the given telemetry data, and prunes some fault conditions to decrease computing time. Then the system simulates the satellite based on the given model. Next, it calculates distances between the generated data and the actual data. Close simulation results are assumed to be similar to the real satellite data. The fault diagnostic system displays graphs of simulation results and candidates for the current satellite condition.

2.2 Satellite Model

The model that the fault diagnostic system uses should be able to generate data almost identical to the telemetry data obtained from the satellite. Therefore, each component in the model receives values from other components, operates based on those values, and influences other components, just like each component of an actual satellite. For instance, a thruster (small rocket engine for course correction) consumes fuel (decreasing the amount of fuel stored in a tank) when the input value of the thruster firing command shows that it is turned on, and changes the orbit (changes reading values of gyro-compasses and earth sensors). It is necessary for each component in the model to describe how it works, and what it is connected to. Moreover, it is necessary to simulate extraordinary conditions, so the model contains not only the normal state descriptions but also abnormal state descriptions.

It is necessary to generate data at high speed because multiple simulations are necessary for forward reasoning; thus, the operation code and the operating mode of each component are written in C.

3 IDIMS

The Integrated Design Information Management System (IDIMS) is a system that collects, restructures, and provides design rationales generated during the design, examination, and operation processes of artifacts. In this paper, we use a subsystem of IDIMS that acquires design rationales from e-mails. It collects sentences that indicate “Issues” or “Decisions” from e-mail texts.

Fig. 2 shows the outline of this subsystem. Users tag sentences that describe issues with “Issue” tags, and sentences that describe decisions with “Decision” tags in e-mail messages. The program is built in the mail server, and identifies Issues and Decisions in a message, makes a template for a reply, and transmits it.
to the other user. IDIMS attempts to reduce user loads by reducing the number of tag types compared with gIBIS[4].

Without Issue/Decision tags, the dependencies between e-mail messages could be understood as dependency problems. However, two or more issues tend to be discussed in one e-mail that identifies problems. Then, when a reply about one issue is contained in an e-mail, it is often assumed that the other issues have been solved. Therefore, it is useful for users to describe Issue/Decision tags. In this work, we construct a mechanism that collects Issues and Decisions according to context in a problem. This clarifies the user’s position in a problem structure through a problem unit of the complex problem management system.

Although IDIMS has its own knowledge display system, we constructed a new interface to access knowledge stored in IDIMS through SOAP, which is the protocol for Web services. The complex problem management system can acquire the knowledge required to construct the satellite model through this interface.

4 Complex Problem Management System

The complex problem management system is a system that allows many people to participate in the solution of large, complex problems [5]. The structure of a large-scale problem is divided in a top-down way. The system creates problem units according to this approach. For instance, the initial problem of designing a space satellite is divided into such problems as designing a power supply subsystem, a communication subsystem, and the structure, because the entire space satellite cannot be designed as a whole. A person is put in charge of each problem, and each person is encouraged to solve his or her own problem. Then, because even the design of the power supply subsystem is complex, it is divided recurrently into the design of a battery, distributors, and other components. As a result, the structure of a problem is determined in a top-down way, like a tree.

The complex problem management system defines each sub-problem, such as the design of a power supply subsystem or design of a battery as a problem unit. The system handles a large-scale problem as a set of problem units.
It is possible to recycle a problem unit, thus preserving information from past problem solutions. For instance, a problem unit concerning design of a battery and problem units for sub-problems can be recycled in the design of a power supply system on another device. Knowledge and information stored in problem units for battery design, such as decisions, discussions or information such as who dealt with the problem are reusable.

4.1 Communication Between Users

In this system, a user is located as a part of a problem unit. A user may have authority to control one or more problem units, and two or more users can participate in a problem unit. When an adjustment with another problem unit is required, for instance, when competing for resources, users can send e-mails from a problem unit to the other problem unit. Each person's name listed in a problem unit's user interface has a “Mailto:” link. In the link, an ID for the problem unit is embedded in the subject of the e-mail, and it enables identification of where the discussion was held. In addition, by describing Issue/Decision tags in e-mail messages, the collection of issues and decisions corresponding to problems is possible. Refer to Fig. 2 for an example of an e-mail message.

4.2 Autonomous Problem-Solving

It is possible to call a Web service as a sub-problem solution process. The Web service is called when the execution button prepared in the user interface is selected, and the result is substituted for the variable. It is also possible to call a system with a Web-based user interface. The page is displayed by clicking a URL. A problem unit works as an HTTP proxy, and results can be automatically acquired by a unit from hidden tags in the user interface that a problem unit can interpret.

5 Case Study

A problem occurred in the down-link when a communication subsystem experiment was performed with a balloon. No result that would satisfy the user was obtained from the diagnostic that used the model-based fault diagnostic system. When the model for the diagnostic system was reconstructed with the complex problem management system and IDIMS for referring to design rationales, a temperature dependency of the communication subsystem equipment was discovered. A satisfying result was obtained with this new model.

Preparation. Tags such as Issue or Decision, and object names assuming that problem units were used, were added to 267 e-mail messages exchanged within the design phase of the satellite communication subsystem. These messages were analyzed, and issues and decisions were stored in IDIMS. The communication subsystem model (Fig. 1(a)) was converted into the problem structure (Fig. 1(b)), and it was stored as a case of a satellite design in the complex problem management system. Each design of a component was defined as a problem unit.
Model Building. The leader of the subsystem design is assigned to the top problem, and then assigns other users as owners of up-link or down-link design. The user in an upper-level problem assigns an eligible person to each subproblem.

Each component designer describes the model through the problem unit assigned to the problem corresponding to that component design. An example of a problem unit user interface is shown in Fig. 3. The model consists of behavior code and behavior modes and connections to other components. The upper-level problem unit collects these code fragments, brings them together in one file, and passes them to the problem unit that controls the model-based fault diagnostic system. The unit compiles these files, and generates the diagnostic system.

Diagnosis. The model-based fault diagnostic system is called from a problem unit and performs diagnosis, and the user checks the result. In this case, the first diagnosis performed based on the model was constructed without the design rationales stored in IDIMS. All the possible diagnoses were too distant from the telemetry data, so the user could not find a satisfactory result. The model was then reconstructed with the complex problem-solving system and IDIMS. An example screen for the problem unit that calls the IDIMS Web service interface and displays results is shown in Fig. 3. As a result, it becomes apparent that a temperature range for the FM receiver to guarantee operation was proposed by the designer, but the other users had forgotten this. This temperature dependency was built into the model, and the model-based fault diagnostic system was used again. A result with only a small distance from the telemetry data was found, and the user judged that this result was satisfactory.
6 Summary

In this paper, a system was constructed that collects design rationales during the design process and provides them to the model construction process of a model-based diagnostic system. The model-based diagnostic uses forward reasoning that leads to a present fault state from the model of the target system. The knowledge required for this inference is that which allows a model that faithfully expresses the behavior of the target to be constructed. It is simply the knowledge required to redesign the component. Then, to make a system that provides knowledge used in design processes for model constructions, IDIMS and the complex problem management system were integrated.

Designers have discussions with other participants via e-mail through problem units of the complex problem management system. IDIMS acquires design rationales from these e-mail messages with information about which part of the problem they relate to. Design rationales are provided to designers through a problem unit corresponding to the context of a problem. The model used by the model-based fault diagnostic system is constructed using this mechanism. As a result, a cause of a fault that had not been found in other tests was discovered.

The problem structure used within this paper relates to the design of a satellite that had already been completed, and design rationales were acquired from the e-mail messages that were exchanged during the design process. Therefore, neither structure nor rationales were acquired in real time through this system. We plan to investigate the efficiency of this system on actual problems.

References

1. B. Chandrasekaran, A.K. Goel and Y. Iwasaki, Functional Representation as Design Rationale, IEEE Computer, Vol. 26, No.1, pp. 48–56, 1993.
2. S. Nakasuka, S. Ogasawara, M. Takata and T. Yamamoto, Model-based Autonomy – How Model without Human Expertise Can Facilitate Fault Diagnosis and Other Spacecraft Tasks?, 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS’01), Montreal, Canada, June 2001.
3. Y. Kato, T. Shirakawa, K. Taketa and K. Hori, An Approach to Discovering Risks in Development Process of Large and Complex Systems, New Generation Computing, 21(2), pp. 161–174, Ohmsha, 2003.
4. J. Conklin and M.L. Begeman, gIBIS: A tool for all reasons, Journal of the American Society for Information Science, 40(3), pp. 200–213, 1989.
5. K. Tanaka and S. Ohsuga, Model-Based Creation of Agents and Distribution of Problem Solving, In Proceedings of the 2001 International Conference on Intelligent Agent Technology, Maebashi, Japan, October 2001.
6. ISSL CubeSat Project Communication Subsystem Group, CubeSat Balloon Experiments Data Analysis Report, 2001, http://www.space.t.u-tokyo.ac.jp/cubesat/publication/balloon.pdf (in Japanese)
7. Y. Kato, T. Shirakawa and K. Hori, Utilizing Fault Cases for Supporting Fault Diagnosis Task, In Proceedings of the Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Crema, Italy, September 2002.
Fault Detection and Diagnosis Using the Fuzzy Min-Max Neural Network with Rule Extraction

Kok Yeng Chen¹, Chee Peng Lim¹, and Weng Kin Lai²

¹School of Electrical and Electronic Engineering, University of Science Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia
cplim@usm.my
²MIMOS Berhad, Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia
lai@mimos.my

Abstract. In this paper, a symbolic rule learning and extraction algorithm is proposed for integration with the Fuzzy Min-Max neural network (FMM). With the rule extraction capability, the network is able to overcome the “black-box” phenomenon by justifying its predictions with fuzzy IF-THEN rules that are comprehensible to the domain users. To assess the applicability of the resulting network, a data set comprising real sensor measurements for detecting and diagnosing the heat transfer conditions of a Circulating Water (CW) system in a power generation plant is collected. The rules extracted from the network are found to be compatible with the domain knowledge as well as the opinions of domain experts who are involved in the maintenance of the CW system. Implication of the FMM neural network with the rule extraction capability as a useful fault detection and diagnosis tool is discussed.

1 Introduction

Fault detection and diagnosis (FDD) in complex processes such as those in a power generation plant is a very crucial task. A good FDD system is of great value in many industrial applications. Early prediction of possible fault states allows maintenance work to take place before the machine/system breaks down. From the literature review, there is evidence that neural network (NN)-based systems are effective in tackling FDD problems. A neural-fuzzy system was used for intelligent diagnosis of turbine blades [1]. A backpropagation NN was employed for fault detection in pneumatic control valve actuators [2]. Applicability of NN-based systems to FDD of helicopter gearbox [3] and diesel engine [4] has also been reported.

One of the limitations of NN models is the so-called “black-box” phenomenon, i.e., no justifications are given for their predictions. Indeed, it is hard for users to accept NN results if no rationale/justification is given on how a prediction is reached. A number of techniques have been proposed for rule extraction from NNs, e.g., the KT extraction algorithm [5], the subset algorithm [6], and the rule-extraction-as-learning technique [7]. The main focus of all these is to provide justification for NN predictions in the form that users can easily comprehend.
In this paper, the Fuzzy Min-Max neural network (FMM) [8] (hereafter refer to as FMM) is enhanced with a rule learning and extraction capability. There are several reasons that motivate the use of FMM. First, it has the capability of learning in a single pass through the data samples and is able to build and fine-tune the decision boundaries of different classes online. Second, the proposed rule extraction procedure, which can be categorized under the decompositional approach [9], is able to extract knowledge from FMM in a straightforward manner. In the rest of this paper, Fuzzy Min-Max neural network will shorten to FMM. The applicability of FMM with symbolic rule extraction to FDD tasks is then investigated.

This paper is organized as follows. Section 2 briefly introduces FMM and its rule extraction algorithm. Section 3 describes a case study on FDD and the experimental results with discussion. Conclusions are drawn in section 4.

2 The Fuzzy Min-Max Neural Network

FMM has the ability to add new pattern classes online and to refine the existing pattern classes as new data samples are received. The main components of this model are the hyperboxes that are created and defined by pairs of min-max points and their corresponding membership functions that describe the degree to which an input pattern fits within the min-max points [8]. Figure 1 shows a three-layer FMM while Figure 2 illustrates the segregation of several (hyper) boxes for a two-class problem in a two-dimensional space. In general, each $F_b$ node represents a hyperbox fuzzy set where the $F_a$ to $F_b$ connections are the min-max points and the $F_b$ transfer function is the hyperbox membership function. Each hyperbox fuzzy set, $B_j$, is defined by

$$B_j = \{A, V_j, W_j, f(A, V_j, W_j)\} \quad \forall A \in I^n$$

where $A = (a_1, a_2, \ldots, a_n)$ is the input pattern $V_j = (v_{j1}, v_{j2}, \ldots, v_{jn})$ is the minimum point of $B_j$, $W_j = (w_{j1}, w_{j2}, \ldots, w_{jn})$ is the maximum point of $B_j$. The membership function, $f(A, V_j, W_j)$, is defined by

$$b_j(A_h) = \frac{1}{2n} \sum_{i=1}^n \left[ \max(0, 1 - \max(0, \gamma \min(1, a_{hi} - w_{ji})) + \max(0, 1 - \max(0, \gamma \min(1, v_{ji} - a_{hi})))) \right]$$

where $A_h$ is the $h$-th input pattern, $\gamma$ is the sensitivity parameter that regulates how fast the membership values decrease as the distance between $A_h$ and $B_j$ increases. The membership function indicates the measurement of how far each component is greater (lesser) than the maximum (minimum) point value along each dimension. If $b_j(A_h) \rightarrow 1$, the point should be more “contained” by the hyperbox. When $b_j(A_h) = 1$, it represents complete hyperbox containment. Each $F_c$ node represents a class, and output of the $F_c$ node represents the degree to which the input pattern $A_h$
fits within class \( k \). The connections between the \( F_a \) and \( F_c \) nodes are binary valued and are stored in matrix \( U \) as shown in equation (3)

\[
u_{jk} = \begin{cases} 
1 & \text{if } b_j \text{ is a hyperbox for class } c_k \\
0 & \text{otherwise}
\end{cases}
\]

(3)

where \( b_j \) is the \( j \)-th \( F_a \) node and \( c_k \) is the \( k \)-th \( F_c \) node. The transfer function for each \( F_c \) node performs the fuzzy union operation of the appropriate hyperbox fuzzy set values as defined by

\[
c_k = \max_{j=1}^m (b_j \times u_{jk})
\]

(4)

Thus the class whose hyperbox is best-suited to the input vector is chosen as the predicted class for the input pattern.

The learning algorithm of FMM comprises an expansion/contraction process. The algorithm first searches a hyperbox of the same class that can expand (if necessary) to include incoming input pattern. If the hyperbox that meets the expansion criteria, as described in equation (5), the input pattern will be included into the hyperbox. Otherwise, a new hyperbox is formed and added to the network. The maximum size of a hyperbox is bounded by \( 0 \leq \theta \leq 1 \), a user defined value.

\[
n \theta \geq \sum_{j=1}^n \left( \max(w_{ji}, a_{ji}) - \min(v_{ji}, a_{hi}) \right)
\]

(5)

In summary, the dynamics of FMM comprises a three-step process:

1. **Expansion**: Identify the hyperbox that can expand and expand it. If an expandable hyperbox cannot be found, a new hyperbox for that class will be added.
2. **Overlap Test**: Determine if any overlap exists between hyperboxes from different classes.
3. **Contraction**: If overlap between hyperboxes of different classes does exist, eliminate the overlap by minimally adjusting each of the hyperboxes.

The detailed description of FMM can be found in [8].

### 2.1 Rule Extraction from the FMM Neural Network

To overcome the “black-box” phenomenon, a rule learning and extraction algorithm, as proposed in [10], is introduced for integration with FMM. Though a large number of hyperboxes are created, some of them are rarely used during prediction. To reduce the complexity of FMM, network pruning is first performed. The objective is to remove those hyperboxes that have low confidence factors, so that a high accuracy rate of prediction is attained. A hyperbox is eliminated when its confidence factor is lower than a user-defined pruning threshold, \( \tau \). The confidence factor (CF) of each hyperbox is expressed as [10].
\[ CF_j = \lambda U_j + (1 - \lambda) A_j \]  

where \( U_j \) and \( A_j \) are, respectively, Usage and Accuracy of the \( j \)-th hyperbox, and \( \lambda \in [0, 1] \) is a weighting factor. Usage, \( U_j \) is defined as the fraction of the number of training patterns coded by any hyperboxes \( b_j (C_j) \) that predict a particular outcome over the maximum number of training patterns coded by any hyperboxes \( J (C_j) \) that predict the same outcome [10], i.e.,

\[ U_j = \frac{C_j}{\max[C_j]} \]  

Accuracy, \( A_j \) is defined as the fraction of the percentage of prediction patterns predicted by hyperboxes \( j (P_j) \) over the maximum percentage of prediction patterns predicted by any hyperboxes \( J (P_j) \) that predict the same outcome [10], i.e.,

\[ A_j = \frac{P_j}{\max[P_j]} \]

To further facilitate the rule interpretation in a comprehensible form, weight quantization by truncation [10] is applied. The method divides the range \([0, 1]\) into \( Q \) intervals, and assigns a quantization point to the lower bound of each interval

\[ V_q = \frac{(q - 1)}{Q} \]

for \( q = 1, 2, \ldots, Q \), where \( Q \) is the quantization level. Details of the algorithm for network pruning and rule extraction can be found in [10].
3 A Case Study

In this work, applicability of FMM with symbolic rule extraction to FDD tasks was investigated by using a set of real sensor data collected from a power generation plant in Penang, Malaysia. The system under scrutiny was the Circulating Water (CW) system [11]. The function of the CW system is to supply a continuous amount of cooling water to the main turbine condenser to condense steam from the turbine exhaust and other steam flows into the condensers. The CW system contains all piping and equipment (such as condensers and drum strainer) between the seawater intake and the outfall where water is returned to the sea.

A set of 2439 samples database on 80MW of the targeted power generation was collected. Each data sample consisted of 12 features, as shown in Table 1. The case study involved detecting and diagnosing the heat transfer efficiency in the condensers of the CW system. The heat transfer conditions were categorized into 2 classes, i.e., efficient or inefficient heat transfer in the condensers. From the dataset, 1224 samples showed inefficient heat transfer conditions while the remaining samples showed efficient heat transfer conditions. For experimentation, all the data samples were equally divided into 3 subsets for training, prediction, and testing.

After several trial runs, FMM produced the best test results when $\theta$ was varied from 0.01 to 0.10. As shown in Table 2, the best accuracy rate achieved was 97.66%. Bootstrapping [12] was applied to the results to determine the 95% confidence intervals of the network performance. The results are shown in Table 3. Notice that the difference between the lower and upper limits of the confidence interval was small, suggesting that the network performance was stable.

To facilitate rule extraction, network pruning was then performed to remove those hyperboxes (prototypes) that had a low confidence factor. The pruning threshold, $\tau$, was varied from 0.0 to 0.8, with $\theta$ set at 0.04. Table 4 shows the test accuracy rates and the number of hyperboxes after pruning. The results from $\tau = 0.7$ was selected for rule extraction as it yielded a smaller-sized network with a high accuracy rate.

### Table 1. Parameters used for the case study

| No | Parameter          | Description                               |
|----|--------------------|-------------------------------------------|
| 1  | LPT A              | Low Pressure Cylinder Exhaust Temperature A|
| 2  | LPT B              | Low Pressure Cylinder Exhaust Temperature B|
| 3  | GEN                | Generator                                 |
| 4  | CWMT A             | Condenser Circulating Water Inlet Temperature A|
| 5  | CWMT B             | Condenser Circulating Water Inlet Temperature B|
| 6  | CWOT A             | Condenser Circulating Water Outlet Temperature A|
| 7  | CWOT B             | Condenser Circulating Water Outlet Temperature B|
| 8  | CMP A              | Condenser Circulating Water Inlet Pressure A|
| 9  | CWOP A             | Condenser Circulating Water Outlet Pressure A|
| 10 | CMP B              | Condenser Circulating Water Inlet Pressure B|
| 11 | CWOP B             | Condenser Circulating Water Outlet Pressure B|
| 12 | VAC                | Condenser vacuum                          |

### Table 2. Test results

| $\theta$ | Test Accuracy (%) | $\theta$ | Test Accuracy (%) |
|----------|-------------------|----------|-------------------|
| 0.01     | 97.17             | 0.06     | 96.06             |
| 0.02     | 97.05             | 0.07     | 96.06             |
| 0.03     | 97.66             | 0.08     | 95.33             |
| 0.04     | 97.66             | 0.09     | 94.34             |
| 0.05     | 95.85             | 0.10     | 95.08             |
Table 3. Bootstrapped results

| r  | Test Accuracy (%) | No of hyperboxes |
|----|-------------------|------------------|
| 0.0| 97.06             | 499              |
| 0.1| 97.05             | 261              |
| 0.2| 97.05             | 196              |
| 0.3| 97.05             | 196              |
| 0.4| 96.93             | 194              |
| 0.5| 96.93             | 194              |
| 0.6| 96.68             | 192              |
| 0.7| 96.06             | 100              |
| 0.8| 79.09             | 27               |

Table 4. Pruning results

| Number of Resamplings | Confidence Intervals | Mean (%) |
|-----------------------|----------------------|----------|
|                       | Lower (%)            | Upper (%)|
| 200                   | 96.695               | 97.301   |
| 400                   | 96.686               | 97.285   |
| 600                   | 96.694               | 97.310   |
| 800                   | 96.686               | 97.309   |
| 1000                  | 96.685               | 97.301   |

Table 5 shows six rules from each class that have the best confidence factor while Table 6 shows the interpretation of the first positive rule and the second negative rule. According to the domain knowledge and opinions from the domain experts, the heat transfer condition can be ascertained by monitoring certain parameters such as LPT A, LPT B, CWIT A, CWIT B, CWOT A, CWOT B, and VAC. Note that VAC for most of the positive rules ranged from 3 (medium) to 4 (high) while the negative rules ranged from 1 (very low) to 2 (low). Indeed, in order to have efficient heat transfer in the condensers, the condenser vacuum should be preserved at a low level. If CWIT A and CWIT B were increased, the steam temperatures (LPT A and LPT B) exiting the turbine would increase to establish the needed differential temperature for continuous heat transfer during power generation [14]. This situation can be identified in most of the extracted rules, where LPT A and LPT B were less than 2 (low) while CWIT A and CWIT B ranged from 2 (low) to 3 (medium) in the negative rules. On the other hand, if LPT A and LPT B increased from 3 (medium) to 4 (high), CWIT A and CWIT B would increase from 3 (medium) to 5 (very high) in most positive rules. The rules extracted were found to be compatible with information in [11] as well as the experts’ opinions in maintaining the CW system.

Table 5. Example of the extracted rules, where positive (+) and negative (–) rules indicate inefficient and efficient heat transfer conditions in the CW system
Table 6. Interpretation of the first positive rule and the second negative rule

| IF            | IF            |
|---------------|---------------|
| LPT A = medium to high | LPT A = very low to low |
| LPT B = medium to high | LPT B = very low |
| GEN = medium   | GEN = medium to high |
| CWIT A = medium | CWIT A = medium |
| CWIT B = low   | CWIT B = very low |
| CWOT A = low to medium | CWOT A = low |
| CWOT B = medium | CWOT B = very low |
| CWIP A = very low | CWIP A = medium |
| CWIP B = low   | CWIP B = low |
| CWOP A = very low | CWOP A = high |
| CWOP B = very low | CWOP B = very low |
| VAC = medium   | VAC = low |

THEN Heat transfer is not efficient  THEN Heat transfer is efficient

4 Summary

In this paper, a rule learning and extraction algorithm has been incorporated into FMM. The resulting network allows the extraction of fuzzy IF-THEN rules from FMM for it to explain its predictions, thus overcoming the black-box phenomenon as suffered by most neural network models. The applicability of FMM to FDD tasks has been examined. The performance was evaluated using a case study with real sensor records. The results demonstrate the potentials of FMM in learning and predicting faults in complex processes as well as in providing a comprehensible explanation for its predictions. The proposed rule extraction algorithm is able to reduce the complexity of FMM and to yield a compact rule set for justifying its predictions. The extracted rules have also been verified as meaningful and are in line with the domain knowledge as well as experts’ opinions. Further research work will concentrate on the aspect of implementation, evaluation, and validation of FMM as a usable and useful FDD tool in a variety of application domains.

References

1. R.J. Kuo, “Intelligent diagnosis for turbine blade faults using artificial neural networks and fuzzy logic”, Engng Applc Artif Intell, vol. 8, 1995, pp. 25-34.
2. J. F. G. de Freitas, I. M. MacLeod, and J. S. Maltz, “Neural networks for pneumatic actuator fault detection,” Trans of the SAIEE, vol. 90, 1999, pp. 28-34.
3. M. R. Dellomo, “Helicopter gearbox fault detection: a neural network based approach,” Journal of Vibration and Acoustics, vol. 121, 1999, pp. 265-272.
4. A. J. C. Sharkey, G. O. Chandroth, and N. E. Sharkey, “A multi-net system for fault diagnosis of a diesel engine,” Neural Computing and Applications, vol. 9, 2000, pp. 152-160.
5. L.M. Fu, “Rule learning by searching on adapted nets”, in Proc. 9th Nat. Conf. Artif. Intell., Anaheim, CA, 1991, pp. 590-595.
6. G. Towell and J. Shavlik, “The extraction of refined rules from knowledge based neural networks”, Mach. Learning, vol.13, no. 1, 1993, pp. 71-101
7. M.W. Craven and J.W. Shavlik, “Using sampling and queries to extract rules from trained neural networks”, in *Mach. Learning: Proc. 11th Int. Conf.*, San Francisco, CA, 1994.
8. P. Simpson, “Fuzzy Min-Max Neural Networks – Part 1: Classification”, *IEEE Trans on Neural Networks*, vol. 3, 1992, pp 776-786.
9. R. Andrews, J.Diederich, and A. B. Tickle, “A survey and critique of techniques for extracting rules from trained artificial neural networks”, *Knowl.-Based Syst.*, vol. 8, no. 6, pp. 373-389, 1995.
10. G. A. Carpenter and A. H. Tan, “Rule extraction: From neural architecture to symbolic representation”, *Connection Sci.*, vol. 7, 1995, pp. 3-27.
11. “System description and operating procedures,” *Prai Power Station Stage 3*, vol. 14, 1999.
12. B. Efron, “Bootstrap Methods: Another Look at the Jackknife”, *The Annals Of Statistics*, vol. 7, 1979, pp. 1-26.
Refinement of the Diagnosis Process Performed with a Fuzzy Classifier

C. D. Bocaniala¹, J. Sa da Costa², and V. Palade³

¹University “Dunărea de Jos” of Galati, Computer Science and Engineering, Department Domneasca 47, Galati 6200, Romania
Cosmin.Bocaniala@ugal.ro

²Technical University of Lisbon, Instituto Superior Tecnico, Department of Mechanical Engineering, GCAR/IDMEC, Avenida Rovisco Pais, Lisboa 1096, Portugal
sadacosta@dem.ist.utl.pt

³Oxford University, Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK
Vasile.Palade@comlab.ox.ac.uk

Abstract. This paper presents a refinement of the diagnosis process performed with a fuzzy classifier. The proposed fuzzy classifier demonstrated high accuracy in recognizing faults. In our previous work, when using this classifier, one single category has been considered for each one of the faults under observation. However, 20 levels of fault strength have been considered for each fault, ranging from small and often unnoticeable effects up to large effects. The present work proposes three categories to be considered for each fault, corresponding to small, medium and respectively large faults. Better diagnosis results are obtained. Moreover, the proposed refinement offers a new insight and more information on the behavior of the faults, that improve the final outcome of the diagnosis process.

1 Introduction

A fault diagnosis system is a monitoring system that is used to detect faults and diagnose their location and significance in a system [6]. The diagnosis system performs mainly the following tasks: fault detection – to indicate if a fault occurred or not in the system, and fault isolation – to determine the location of the fault. One of the main perspectives on fault diagnosis is to consider it as a classification problem [12]. The symptoms are extracted based on the measurements provided by the actuators and sensors in the monitored system. The actual diagnostic task is to map data points of the symptoms space into the set of considered faults.

The research literature offers three possible directions to develop fuzzy classifiers for fault diagnosis: mixtures of neural networks and fuzzy rules [5] [14], sets of fuzzy rules that describe the symptoms-faults relationships using transparent linguistic terms [8] [11], and collections of fuzzy subsets that represent the normal state and all faulty states of the system [4]. The fuzzy classifier addressed in this paper follows the third approach and it was proposed and described in detail in [2] [3]. The main advantages
of the classifier are the high accuracy with which it delimits the areas corresponding to different categories, and the fine precision of discrimination inside overlapping areas. In our earlier work [2], the parameters of the classifier have been tuned using genetic algorithms. In this paper, particle swarm optimization (PSO) is used instead of genetic algorithms.

This paper presents a refinement of the diagnosis process performed with a fuzzy classifier. In the previous work [2], when using the classifier, one single category has been considered for each one of the considered faults. The present work proposes three categories to be considered for each fault, corresponding to small, medium and, respectively, large fault strengths. This increase in the number of categories considered within the fuzzy classifier does not affect its previous good performances. Moreover, the proposed refinement offers a new insight and more information on the behavior of the faults that improve the diagnosis results.

The paper is structured as follows. Section 2 presents the main theoretical aspects of the fuzzy classifier. Section 3 briefly describes the PSO technique and how it was used in this paper. Next section, Section 4, introduces the case study - the DAMADICS benchmark [7]. Section 5 presents the better diagnosis results obtained with the proposed refinement. Finally, a brief abstract of the paper contribution and some further development directions are given in the last section.

2 The Fuzzy Classifier

The fuzzy classifier addressed in this paper has been recently introduced in [2] [3]. The classifier performs its task using a measure of similarity between points in the space associated with the problem [1]. The similarity of data points within the same category is larger than the similarity of data points belonging to different categories. The similarity between two points $u$ and $v$, $s(u,v)$, will be expressed using a complementary function, $d(u,v)$, expressing dissimilarity. The dissimilarity measure is encoded via function $h^\beta(\delta(u,v))$ that depends on one parameter, $\beta$, and that maps the distance between $u$ and $v$, $\delta(u,v)$, into $[0,1]$ interval (Eq. 1).

$$h^\beta(\delta(u,v)) = \begin{cases} \delta(u,v)/\beta, & \text{for } \delta(u,v) \leq \beta \\ 1, & \text{otherwise} \end{cases}$$

(1)

The similarity measure between two data points can be extended to a similarity measure between a data point and a subset of data points. The similarity between a data point $u$ and a subset is called the subset affinity measure. Let $C\{C_i\}_{i=1, \ldots, m}$ be the partition of a set of data points according to the category they belong to. The subset affinity measure between a data point $u$ and a category $C_i$ is given by Eq. 2, where $n_i$ denotes the number of elements in $C_i$.

$$r(u, C_i) = \frac{\sum_{v \in C} s(u,v)}{n_i} = 1 - \frac{1}{n_i} \sum_{v \in C_i} h^\beta(\delta(u,v)).$$

(2)
The effect of using the $\beta$ parameter is that only those data points from $C_i$, whose distance to $u$ is larger than $\beta$, contribute to the affinity value. The explanation is that only these points have a non-zero similarity with $u$. It follows that the affinity of a data point $u$ with different categories in the partition is decided within the neighborhood defined by $\beta$.

The natural belongingness of a data point to a category varies between a maximum value and a minimum value (corresponding to non-belongingness) and it can be approximated using the subset affinity measure. Therefore, each category $C_i$ (which represents a classical set) is replaced by a fuzzy set because the belongingness to this type of sets varies inside [0,1] interval. The fuzzy sets are induced by the corresponding categories as denoted by Eq. 3. The term $r(u,C)$ expresses the affinity of $u$ to the whole set $C$, the $n$ value represents the cardinal of $C$, and the $n_i$ value represents the cardinal of $C_i$. Notice that each category has associated a dedicated $\beta$ parameter.

$$
\mu_i(u) = \frac{\sum_{v \in C_i} h^\beta(\delta(u,v))}{r(u,C)} = \frac{n_i - \sum_{v \in C} h^\beta(\delta(u,v))}{n - \sum_{v \in C} h^\beta(\delta(u,v))}.
$$

An object $u$ presented at the input of the classifier is assigned to the category $C_i$ whose corresponding degree of assignment $\mu_i(u)$ is the largest (Eq. 4). In case of ties, the assignment to a category cannot be decided and the object is rejected.

$$
u \in z\text{-th category } \iff \mu_i(u) = \max_{j=1,\ldots,m} \mu_j(u).
$$

In order to design and test the classifier, the set of all available data for the problem to be solved is split in three distinct subsets: the reference patterns set, the parameters tuning set, and the test set. For more details on the selection methodology for the constituent elements of the previous three sets and on the role played by each of these sets, see [2]. The set of $\beta$ parameters is tuned using particle swarm optimization (PSO), described in Section 3. The time needed for tuning when using PSO is substantially smaller than the time needed in our previous work [2] that used genetic algorithms for tuning.

Using only one similarity measure does not always provide satisfactory results [2] [3]. Thus, the advantages brought by two or more similarity measures may be combined in order to improve the performance of the classifier, i.e. a hybrid approach may be used. The similarity measures may be combined following next suggestions. First, the $\beta$ parameter may be applied only to one of the similarity measures used. In this case, all subset affinity measures will be computed inside the neighborhood defined by this $\beta$ parameter. Second, there may be only one subset affinity measure resulted from the combination of all similarities used or, there may be one subset affinity measure for each similarity measure used. Finally, the fuzzy membership functions represent combinations of subset affinity measures if more than one such a measure exists.

In this paper, a hybrid approach based on Euclidean distance and Pearson correlation [16] is used. The $\beta$ parameter will be applied only to the similarity measure induced by the Euclidean distance. Two subset affinity measures are used,
based on the two similarity measures induced by the Euclidean distance and respectively Pearson correlation. Finally, the fuzzy membership functions represent the combinations of the two subset affinity measures.

3 Particle Swarm Optimization

The PSO methodology has been recently introduced in the field of Evolutionary Computing [10]. The main idea is to use mechanisms found by studying the flight behaviour of bird flocks [9]. The method may be used to solve optimization problems using the next analogy. If a roosting area is set, then the birds will form flocks and will fly towards this area, “landing” when they arrived there. The roosting area may be seen as an optimal or a near-optimal solution in the search space. The birds may represent points in the search space that will move in time towards this solution. The search process is guided by an objective function and each point is able to evaluate the value of this function (the fitness) for its current location.

The variant of PSO used in this paper starts from a set of points around origin in the $\beta$ parameters $m$-dimensional space. Repeated experiments showed that the probability to find points with large fitness around origin is very high. This means that it is very likely that the search process starts with particles located very close to the optimal solutions. The exploration of the search space combines the advantages of two techniques: the global guiding provided by the global best solution, and the local guiding provided for each particle by a small neighborhood. Using the previous combination has produced better performance than using only one singular technique. However, this is done at the cost of a larger number of iterations. The paper does not present a comparison between the performances of using single or combined techniques as it is out of the scope of this paper. For the stopping condition, it was noticed that if the global best location does not modify for a relatively small number of iterations, then this location represents an optimal solution. On the basis of this observation, the search process will be stopped if the global best location does not change for six iterations. Using the analogy from above, if the roosting is found, then the global best location will not further modify and, therefore, there is no reason to wait until all the birds landed.

4 Case Study

The DAMADICS benchmark flow control valve was chosen as the case study for this method. More information on DAMADICS benchmark is available via web [7]. The valve was extensively modeled and a MATLAB/SIMULINK program was developed for simulation purposes [15]. The data relative to the behavior of the system while undergoing a fault was generated using, as inputs to the simulation real data, normal behavior and some faulty conditions, collected at the plant. This method provides more realistic conditions for generating the behavior of the system, while undergoing a fault. It also makes the FDI task more difficult because the real inputs cause the system to feature the same noise conditions as those in the real plant. However, the resulting FDI systems will deal better when applied to the real plant.

The system is affected by a total of 19 faults. In this paper only the abrupt manifestation of the faults has been considered. A complete description of the faults
Refinement of the Diagnosis Process Performed with a Fuzzy Classifier

and the way they affect the valve can be found in [13]. There are several sensors included in the system that measure variables that influence the system, namely the upstream and downstream water pressures, the water temperature, the position of the rod, and the flow through the valve. These measurements are intended for controlling the process but they can also be used for diagnosis purposes, which means that the implementation of this sort of system will not imply additional hardware. Two of these sensors, the sensor that measures the rod position \(x\) and the sensor that measures the flow \(F\) provide variables that contain information relative to the faults. The difference \(dP\) between the upstream pressure \(P1\) sensor measurement and the downstream pressure \(P2\) sensor measurement is also considered (besides \(F\) and \(x\)), as it permits to differentiate \(F17\) from the other faults. For the rest of the faults, the previous difference has always negligible values (close to zero).

The effects of six out of the 19 faults on this set of sensor measurements are not distinguishable from the normal behavior: \(F4, F5, F8, F9, F12, F14\). Therefore, in the following, these cases are not studied. They can be dealt with if further sensors are added to the system. Also, there can be distinguished three groups of faults, \(\{F3, F6\}\), \(\{F7, F10\}\), and \(\{F11, F15, F16\}\), that share similar effects on the measurements. Due to the large overlapping, a fault member in one of the previous groups can be easily mistaken with faults in the same group. This problem is solved in recent studies by using a hybrid similarity measure based on Euclidean distance and Pearson correlation, in order to distinguish between elements in the previous three groups of faults.

5 Diagnosis Results

The 13 faults distinguishable from the normal state were simulated two times for 20 values of fault strength, uniformly distributed between 5% and 100%, and different conditions for the reference signal. For each point, a time window comprising the previous 5 seconds, and including the current point, is considered. The strength of a fault represents the intensity with which the fault acts on the valve. The previous settings approximate very well all possible faulty situations involving the 13 faults. The data obtained during the first simulation have been used to design the classifier, i.e. 50% for the \textit{REF} set, 50% for the \textit{PARAM} set [2]. The data obtained during the second simulation have been used as the \textit{TEST} set.

In previous work [2] [3], only one single category has been considered within the fuzzy classifier for the normal state and each one of the 13 visible faults. In this paper, in order to obtain a much more refined diagnosis of the faults, three categories have been considered for each one of the thirteen faults. The three categories correspond to small, medium and respectively large fault strengths, i.e. for fault \(F1\) there are three categories \(F1S\) for fault strengths between 5% and 30%, \(F1M\) for fault strengths between 35% and 60%, and \(F1L\) for fault strengths between 65% and 100%. For the subset of faults \(\{F1, F3, F6, F10, F11, F15, F16\}\), the effects of small fault strengths are not distinguishable from the normal state. Also, the effects of the medium fault strengths are not visible in the case of \(F6\) and \(F16\). The categories corresponding to previous cases have not been considered here. The confusion matrix obtained applying the classifier is shown in Table 1. The columns of the table, except the elements of the main diagonal, contain the percentage of the misclassified data, for the fault shown in the column heading. The performances per class are shown in the
main diagonal. Due to the space limitations, only a reduced confusion matrix may be shown. However, the selected categories are chosen so that they illustrate all possible situations in the original confusion matrix.

Table 1. The confusion matrix for the proposed refinement

|     | F1M | F1L | F2S | F2M | F2L | F7S | F7M | F7L | F19S | F19M | F19L | Remaining Categories |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----------------------|
| F1M | 96.0| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 4.0                   |
| F1L | 0   | 77.3| 0   | 0   | 0   | 0   | 21.2| 0   | 0   | 0   | 0   | 1.5                   |
| F2S | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 23.1                  |
| F2M | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 76.9                  |
| F2L | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0.1                   |
| F7S | 0   | 4.2 | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 87.5                  |
| F7M | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0.0                   |
| F7L | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0.0                   |
| F19S| 0   | 0   | 0   | 2.9 | 41.2| 0   | 0   | 0   | 0   | 47.0| 0   | 2.9                   |
| F19M| 0   | 0   | 0   | 4.2 | 0   | 0   | 0   | 0   | 0   | 4.2 | 0   | 91.6                  |
| F19L| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 100| 0.0                   |

Table 2. The confusion matrix without the proposed refinement

|     | F1 | F2 | F7 | F19 | Remaining Categories |
|-----|----|----|----|-----|---------------------|
| F1  | 78.2| 21.8| 0  | 0.0 |                     |
| F2  | 0   | 88.9| 0  | 9.3 | 1.8                 |
| F7  | 0   | 0   | 100| 0   | 0.0                 |
| F19 | 0   | 16.9| 81.5| 1.6 |                     |

An analysis of the results in the Table 1 reveals the next facts: (a) for fault F1 the effects for medium and large fault strengths are distinguishable one from each other (the small fault strengths are not distinguishable from the normal state), (b) for fault F7 the effects of all three subsets of fault strengths are the same, and (c) there are faults that overlap for some subsets of fault strengths: the large fault strengths of F1 overlap with F7, F2 overlaps with the small fault strengths of F19. The three situations correspond to the trends found in the sensor signals for the respective faults.

The proposed refinement offers a new insight and more information on the behavior of the faults that improve the diagnosis results. The confusion matrix obtained in previous work [2] is shown in Table 2. (This confusion matrix has been obtained using a single similarity measure based on the Euclidean distance.) The overlapping between F1 and F7, and between F2 and F19 in Table 2 are now better explained by Table 1. Fault F1L is partially misclassified as F7. The medium fault strengths of F1 are now clearly separable from all other faults. Fault F2 is misclassified only with F19S, and viceversa. The medium and large fault strengths, F19M and F19L, are now clearly separable from other faults.
The classifier performs detection and isolation in one single step. If the classifier outputs the same fault label for two consecutive seconds, then the system is diagnosed as being affected by that fault. The isolation matrix for F1, F2, F7 and F19 is given in Table 3 (small and medium fault strengths) and Table 4 (large fault strengths). The columns of the matrix stand for the fault strengths considered. The normal state (N) is separable from the faulty states (well-classified) in proportion of 99.60%.

### Table 3. The isolation matrix for small and medium fault strengths

|       | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% | 55% | 60% |
|-------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| F1    | N  | N   | N   | N   | N   | N   | N   | N   | F1M | F1M | F1M | F1M |
| F2    | N  | N   | N   | N   | F19S| F19S| F19S| F19S| F2L | F19S| F19S| F2L |
| F7    | F7 | F7  | F7  | F7  | F7  | F7  | F7  | F7  | F7  | F7  | F7  | F7  |
| F19   | N  | N   | F19S| F19S| F2L | F19L| F19L| F19L| F19L| F19L| F19L| F19L|

### Table 4. The isolation matrix for large fault strengths

|       | 65% | 70% | 75% | 80% | 85% | 90% | 95% | 100% |
|-------|-----|-----|-----|-----|-----|-----|-----|------|
| F1    | F1L | F1L | F1L | F1L | F1L | F7  | F7  | F7   |
| F2    | F7  | F7  | F7  | F7  | F7  | F7  | F7  | F7   |
| F7    | F2L | F2L | F2L | F19S| F2L | F2L | F2L | F2L  |
| F19   | F19L| F19L| F19L| F19L| F19L| F19L| F19L| F19L |

The classifier performs detection and isolation in one single step. If the classifier outputs the same fault label for two consecutive seconds, then the system is diagnosed as being affected by that fault. The isolation matrix for F1, F2, F7 and F19 is given in Table 3 (small and medium fault strengths) and Table 4 (large fault strengths). The columns of the matrix stand for the fault strengths considered. The normal state (N) is separable from the faulty states (well-classified) in proportion of 99.60%.

### 6 Conclusions

This paper proposed a refinement of the diagnosis process performed with a fuzzy classifier. Three categories, instead of only one as in our previous work, have been considered for each fault, corresponding to small, medium and, respectively, large fault strengths. This increase in the number of categories considered within the fuzzy classifier provides better diagnosis performances. The proposed refinement offers a new insight on the behavior of the faults that significantly improves the diagnosis results by reducing the overlapping areas occurring in previous work. It is to be noticed that the refinement scheme adopted in the paper, i.e. three subsets of fault strengths, 5%-30%, 35%-60% and 65%-100%, is quite rigid and it does not take into account the characteristics of the data. Further research may focus on using more flexible refinement schemes that will further reduce the overlapping areas between different faults.

### Acknowledgements

This work was supported by the European Commission’s FP5 Research Training Network Program – Project “DAMADICS – Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems”.

---

**Table 3.** The isolation matrix for small and medium fault strengths

**Table 4.** The isolation matrix for large fault strengths
References

1. Baker, E.: Cluster analysis by optimal decomposition of induced fuzzy sets (PhD thesis). Delftse Universitaire Pres, Delft, Holland (1978).
2. Bocaniala, C.D., Sa da Costa, J. and Palade, V.: A Novel Fuzzy Classification Solution for Fault Diagnosis. Paper accepted to the International Journal of Fuzzy and Intelligent Systems, IOS Press (2004).
3. Bocaniala, C.D., Sa da Costa, J. and Louro, R.: A Fuzzy Classification Solution for Fault Diagnosis of Valve Actuators. In: Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Oxford, UK, September 3-5, Part I. LNAI Series, Springer-Verlag (2003) 741-747.
4. Boudaoud, N., Masson, M.: Diagnosis of transient states using pattern recognition approach. JESA – European Journal of Automation 3 (2000) 689-708.
5. Calado, J. M. G., Korbicz, J., Patan, K., Patton, R., Sa da Costa, J. M. G.: Soft Computing Approaches to Fault Diagnosis for Dynamic Systems. European Journal of Control 7 (2001) 248-286.
6. Chen, J., Patton, R. J.: Robust Model-Based Fault Diagnosis for Dynamic Systems. Asian Studies in Computer Science and Information Science, Kluwer Academic Publishers, Boston, USA (1999).
7. European Community’s FP5, Research Training Network DAMADICS Project, http://www.eng.hull.ac.uk/research/control/damadics1.htm.
8. Frank, P.M.: Analytical and qualitative model-based fault diagnosis – a survey and some new results. European Journal of Control 2 (1996) 6-28.
9. Heppner, F., Grenander, U.: A stochastic nonlinear model for coordinated bird flocks. In: The Ubiquity of Chaos. AAAS Publications, Washington, DC (1990).
10. Kennedy, J. and Eberhart, R.: Particle Swarm Optimization. In: Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia (1995).
11. Koscielny, J.M., Syfert, M., Barty, M.: Fuzzy-logic fault diagnosis of industrial process actuators. International Journal of Applied Mathematics and Computer Science 9 (1999) 653-666.
12. Leonhardt, S., Ayoubi, M.: Methods of fault diagnosis. Control Engineering Practice 5 (1997) 683-692.
13. Louro, R.: Fault Diagnosis of An Industrial Actuator valve (MSc dissertation). Technical University of Lisbon, Lisbon, Portugal (2003).
14. Palade, V., Patton, R. J., Uppal, F. J., Quevedo, J., Daley, S.: Fault diagnosis of an industrial gas turbine using neuro-fuzzy methods. In: Preprints of the 15th IFAC World Congress, Barcelona, Spain,CD-ROM (2002).
15. Sa da Costa, J., Louro, R.: Modelling and simulation of an industrial actuator valve for fault diagnosis benchmark. In: Proceedings of the Fourth International Symposium on Mathematical Modelling, Vienna, Austria (2003).
16. Weisstein, E.W.: Correlation Coefficient. From MathWorld--A Wolfram Web Resource, http://mathworld.wolfram.com/CorrelationCoefficient.html.
ANN-Based Structural Damage Diagnosis Using Measured Vibration Data

Eric W.M. Lee and H.F. Lam

Department of Building and Construction, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong (S.A.R.), P.R. China
{ericlee, paullam}@cityu.edu.hk
http://www.cityu.edu.hk/cityu/dpt-acad/fse-bc.htm

Abstract. This paper presents the application of a novel Artificial Neural Network (ANN) model for the diagnosis of structural damage. The ANN model, denoted as the GRNNFA, is a hybrid model combining the General Regression Neural Network Model (GRNN) and the Fuzzy ART (FA) model. It not only retains the important features of the GRNN and FA models (i.e. fast and stable network training and incremental growth of network structure) but also facilitates the removal of the noise embedded in the training samples. Structural damage alters the stiffness distribution of the structure and so as to change the natural frequencies and mode shapes of the system. The measured modal parameter changes due to a particular damage are treated as patterns for that damage. The proposed GRNNFA model was trained to learn those patterns in order to detect the possible damage location of the structure. Simulated data is employed to verify and illustrate the procedures of the proposed ANN-based damage diagnosis methodology. The results of this study have demonstrated the feasibility of applying the GRNNFA model to structural damage diagnosis even when the training samples were noise contaminated.

1 Introduction

The occurrence of damage during the lifetime of all load-carrying structures such as buildings and bridges is unavoidable. To ensure the safety and so as to minimize the possible financial losses, the existence and location of structural damage must be detected as early as possible. This paper aims at the development of a practical Artificial Neural Network (ANN) based methodology for the detection of damage location at an early state following the pattern matching approach. The application of the pattern matching approach in structural damage detection was proposed by Cawley and Adams [1]. They proved that the ratio between the measured eigenvalue perturbations due to structural damage depends only on the damage location but not the damage extent. Therefore, this ratio can be treated as a “pattern” for the detection of damage location. Since only natural frequencies are used, the method fails to distinguish the...
difference among damages at symmetrical locations on the structure. Lam et al. [2] extended the work of Cawley and Adams [1], and proved that the ratio between the eigenvector perturbation and the eigenvalue perturbation due to structural damage is a function of damage location but not damage extent. Lam defined this ratio as damage signature [2], and proposed to use it as a “pattern” for damage detection following the pattern matching approach.

It is well known that the design of ANN models has significant effects on both the training of the ANN and the performance of the trained ANN. In this paper, a hybrid ANN model denoted as GRNNFA [3-7] is introduced. It employs the Fuzzy ART (FA) [8] as a pre-processor of the General Regression Neural Network (GRNN) [9] to compress the training samples to fewer numbers of representative kernels. Since the prototypes created by the FA are represented by hyper-rectangles, we propose a data compression scheme to convert the information of the prototypes to the centers, labels and widths of the kernels in the GRNN model. This compression scheme also facilitates the removal of symmetrically distributed noise embedded in the training samples. The damage detection method presented in this paper follows the approach of multidimensional classification. The damage locations were identified using the GRNNFA model [3]. Damage signatures [2] were employed as the training and testing samples. In order to demonstrate the proposed methodology, a simple numerical example about the detection of damage locations of a 10-story shear building model is given.

2 Methodology

2.1 Damage Detection Method

Damage signature [2] is employed as the patterns in the proposed structural damage diagnosis methodology. Since damage signature is not a function of fault extent, the effect of different levels of fault extent can be eliminated. As a result, the number of patterns to be considered in the damage detection process can be limited to a reasonable value. The measured damage signature is defined as:

$$\hat{\Lambda} = \frac{\Delta \Phi}{\Delta \hat{\omega}_r}$$

(1)

where $$\hat{\omega}_r$$ is the measured eigenvalue with unit (rad/s)^2 of the r-th mode and r is the reference mode number; $$\Delta \Phi$$ is a column vector of measured eigenvector perturbation for all selected modes. By computer simulation, the eigenvalue and eigenvector perturbations for different damage scenarios can be calculated. The calculated damage signatures for different damage scenarios can be calculated as:

$$\Lambda(k) = \frac{\Delta \Phi(k)}{\Delta \hat{\omega}_r(k)} \quad \text{for } k = 1, ..., K$$

(2)

where k is the damage scenario number; and K is the total number of damage scenarios to be considered. The proposed method matches the measured damage signature in Equation (1) to the K calculated damage signatures in Equation (2). The
damage scenario corresponding to the best fitted calculated damage signature is the most possible damage scenario. One difficulty in the implementation of the pattern matching approach is the lack of a systematic and intelligent way to match the measured pattern with all the calculated patterns in the database. The proposed structural damage diagnosis methodology takes the advantage of ANN to overcome this problem. There are many different ways to employ ANN in the detection of structural damages. In this paper, the ANN is trained by the calculated damage signatures in Equation (2) for all considered damage scenarios as inputs and the corresponding fault location index vector $L_i$ in Equation (3) as targets.

$$L = \{L_1, L_2, \ldots, L_i, \ldots L_{N_d}\}^T$$  \hspace{1cm} (3)

where $N_d$ is the total number of possible fault locations to be considered, and $L_i$ is the fault location index for the $i$-th possible damage location. An index of value 1 stands for a fault at the corresponding location, while 0 represents an undamaged situation. According to the generalization property of ANN, the trained ANN can be used to “estimate” (or “approximate”) the fault location index vector of the damaged structure by fitting the measured damage signature in Equation (1) to the ANN input.

### 2.2 GRNNFA Model Development

The development of the GRNNFA [3] model is based on the clustering version of the GRNN model with multiple hyper-spherical kernels [10]. The GRNNFA architecture consists of two modules i.e., FA is employed for training whereas the GRNN is employed for prediction. The basic approach of combining the GRNN and FA models is to first cluster all training samples to fewer numbers of prototypes by FA. Then, the FA prototypes are converted into the GRNN kernels. For detail formulations of the GRNNFA models, the readers may refer to reference [3]. Since each FA prototype is originally represented by two vertices of the hyper-rectangle, the scheme as detailed below is proposed to obtain the three parameters of each Gaussian kernel (i.e. center, label and width) from the respective hyper-rectangle. The kernel center is estimated by applying the FA to establish prototypes in the input domain according to the distribution of input samples. However, the prototypes created by the FA cannot be used directly as the GRNN kernels since these prototypes, in accordance with the FA learning algorithm, only represent the vertices of hyper-rectangles. As a result, the method proposed by Lim and Harrison [11] for estimating kernel centers of the prototypes is adopted in GRNNFA. The kernel label is estimated by comparing the format of the GRNN model and a statistical regression model developed from Bayesian theory. The form of the regression model is similar to that of the GRNN model. It was interpreted by Lee et al [3] that, by taking the centroid of the outputs of the samples within a prototype created by the FA to be label of the corresponding kernel of the GRNN, the GRNNFA prediction is statistically justified. Multiple hyper-spherical kernels similar to that in reference [10] are adopted in the development of the GRNNFA model. Every kernel has its own radii of spread. Hence, the total number of kernel widths to be determined equals to the total number of kernels. Traditional error gradient driven approaches become inefficient in determination of large set of
parameters. Here, K-nearest-neighbours approach is proposed to evaluate the widths of the multiple hyper-spherical kernels by determination of a single parameter. Each kernel width is determined by the scheme proposed by Lim and Harrison [11] but with the number of the nearest neighbours varied.

3 Numerical Example

3.1 Structural Modelling

A 10-story shear building model was employed to illustrate the proposed structural damage diagnosis methodology. The interstory stiffness of the “undamaged” structure at the \(i\)-th story is taken to be \(k_i = 1000 \text{ kN/m}\) for \(i = 1, \ldots, 10\). The lumped mass at each story is \(m_i = 1 \text{ kg}\) for \(i = 1, \ldots, 10\). In this example, damage at the \(i\)-th story is simulated by a 20% reduction in the corresponding interstory stiffness \(k_i\). The total number of possible damage location is 10, and the total number of damage patterns are 1023 (\(= 2^{10} - 1\)). The measured damage signature can be calculated using Equation (1). In real situation, only the first few modes can be identified from field tests. Therefore, only the first mode is employed in the calculation of damage signatures in this example. That is, the value of \(r\) in Equation (1) is equal to 1.

3.2 GRNNFA Model Training

The performances of the GRNNFA model in the application to structural damage detection were evaluated. The samples for network training were artificially generated by computer simulation. In order to study the effect of noise on the proposed methodology, different levels of noise were introduced to the input components of each sample. There were total 1023 samples generated from which 767, 128 and 128 samples were randomly extracted for network training, validation, and testing, respectively. The performance of the trained model was evaluated by the unseen noise-free testing samples. The randomization procedures involved in the generation of noise-contaminated data and selection of data for network training and testing are mitigated by the application of the bootstrapping [12] technique, which was employed to quantify the performance indicators statistically. Instead of a single run, several experiments were conducted and bootstrapping was applied to obtain the prediction errors in the form of means and 95% confidence intervals.

4 Results and Discussion

The prediction results with different noise levels are shown in Fig. 1. Each line in the figure indicates the bootstrap mean of percentage of the samples with the number of the correctly predicted locations more than that shown on the horizontal axis. It can be observed that the predicted results with different noise levels are close to each other. In general, the percentages of a given number of correct predictions for low noise levels are slightly larger than that of high noise levels. From the dashed circle, it can be observed that the percentages of 9 correct predictions for 10% and 15% noise
levels are higher than that for 5% noise level. This shows the superior performance of the GRNNFA model in handling noisy data. also shows that more than 95% of the testing samples were correctly predicted for at least 7 locations, and all testing samples were predicted correctly for at least 6 locations.

Let’s consider the situation when the actual damage is at, say, the second story and the prediction of the GRNNFA model indicates both of the second and third stories are damaged. Although the prediction at the third story is not correct, this result is on the safe side, and it is acceptable from the structural engineering viewpoint. The wrong predictions, at which some of the identified undamaged locations are, in fact, damaged, are dangerous. Here we define the so-called “safe” predictions as all the correct predictions together with the wrong predictions, at which all the identified undamaged locations are not damaged. Fig. 2 shows the predicted results of the GRNNFA model under different noise levels. There are two sets of curves shown in each of the sub-figures of Fig. 2. One set of the curves indicates the correct predictions, while the other set shows the safe predictions with 95% bootstrap confidence intervals. In Fig. 2, the 95% confidence intervals of the bootstrap means of the safe predictions are lower than that of the correct predictions for 9 or lower correct/safe predicted location. However, the numbers of samples with all correct/safe predicted locations are highly increased. Furthermore, the narrow widths of the confidence intervals as shown in Fig. 2 demonstrate the stable performance of the GRNNFA model.

5 Conclusions

Structural health monitoring using Artificial Neural Network (ANN) has attracted much attention in the last decade. However, only very little publications address the problem of ANN design. The lack of a practical and robust ANN design method is one of the main reasons for the delay in the applications of ANN in the structural damage detection industry. A new structural damage detection method, which follows the pattern matching approach, is presented in this paper. With the help of the damage signature and ANN techniques, the proposed damage detection method solves many problems of the original pattern matching approach, such as the problem of large number of damage patterns, and the lack of systematic matching method.

Apart from the damage detection method, a general ANN design method is also presented in this paper. Since the actual structural measurements are usually noise contaminated, the performance of the ANN model working in noisy environment was examined by the introduction of the Gaussian noise into the training and validation samples for network training. The trained ANN model was applied to the noise-free testing samples to evaluate the prediction error. Since the randomization processes (e.g. data extraction, noise generation and data presentation order, etc.) were involved in the course of the network training, bootstrapping technique was employed to mitigate these randomization effects and to obtain a statistically justified result.

The proposed structural damage diagnosis methodology consists of the new structural damage detection and ANN design methods. One of the objectives of this paper is to study the performance of the GRNNFA model in structural damage diagnosis following the pattern matching approach. The GRNNFA model was developed for
Fig. 1. Bootstrap mean of percentage of the samples with the number of the correctly predicted locations more than that shown on the horizontal axis

Fig. 2. Number of samples of different numbers of correct predictions and noise level
noisy data regression. It has been proven to be effective in the tasks of regression and classification in noisy environment. In the numerical example, the proposed structural damage diagnosis methodology is tested with different noise levels. The results from different noise levels are consistent, and it can be concluded that the GRNNFA prediction is not sensitive to the noise level. It is one of the outstanding advantages of the proposed methodology. Furthermore, the narrow width of the 95% confidence intervals of the bootstrap means demonstrates the stable performance of the proposed methodology. It is concluded that the proposed methodology is feasible, and further studies are required to be carried out to confirm the applicability of the proposed methodology using real experimental measurements.

References

1. Cawley, P., Adams, R. D.: The Location of Defects in Structures from Measurements of Natural Frequencies. Journal of Vibration and Acoustics 14(2) (1979) 49-57.
2. Lam, H. F., Ko, J. M., Wong, C. W.: Localization of Damaged Structural Connections based on Experimental Modal and Sensitivity Analysis. Journal of Sound and Vibration 210(1) (1998) 91-115.
3. Lee, E.W.M., Lim, C.P., Yuen, R.K.K., Lo, S.M.: A hybrid neural network for noisy data regression. IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, 34(2) (2004) 951-960.
4. Lee, E.W.M., Yuen, R.K.K., Lo, S.M., Lam, K.C. and Yeoh, G.H.: A Novel Artificial Neural Network Fire Model for Prediction of Thermal Interface Location in Single Compartment Fire. Fire Safety Journal 39 (2004) 67-87.
5. Yuen, R.K.K., Yuen, K.K.Y, Lee, E.W.M. and Cheng, G.W.Y.: A Novel Artificial Neural Network Fire Model for Determination of Thermal Interface in Compartment Fire. Proceedings of the International conference on building fire safety, November 2003, Brisbane, Australia. (2003) 25-32.
6. Lee, E.W.M., Lim, C.P. and Yuen, R.K.K.: A Novel Neural Network Model for Online Noisy Data Regression and Classification. Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation, February 2003, Vienna, Austria (2003) 95-105.
7. Lee, E.W.M., Yuen, R.K.K., Cheung, S.C.P, Cheng, C.C.K.: Application of Artificial Neural Network for Determination of Thermal Interface in Single Compartment Fire. Proceedings of the International Conference on Robotics, Vision, Information and Signal Processing, January 2003, Penang, Malaysia (2003) 588-593.
8. Carpenter, G.A., Grossberg, S., David, B.R.: Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Network 4, (1991) 759-771.
9. Specht, D.F.: A general regression neural network. IEEE Transaction on Neural Networks 2(6) (1991) 568-576.
10. Tomandl, D., Schober, A.: A modified general regression neural network (MGRNN) with new, efficient training algorithms as a robust ‘black box’ –tool for data analysis. Neural Networks 14 (2001) 1023-1034.
11. Lim, C.P., Harrison, R.F.: Modified Fuzzy ARTMAP approaches Bayes Optimal Classification Rates: An Empirical Demonstration. Neural Network 10(4) (1997) 755-774.
12. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. The Annals Of Statistics 7 (1979) 1-26.
Induction Machine Diagnostic Using Adaptive Neuro Fuzzy Inferencing System

Mohamad Shukri, Marzuki Khalid, Rubiyah Yusuf, and Mohd Shafawi
Center for Artificial Intelligence and Robotic (CAIRO),
Universiti Teknologi Malaysia,
Jalan Semarak, Kuala Lumpur,
Malaysia
marzuki@utmkl.utm.my

Abstract. Electrical machines are subjected to wear and tear after being used for sometime and proper maintenance is required to prevent breakdown. One of the main maintenance efforts is to detect fault occurring in the electrical machines. Some of these faults are slowly developing faults and early detection of these faults is crucial to prevent machine breakdown. In this paper, we investigate the effectiveness of a fault detection and diagnosis system using adaptive neuro fuzzy inferencing system (ANFIS) on a simulated three-phase induction motor. Several parameters of the induction motor are adjusted to represent faulty conditions. The experimental results obtained show that the algorithm has good fault detection and diagnosis ability.

1 Introduction

Induction motors are the workhorse of many different industrial applications due to their ruggedness and versatility. Although the induction motor is well constructed and robust, the possibility of faults is inherent due to stresses involved in the conversion of electrical to mechanical energy and vice verse. For productivity and safety reason, there has been an increasing demand for automated preventive maintenance and fault diagnostic system.

Neural networks and some related methods are already proven as a reliable technique to monitor the condition of a motor [9],[15], [3], [13] and [1]. However, it cannot provide the general heuristic or qualitative information about what contributes to faults. This inability is due to the ‘black box’ feature of the neural network. Even though neural network can perform the correct input output relationship for a given problem, it cannot perform this function in a manner which makes heuristics sense. Fuzzy logic is another method, which has been constantly used for fault detection and diagnosis [14], [8],[10],[11] and [12]. It has the capability of transforming heuristic and linguistic term into numerical values for use in complex machine computation via fuzzy rules and membership functions [16]. However fuzzy logic requires fine-tuning in order to get acceptable rule base for each faulty condition. The process can be time consuming if there are a lot of faulty conditions. Another drawback of fuzzy logic is exact solution, which is quite essential for motor fault detection is not provided due to the fuzzy nature of the solutions.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 380–387, 2004.
© Springer-Verlag Berlin Heidelberg 2004
The problems arise from fuzzy logic or neural network alone can be solved by the integration of both methods and is proven for motor fault diagnostic [5] and [6]. They used neural network to tune and optimize the fuzzy sets for the fuzzy logic algorithm. Besides that, the neural network is also used to optimize the fuzzy rule base, which is determined earlier based on the minimal knowledge of the system. However, the fuzzy rules have to be adjusted manually because trained neural network weights only show the level of relevance of the initial rules. After the rule base has been modified, the training process is repeated.

2 Adaptive Neuro Fuzzy Inferencing System (ANFIS)

The adaptive neuro fuzzy inferencing system (ANFIS) architecture is an integration of both fuzzy logic and neural network algorithm [7]. The system is neural network structure upon fuzzy logic principles, which enable the system to provide qualitative description about the motor condition and the fault detection process.

In conventional neural networks, the back propagation algorithm is used to learn, or adjust, weights on connecting arrows between neurons from input-output training samples. Refer to figure 1 below. In the ANFIS structure, the parameters of the premises and consequents play the role of weights. Specifically, when the membership functions $A^j_i$ used in the “If” part of the rules are specified parametrically that is, the shape is specified and are called premise parameters, whereas the parameters, $a_i, b_i, c_i, i = 1, 2$ in the “then” part of the rules are referred to as consequent parameters. The ANFIS learning algorithm consists of adjusting the above set of parameters from sample data

$$\left(\left(x^k_1, x^k_2\right), y^k\right), k = 1, ..., N.$$  

The effectiveness of fuzzy models representing nonlinear input-output relationships depends on the membership functions involved. Thus, the tuning of membership functions is an important issue in fuzzy modeling. This tuning task is viewed as an optimization problem, in which neural networks offer a possibility to solve this problem. In order to train a fuzzy-neural network, a set of training data is needed in the form of input-output couples, and a specification of the rules, including a preliminary definition of the corresponding membership functions. A standard approach is to assume a certain shape for the membership functions so that the membership functions depend on parameters that can be learned by a neural network.

An unknown function, or control law, to be realized by a fuzzy inference system is known only through the training set

$$\left\{\left(x^1, y^1\right),...,\left(x^K, y^K\right)\right\}$$

(1)

where $x^K = \left(x^k_1, ..., x^k_n\right) \in R$. To model the unknown function, the fuzzy “If… then…” rules $R_i, i = 1, ..., m$, of the following type is used
Fig. 1. Adaptive Neuro Fuzzy Inferencing System (ANFIS) model

\[ R_i: \text{If } x_1^k \text{ is } A_i^1 \text{ and } \ldots \text{ and } x_n^k \text{ is } A_i^n \text{ then } y = \sum_{j=1}^{n} z_i^j x_j^k + z_i \]

(2)

where \( A_i^j \) are fuzzy membership functions and \( z_i^j \) are real numbers. Let \( O^k \) be the output from the fuzzy system corresponding to the input \( x^k \). The fuzzy AND of each rule is implemented by the product, and the antecedent of the \( i^{th} \) rule is given by

\[ \alpha_i^k = \prod_{j=1}^{n} A_i^j(x_j^k) \]

(3)

The other t-norms for modeling the logical connective AND can be used. The computed output of the system is

\[ O^k = \frac{\sum_{i=1}^{m} \alpha_i^k \left( \sum_{j=1}^{n} z_i^j x_j^k + z_i^0 \right)}{\sum_{i=1}^{m} \alpha_i^k} = \frac{\sum_{i=1}^{m} \left( \prod_{j=1}^{n} A_i^j(x_j^k) \right) \left( \sum_{j=1}^{n} z_i^j x_j^k + z_i^0 \right)} {\sum_{i=1}^{m} \prod_{j=1}^{n} A_i^j(x_j^k)} \]

(4)

and define the measure of error for the \( k^{th} \) training pattern is

\[ E^k = \frac{1}{2} (O^k - y^k)^2 \]

(5)

where \( O^k \) is the computed output from the fuzzy system corresponding to the input pattern \( x^k \), and \( y^k \) is the desired output, \( k = 1, \ldots, K \). Standard neural network learning methods are used to learn \( z_i^j, j = 0,1,\ldots,n \) in the consequent part of the fuzzy rule \( R_i \).
3 Induction Motor Faults

The use of induction motor in industries is very extensive and the motors are exposed to many type of environments, miss operation and manufacturing defects. A few examples of motor fault such as short circuit of motor leads, inter turn short circuits, ground faults, worn out or broken bearings and broken rotor bars, as phase failure, asymmetric of main supply, mechanical overload and blocked rotor are inevitable [4]. Furthermore, operation within hostile environment can accelerate aging of the motor and make it more susceptible to incipient faults. This paper will look into a common induction machine fault, which is bearing failure and how this fault can be simulated for experiment.

Motor power loss, is composed of stator loss, core loss, stray loss, friction and winding loss and windage loss. The friction and winding losses are mechanical losses. Most friction is generated by the bearing, which comprises 5% to 10% of the overall losses experienced by a healthy motor. An increase in the friction is caused by many factors such as bearing wear, aging greases and lubrication leakage. This mechanical friction may cause excessive heat to burn the motor or to a stage that locks the motor. To simulate different levels of friction–related faults, the friction coefficient can be varied accordingly [2].

4 ANFIS Setup and Training

The experiment is done in MATLAB-SIMULINK environment. The model used is an asynchronous machine, which is provided in SimPowerLib block set. The model is a three phase 220 volt, 60 Hz, 1725-rpm motor configured to squirrel-cage rotor type and the reference frame is set to rotor type.

The adaptive neuro fuzzy inferencing system (ANFIS) function in Fuzzy Logic Toolbox is used as the neuro-fuzzy engine. The ANFIS function is configured to two inputs: for stator current and shaft speed. Only one phase current is measured because the experiment is done on balanced magnetic field. The effect of the simulated faults is similar on the other phase current. The ANFIS output is configured to a motor

![ANFIS network and structure](image-url)
condition, which is partition into three fuzzy singletons: good, fair and bad. The ANFIS network and rules are shown in figure 2. Three gaussian fuzzy sets are used to partition the inputs to low, medium and high as shown in figure 3. The inferencing rules are selected according to basic and minimal knowledge known from the system.

**Fig. 3.** Input fuzzy sets for stator current and shaft speed

**Fig. 4.** Training data for the ANFIS engine

**Fig. 5.** Input fuzzy membership for stator current and shaft speed after training
The data used to train the adaptive neuro fuzzy inferencing system is shown in figure 4 below. With the error tolerance set to 0.001, the training data has optimized the input fuzzy membership for current and speed through 2000 epoch and the result is shown in figure 5.

5 Experiment and Result

The objectives of this experiment are to test the performance and capability of the proposed fault diagnostic method and show how the method is being applied. Bearing fault is simulated by changing the value of friction coefficient. Initially, the motor starts on normal condition where the friction coefficient value is very low. When the bearing fault starts, the friction coefficient increases and this will affect its speed and phase current. During medium friction coefficient value, the phase current increases while the shaft speed reduces. When the fault is getting worse, the phase current increases to a higher value while the shaft speed reduces to a lower value. The simulated phase current and shaft speed data for the experiment are shown in figure 6.

The trained ANFIS engine is tested with an experimental data. This data is generated by simulating the asynchronous motor model with different friction coefficient value, which is configured to low, medium and high. Each of these conditions is divided into two levels, small (1) and big (2), in order to generate the effect of slowly developing bearing fault.

From the experiment result shown in figure 7, the trained ANFIS engine is able to identify the motor operating condition. During normal operation where the friction is very low, the ANFIS engine defines the induction machine is in a good condition. Although the friction starts to increase, within acceptable value, it still remains good. However, when the friction increases higher, the ANFIS alarmed with fair condition, and to a certain tolerance it remains. But this is the stage where the machine needs to be repaired. When the friction increases to a dangerous level the ANFIS detect and define the induction machine condition as bad. The result also shows a small variation between the actual output and the experiment (testing) output during good and fair conditions. This is because the shaft speed and phase current data are measured and experimented during transient period.

![Fig. 6. Phase current and shaft speed experimental data](image-url)
6 Conclusion

A neural fuzzy system, adaptive neuro fuzzy inferencing system (ANFIS) has been presented to perform three-phase induction motor fault diagnostic. The ANFIS provide quantitative description of the motor faults under different operating conditions as well as qualitative heuristic explanation of these operating conditions through fuzzy rules and membership functions. In this paper, ANFIS is being used to predict the motor friction coefficient value from the measured phase current and shaft speed data. The changing value of friction coefficient determines that the motor is in faulty condition. This incipient mechanical problem is detected at very early stage and precaution measure can be taken immediately.

Fig. 7. Experiment result for good, fair and bad motor condition

References

1. Chow, M. –Y. and S. –O. Yee (1991). “Using Neural Network to Detect Incipient Faults in Induction Motor” Journal of Neural Network Computing 2(3): 26-32
2. Chow. M –Y (1997). Methodologies of Using Neural Network and Fuzzy Logic Technologies for Motor Incipient Fault Detection. World Scientific. Singapore
3. Demian, C.; Cirrincione, G.; Capolino, G.A.(2002). “A neural approach for the fault diagnostics in induction machines” IECON Industrial Electronics Society, IEEE 2002 28th Annual Conference, Nov. 2002 Pg:3372 - 3376 vol.4
4. Gupta, B. K. and I. M. Culbert (1992), “Assessment of Insulation condition in rotating machine stators”, IEEE Transaction on Energy Conversion 7(3)
5. Goode, P. and M. –y. Chow (1995). “Using a Neural/Fuzzy to Extract Knowledge of Incipient Fault in Induction Motor Part 1 – Methodology.” IEEE Transaction on Industrial Electronics 42(2): 131-138
6. Goode, P. and M. –y. Chow (1995). “Using a Neural/Fuzzy to Extract Knowledge of Incipient Fault in Induction Motor Part 2 – Application.” IEEE Transaction on Industrial Electronics 42(2): 139-146
7. J.S.R. Jang (1993). “ANFIS: adaptive –network-based fuzzy inference system”, IEEE Transaction on System, Man and Cybernetics 23(3) 665-684
8. Kiuipel, N.; Frank, P.M.(1993) “Process supervision with the aid of fuzzy logic” Systems, Man and Cybernetics Conference Proceedings., International Conference on, 17-20 Oct. 1993 Pg: 409 - 414 vol.2
9. McCulloch, W.S. and W. Pitts (1943). “A logical calculus of the ideas immanent in nervous activity.” Bulletin of Mathematical Biophysics. 127-147
10. Shukri Z.A. (2001). Design and Development of Intelligent Model Based Fault Diagnostic Algorithm Based on Parameter Estimation and Fuzzy Logic Technique. Master thesis. University of Technology, Malaysia.
11. Shukri Z.A, Marzuki K., Rubiyah Y., Shamsuddin M.A. (2002) “Model Based Fault Detection and Diagnosis Using Parameter Estimation And Fuzzy Inference”, Asean Control Conf. (ASCC) 2002, Singapore
12. Shukri Z.A., Rubiyah Y., Marzuki K., Shamsuddin M.A. (2002) “Application of Model Based Fault Detection and Diagnosis Using Parameter Estimation and Fuzzy Inference to a DC-Servomotor”, IEEE International Symposium on Intelligent Control (ISIC ’02) 2002, Vancouver Canada.
13. Stefano, R.; Meo, S.; Scarano, M. (1994) “Induction motor faults diagnostic via artificial neural network” ISIE ’94 IEEE International Symposium on Industrial Electronics, 25-27 May 1994 Pg:220 - 225
14. Sauter, D.; Mary, N.; Sirou, F.; Thieltgen, A (1994) “Fault diagnosis in systems using fuzzy logic”, Proceedings of the Third IEEE Conference on Control Applications, 24-26 Aug. 1994 Pg:883 - 888 vol.2
15. Tallam, R.M.; Habetler, T.G.; Harley, R.G.(2002) “Self-commissioning training algorithms for neural networks with applications to electric machine fault diagnostics”. Power Electronics, IEEE Transactions on, Vol: 17 Issue: 6, Nov. 2002 Pg:1089 - 1095
16. Zadeh, L.A. (1965). Fuzzy Sets. Information and Control. New York, Academic Press. 8: 338-353
Real Time Stokes Inversion Using Multiple Support Vector Regression

David Rees¹,²,³, Ying Guo¹, Arturo López Ariste³,⁴, and Jonathan Graham³

¹ Centre for Intelligent Systems Design, CSIRO ICT Centre, PO Box 76, Epping, NSW 1710, Australia
² School of Mathematics and Statistics, University of Sydney, NSW 2006 Australia
³ High Altitude Observatory, NCAR** PO Box 3000, Boulder, CO-80305, USA
⁴ THEMIS-CNRS UPS 853 C/Via Lactea sn. 38200 La Laguna, Canary Islands, Spain

Abstract. Solution of the Stokes inversion problem to estimate the vector magnetic field distribution on the sun from the profiles of the Stokes parameters of polarised light in magnetically sensitive spectral lines is a vital task in our understanding of solar activity. Recently machine learning techniques such as principal component analysis and neural networks have led to the development of real time inversion computer codes. This paper presents a novel alternative approach to real time inversion called Multiple Support Vector Regression which is applied here for the first time to synthetic Stokes profile data.

1 Introduction

The magnetic field that permeates the external layers of the sun plays a fundamental role in solar activity. Estimation of the magnetic field distribution near the solar surface is done indirectly using spectropolarimetry, i.e. measurement of the wavelength dependence of the Stokes parameters (or Stokes profiles) of polarised radiation in magnetically sensitive spectral lines. The solution of the inverse problem to infer the vector magnetic field from Stokes profile data is known as Stokes inversion (Socas-Navarro [1]). Modern spectropolarimeters provide accurate measurements of the Stokes profiles of many different spectral lines formed at various atmospheric heights. Future space- and ground-based instruments, such as Solar-B and SOLIS, will achieve unprecedented spatial resolution and coverage of the solar surface. The expected flood of data from such instruments has recently been the catalyst for the development of several new approaches to Stokes inversion based on machine learning, aimed at real time data analysis.

Initial steps towards real time inversion were made by Rees et al [2, 3] who proposed a database search method using principal component analysis (PCA).
Socas-Navarro et al. [4] and López Ariste et al. [5] showed that so-called PCA inversion is over two orders of magnitude faster than traditional nonlinear least squares model fitting (Auer et al. [6]). Essentially real time inversion has been achieved using multilayer perceptron neural networks (Carroll and Staude [7]; López Ariste et al. [5]; Socas-Navarro [8]). In parallel with this PCA and neural network research we have been investigating an alternative approach which we call multiple support vector regression (MSVR) (Rees and Guo [3]) based on support vector machines (SVMs) (Vapnik [10]; Cristianini and Shawe-Taylor [11]). MSVR is the focus of the current paper.

The rest of the paper is structured as follows. Section 2 summarises the MSVR method and Section 3 illustrates its application to synthetic unpolarised intensity profile data. Section 4 presents the first application of MSVR to synthetic Stokes profile data. We conclude in Section 5, setting the research agenda for the next stage of development of MSVR.

2 Multiple Support Vector Regression

In brief the inversion problem can be formulated as follows. Suppose we measure an $N$ dimensional signal $S = (S_1, \ldots, S_N)$ and associate with this signal a set of parameters $p = (p_1, p_2, \ldots)$. Thinking of $S$ as an operator (generally nonlinear) on $p$, the goal is to find the inverse operator $\mathcal{F}$ such that

$$p = \mathcal{F}(S(p))$$

(1)

To approximate $\mathcal{F}$ we use a training set of signals $S_j = S(p_j), j = 1, \ldots, M$ corresponding to $M$ different parameter sets $p_j$. In many cases $N$ is large and it is advantageous to reduce dimensionality by PCA, reconstructing $S$ using eigenvectors estimated from this training set (Rees et al. [2][3]). Then instead of $S$ we can work with the vector $E = (e_1, \ldots, e_n)$ of eigenfeatures or principal components, where $n \ll N$, and the inversion problem can be recast as finding $\mathcal{F}$ such that

$$p = \mathcal{F}(E(p))$$

(2)

For a model with $L$ physical parameters each eigenfeature vector $E_i$ in the training set has an associated parameter set $p_i = (p_{i1}, \ldots, p_{ik}, \ldots, p_{iL})$. In order to estimate a particular parameter $p_k$, we organise the $M$ training examples as $(E_1, p_{i1}), \ldots, (E_i, p_{ik}), \ldots, (E_M, p_{Mk})$, regarding $E_i$ as input vectors and $p_{ik}$ as the associated output values for application of the SVR algorithm.

The goal of SVR is to find a function $f_k(E)$ such that

$$|f_k(E_j) - p_{jk}| \leq \epsilon, \text{ for } j = 1, \ldots, M,$$

(3)

where $\epsilon > 0$. Thus the function value $f_k(E_j)$ has at most $\epsilon$ deviation from the actually obtained targets $p_{jk}$ for all the training data, and at the same time, is as smooth as possible. The SVR function has the form:
\[
f_k(E) = \sum_{i=1}^{M} \alpha_{ik} K_k(E_i, E) + b_k. \tag{4}
\]

where \(\alpha_{ik}\) and \(b_k\) are constants, and \(K_k\) is the kernel function. The index \(k\) emphasises that one is free to choose different kernel functions for different system parameters.

For some cases linear SVR may be sufficient, but in general nonlinear SVR is desired. In the latter case a number of kernel functions have been found to provide good performance, including polynomials, radial basis functions (RBF), and sigmoid functions. The SVR optimisation problem is then solved in accordance with standard techniques (see, for example, Cristianini and Shawe-Taylor [11]).

The regression functions \(p_k = f_k(E)\), for \(k = 1, \ldots, L\), learned by this process, constitute the inverse operator \(F\) in equation (2).

3 Application to Unpolarised Spectra

We now illustrate the method using synthetic unpolarised intensity profiles modelled analytically by

\[
I = 1 + \frac{1}{1 + \eta_0 e^{-(x/\delta)^2}} \tag{5}
\]

where \(x\) is a dimensionless wavelength measured from line centre. This model has two adjustable parameters: \(\eta_0\) which is the line to continuum opacity ratio at line centre and \(\delta\) which controls the line broadening. In terms of the previous notation, \(p = (\eta_0, \delta)\) and \(S\) is composed of values of \(I\) sampled at \(N\) values of the wavelength \(x\). The goal is to find the regression functions \(\eta_0 = f_{\eta_0}(E)\) and \(\delta = f_{\delta}(E)\).

3.1 One Parameter Case

Fixing \(\delta = 1\), we generated a training set of \(M = 19\) profiles using the opacity values \(\eta_0 = 1 : (0.5) : 10\), i.e. from 1 to 10 in steps of 0.5. The profiles were computed at \(N = 61\) wavelengths \(x = -3 : (0.1) : 3\) and just two eigenfeatures, i.e. a 2-dimensional eigenfeature vector \(E = (e_1, e_2)\) was used. The training data were fitted with a polynomial kernel. One result of the fitting is automatic selection of the number of support vectors required for the SVR function. In this case there are 7 support vectors. The SVR function is a smooth interpolating function which can be used for accurate parameter estimation for any eigenfeature vector, not just those in the training set. The training set and the SVR function (a smooth interpolating curve) are shown in Figure 1.

Synthetic test data were generated for a large number of values of \(\eta_0\). These test data were used as “observations” and the parameter values estimated with the SVR function. The errors in these estimated \(\eta_0\) were found to be less than 1% for all test data.
3.2 Two Parameter Case

Here we allow both parameters to vary, generating a training set of \( M = 121 \) profiles for \( \eta_0 = 1 : (0.9) : 10 \) and \( \delta = 0.5 : (0.1) : 1.5 \) for \( N = 81 \) wavelengths \( x = -4 : (0.1) : 4 \). We used a 3-dimensional eigenfeature vector \( E = (e_1, e_2, e_3) \) and fitted the regression functions with an RBF kernel.

The number of support vectors defining the regression functions \( f_{\eta_0}(E) \) and \( f_{\delta}(E) \) were 83 and 72 respectively. The training data and the regression functions (smooth interpolating surfaces) viewed as functions of \( e_1 \) and \( e_2 \) are shown in Figure 2.

Synthetic test data were again generated for a large number of parameter values and the regression functions were used to estimate the parameters from these “observations”. The errors in these estimates were found to be less than 1.3% for \( \eta_0 \) and less than 0.3% for \( \delta \) for all test data.

4 Application to Polarised Spectra

We now apply MSVR to invert Stokes profiles. For simplicity we consider only the spectral profiles of intensity \( I \) and net circular polarisation \( V \). A training set of \( M = 399 \) synthetic \( I \) and \( V \) profiles sampled at 100 wavelengths was generated for a magnetically sensitive spectral line of neutral iron by solving the equations of polarised radiative transfer in a model of the solar photosphere for a range of magnetic field strengths \( B = 0 : (100) : 2000 \) G (Gauss), and inclinations \( \gamma = 0 : (5) : 90 \) degrees to the line of sight; the field azimuth was not varied. Thus in this model the parameter vector is \( p = (B, \gamma) \) and the signal vector is a 200-dimensional concatenation of the \( I \) and \( V \) profiles, which, on applying PCA
Fig. 2. Unpolarised SVRs (smooth interpolating surfaces) $\eta_0 = f_{\eta_0}(E)$ (left) and $\delta = f_\delta(E)$ (right). The level curves are defined by the training data separately to $I$ and $V$ and retaining only the first two eigenfeatures for each, leads to a composite 4-dimensional eigenfeature vector, $E = (e_1, e_2, e_3, e_4)$.

The goal is to estimate the SVR functions $B = f_B(E)$ and $\gamma = f_\gamma(E)$. We fitted the regression functions with an RBF kernel. The number of support vectors varied depending on selection of certain fitting criteria in the SVR algorithm, but averaged about 80. Here we present only the results for $f_B(E)$. Training data and regression function (smooth interpolating surface) are shown in Figure 3 as functions of $e_1$ and $e_2$.

As in the unpolarised case synthetic test data were generated and used as “observations”. The errors in the estimates of $B$ from these data, viewed as an error surface in Figure 4, were less than 10G in magnitude, well under the errors typically found in analysis of real observational data.

Fig. 3. Polarised training data viewed as a dark mesh superposed on SVR function $B = f_B(E)$ (smooth interpolating surface) for magnetic field strength
5 Conclusion

MSVR provides explicit functional representations of model parameters, as does nonlinear regression by a multilayer perceptron neural network, and thus is suitable for real time implementation. The very preliminary tests with synthetic Stokes data in this paper indicate that MSVR will indeed work for Stokes inversion, but much more research and testing are required to decide whether MSVR is a viable alternative to neural network inversion. Issues to be addressed include how best to form the composite signal and associated eigenfeature vectors, especially when all four Stokes profiles are involved, i.e. linear as well as circular polarisation are treated simultaneously.

Neural network inversion is currently emerging as the method of choice for on-board real time data processing, for example on the Helioseismic Magnetic Imager (HMI) experiment on the Solar Dynamics Observatory mission to be launched in 2007. It is worth noting here that HMI is a filtergraph instrument and samples the Stokes spectra at only a small number of wavelengths. Graham et al. [12] showed that even with such limited wavelength coverage it is possible to obtain reliable vector magnetic field estimates by traditional inversion using nonlinear least squares model fitting. Obviously in this case the signal data already is low dimensional and the PCA compression step discussed in this paper is not necessary. It will be interesting to investigate the application of MSVR to such data.

6 Acknowledgements

David Rees gratefully acknowledges the generous support of the Affiliate Scientist Program at the High Altitude Observatory.
References

1. Socas-Navarro H.: Stokes inversion techniques: Recent achievements and future horizons, in: Advanced Solar Polarimetry – Theory, Observation and Instrumentation, ed. M. Sigwarth, (2001)
2. Rees, D. E., Mulcahy, D., and Thatcher, J.: A database technique for fast spectral line inversion, Proc. IASTED Int. Conference on Signal Processing and Communications, Canary Islands (1998) 408–411
3. Rees D., López Ariste A., Thatcher J., Semel M.: Fast inversion of spectral lines using principal component analysis. I. Fundamentals, A&A 355 (2000) 759–768.
4. Socas-Navarro H., López Ariste, A., Lites B. W.: Fast inversion of spectral lines using principal component analysis. II. Inversion of real Stokes data, ApJ 553 (2001) 949-954.
5. López Ariste A., Rees, D., Socas-Navarro, H., Lites, B.W.: Pattern recognition techniques and the measurement of solar magnetic fields, Proc.SPIE Vol 4477 Astronomical Data Analysis, San Diego, August (2001) 96–106.
6. Auer, L. H., Heasley, J. M., House, L. L.: The determination of vector magnetic fields from Stokes profiles. Solar Phys. 55 (1977) 47–61
7. Carroll, T., and Staude, J.: The inversion of Stokes profiles with artificial neural networks. A&A 378 (2001) 316 – 326
8. Socas-Navarro, H.: Measuring solar magnetic fields with artificial neural networks. Neural Networks 16 (2003) 355–363
9. Rees, D.E., and Guo, Y.: Ghosts in the machine: from turbulence, faces and ducks to solar magnetic fields, in: Trujillo Bueno J., Sanchez Almeida J. (eds.), Proceedings of the Solar Polarisation Workshop 3, Tenerife, Oct 2002, ASP Conf. Series (2003)
10. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer, NY (1995)
11. Cristianini, N., and Shawe-Taylor: An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK (2000)
12. Graham, J.D, López Ariste, A., Socas-Navarro H., Tomczyk, T.: Inference of solar magnetic field parameters from data with limited wavelength sampling. Solar Phys. 208 (2002) 211–232
Extracting Stellar Population Parameters of Galaxies from Photometric Data Using Evolution Strategies and Locally Weighted Linear Regression*

Luis Alvarez¹, Olac Fuentes¹, and Roberto Terlevič¹²

¹ Instituto Nacional de Astrofísica Óptica y Electrónica, Luis Enrique Erro # 1 Santa María Tonantzintla, Puebla, 72840, México
lochoa@ccc.inaoep.mx, {fuentes, rjt}@inaoep.mx
² Institute of Astronomy, University of Cambridge
Madingley Road, Cambridge CB3 0HA, UK

Abstract. There is now a huge amount of high quality photometric data available in the literature whose analysis is bound to play a fundamental role in studies of the formation and evolution of structure in the Universe. One important problem that this large amount of data generates is the definition of the best procedure or strategy to achieve the best result with the minimum of computational time.

Here we focus on the optimization of methods to obtain stellar population parameters (ages, proportions, redshift and reddening) from photometric data using evolutionary synthesis models. We pose the problem as an optimization problem and we solve it with Evolution Strategies (ES). We also test a hybrid algorithm combining Evolution Strategies and Locally Weighted Linear Regression (LWLR). The experiments show that the hybrid algorithm achieves greater accuracy, and faster convergence than evolution strategies. On the other hand the performance of ES and ES-LWLR is similar when noise is added to the input data.

1 Introduction

The main aim of this work is to explore automatic techniques for obtaining stellar population parameters (spp) from photometric data of galaxies (pd). Given the huge amount of information available in the form of photometric data, for example Sloan Digital Sky Survey [1], it is necessary to test faster data analysis methods.

Using evolutionary algorithms in different problems of astronomy was suggested in [5]. In [10] genetic algorithms were used to predict parameters of interacting galaxies. The analysis of stellar spectra with evolution strategies is

* This work was partially supported by CONACyT (the Mexican Research Council) under grants 177932 and J31877A.
presented in [8]. The extraction of spp has had several approaches. Some methods for determining age and reddening of stellar clusters are reviewed in [9]. Neural networks have also been used for calculating photometric redshifts [6].

Here we consider the problem as an optimization one, where finding a solution is difficult due to the number of variables involved and the level of noise in the data. Evolution Strategies (ES) seem to perform well in these particular conditions but, for problems with many dimensions ES can be slow. To speed up the convergence we have combined ES with Locally Weighted Linear Regression (LWLR). LWLR creates local linear models of a function around a query point. We used the candidate solutions found in each iteration of the ES for predicting another solution, possibly better, that will be included in the solution set.

This paper is structured as follows: Section 2 describes the procedure for creating data. Section 3 briefly summarizes the methods. Section 4 presents the results of experiments and Section 5 shows the conclusions.

## 2 Data

We have utilized a set of theoretical synthetic spectra, $F$ of simple stellar populations along isochrones of given age and metallicity. They were computed for solar metallicity $Z = 0.02$ with logarithmic ages of: 6, 8, 8.3, 8.6, 9, 9.6, 9.78, 10, 10.2 yr. The resolution is 20Å and all were computed for a Salpeter Initial Mass Function.

Using this set we then form a set of synthetic galactic spectra [4]. First we separated the nine spectra into three age groups: the first contains one spectrum of a young stellar population, the second contains four intermediate-age spectra and the third has four old spectra. In the first step of the procedure we only have to specify $F_2$ and $F_3$. Since the proportions must sum 1, only $p_1$ and $p_2$ need to be specified.

We normalized the spectra, each component of a normalized spectrum $F_{i\text{norm}}$ being given by $F_{i\lambda\text{norm}} = F_{i\lambda}/\sum_{\lambda} F_{i\lambda}$, where $\lambda \in [890 - 23010\text{Å}]$. For clarity we now change $F_{i\text{norm}}$ to $F_i$.

The procedure for forming the galactic spectra and the photometric data is:

1. Randomly select populations $F_i$ and proportions $p_i$.
2. Combine three spectra, $F_i$, of different ages at given proportions $p_i$.
   \[ F_{\text{combined}} = F_1p_1 + F_2p_2 + F_3p_3 . \] (1)
3. Apply a simplified model of reddening, $R$, to $F_{\text{combined}}$
   \[ F_{\lambda,\text{reddened}} = F_{\lambda,\text{combined}} - F_{\lambda,\text{combined}} \left( \frac{kR}{\lambda} \right) . \] (2)
4. Apply redshift, $Z$, to $F_{\text{reddened}}$ according to the formula
   \[ \lambda = \lambda_0(Z + 1) . \] (3)

We obtain a new spectrum, $F_{\text{redshifted}}$. 
Fig. 1. Three spectra from which we construct a galactic spectrum. These are chosen from a grid of nine spectra.

5. Finally we divide, $F_{\text{redshifted}}$, into fifteen intervals of the same width and we average their fluxes in each interval, simulating wide-band filters.

To summarize, the six $spp$ we wish to extract are $[F_2, F_3, p_1, p_2, R, Z]$. Figure 1 shows three selected spectra, and Figure 2 depicts their combination following the procedure just described.

3 Method

In this section we describe ES, LWLR and the ES-LWLR hybrid in the context of the $spp$ problem. Our main aim is to test the combination of ES with LWLR for speeding up the convergence into a suitable solution. This hybrid approach is based on the idea that it is possible to use the solutions generated in each iteration of the ES to form a local linear model of the objective function. For this purpose, we chose LWLR [2]. We then use the current model to predict another, hopefully better, solution. The predicted solution is added to the current population. A similar technique, but in conjunction with Active Learning is reported in [7].

3.1 Evolution Strategies

Among the several varieties of ES [3] we selected the $(\mu + \lambda)$-ES because it adapts better to the requirements of the hybrid algorithm. In this variant of
Fig. 2. A spectrum formed from the spectra of Figure 1 according to the procedure described in Section 2. The parameters used in this case are: $F_2 = 1 \times 10^8 \text{yr}$, $F_3 = 3 \times 10^{10} \text{yr}$, $p_1 = 0.23$, $p_2 = 0.37$, $R = 0.3$ and $Z = 0.2$. The photometric data indicated by triangles result from dividing the spectrum into fifteen intervals of equal width and averaging their corresponding fluxes.

ES, the candidate solutions are chosen from a population formed by $\mu$ parents and $\lambda$ offspring. These results in the population concentrating around one solution allowing the fit of finer models using LWLR. Other variants like $(\mu, \lambda)$-ES that only choose their candidate solutions from $\lambda$ offspring, are slower and their solutions are more sparse, although they are generally more noise tolerant.

How ES implements the principles of evolution in optimization is explained next: Given the problem of finding the $x$ that minimizes the function $f(x)$ in a domain, $M$, where $x$ satisfies the constraints $g_i(x) > 0$, $i \in \{1, \ldots, k\}$, ES perform the following process in order to find a solution.

1. Generate a set of $\mu$ random solutions called initial population, $P^0$.
2. Recombine the population, applying operator $r$, for creating $\lambda$ offspring.
3. Mutate the offspring, applying operator $m$.
4. Select from the total population of parents and offspring the best $\mu$ solutions (the new population), and eliminate the rest.
5. Add a new individual applying the LWLR algorithm to the current population (hybrid part).
6. Go back to the second step if the termination criterion, $t$, is not satisfied.

The hybrid step 5 will be explained in subsection 3.2. All components of the ES are summarized in the tuple $(P^0, \mu, \lambda; r, m, s; \Delta \sigma; f, g, t)$.

The recombination and mutation operators (also called genetic operators) work on extended vectors, $a = (x, \sigma)$, where $\sigma$ has the same length as $x$, and contains the standard deviations for carrying out the mutation.

The recombination operator, $r$, produces one individual, $a'$, by selecting randomly with uniform probability two elements, $a_a$ and $a_b$, from the current population $P$, and mixing them. We use discrete recombination in the $x$’s, which
consists of randomly selecting with uniform probability the components of the offspring from the two parents, (see Equation (5)). For the $\sigma$’s, average recombination is used, (see Equation (6)). The recombination operator resembles the sexual reproduction in biological organisms.

$$r(P^t) = a' = (x', \sigma')$$ .
$$x'_i = x_{i,b} \text{ or } x_{i,a}$$ .
$$\sigma'_i = \frac{1}{2}(\sigma_{i,a} + \sigma_{i,b})$$ .

The mutation operator, $m$, randomly changes the vector $a'$, as occurs in nature. This brings improvements in the population, i.e., better adapted individuals (solutions), those that produce smaller objective function values than past generations did.

$$m(a') = a'' = (x'', \sigma'')$$ .
$$\sigma''_i = \sigma'_i \exp N_0(\Delta\sigma)$$ .
$$x''_i = x'_i + N_0(\sigma''_i)$$ .

$N_0(\sigma)$ is a random number generator from a normal distribution with mean 0 and standard deviation $\sigma$. $\Delta\sigma$ is a meta-parameter that controls the rate of change of the $\sigma$’s.

The selection operator $s$ evaluates the objective function, $f$, for the total population and chooses the best $\mu$ (those that evaluate $f$ to the smallest absolute values), that will in turn form the new generation of solutions. In our problem the $x$ part of the extended $a$ vector corresponds to the $\text{spp}$. The objective function, $f$, is the sum of the quadratic differences among the photometric data given as query, and the photometric data produced by one solution of the population. The stopping criterion, $t$, finalizes the algorithm at the 50th iteration (generation), although it could be also a minimum error or an elapsed time.

### 3.2 Locally Weighted Linear Regression

As mentioned at the beginning of Section 3, the solutions generated in each generation of the ES could be used to form a linear model for predicting another solution. The information available for creating the model are the ordered pairs $(\text{spp}, \text{pd})$. The $\text{spp}$ are generated randomly by the ES and the photometric data $\text{pd}$ are calculated by means of the procedure outlined in Section 2. In this particular case, we need to predict the $\text{spp}$, so we reverse the pairs to obtain $(\text{pd}, \text{spp})$. The linear model is represented by Equation (10), with unknown coefficients, $\beta$.

$$\text{PD} \beta = \text{SPP}$$ .

PD and SPP are matrices that contain all the pairs of a generation. The unknown vector, $\beta$, is found by minimizing the least squares criterion

$$C = \sum_{\text{for each row } i} (\text{PD}_{i}^T (\beta - \text{SPP}_{i})^2$$ .


The solution is also called normal equations

$$\beta = (PD^T PD)^{-1} PD^T SPP .$$

(12)

Vector $\beta$ is used to predict any $spp$ given a new set of $pd$ (here named $pd_{query}$), at the accuracy allowed by a global linear model. The accuracy is improved if we make local models, around the $pd_{query}$, instead of global models. We have achieved this using proportionally weighting the data $(PD, SPP)$ according to the Euclidean distance between each row $PD_i$ and the $pd_{query}$. In this way, the $PD_i$ near $pd_{query}$, will have greater influence in the model than farther $PD_i$.

$$Z = W \cdot PD .$$

(13)

$$V = W \cdot SPP .$$

(14)

Equations (13) and (14) weight the data by multiplying them by the diagonal matrix $W$, containing the inverse Euclidean distances of the $pd_{query}$ to each row vector in $PD$. Replacing these weighted variables in the normal equations, $\beta$ can be determined. An $spp$ is estimated given a $pd_{query}$

$$\beta = (Z^T Z)^{-1} Z^T V .$$

(15)

$$spp(pd_{query}) = pd_{query}^T (Z^T Z)^{-1} Z^T V .$$

(16)

Equation (16) is included in step 5 of the ES algorithm, in this way we construct the hybrid algorithm ES-LWLR.

4 Results

We have run a series of tests in order to critically assess the performance of ES and ES-LWLR. We adopted as the measure of quality the mean absolute errors (MAE) of the $spp$ and the execution time of the algorithms. The parameters for both algorithms are the same: population $\mu = 50$, number of offspring $\lambda = 100$, $\Delta \sigma = 0.25$, stopping criterion $t = 50$ generations for each $spp$.

The different tests arise from variations of the procedure described in Section 2. The first test set consisted of 100 random sets of $spp$ generated according to the mentioned procedure (see Section 2). We then reduced the resolution to just 100 pixels by uniformly sampling the original 1107 pixels spectra. Two additional sets come from the addition of Gaussian noise to both low and high resolution spectra after we have applied the redshift to them. The aggregated noise has mean zero and standard deviation equal to 0.1 times the maximum flux of the spectrum. Thus we have a total of 400 $spp$ for testing. We test both algorithms over these 400 $spp$, giving a total of eight test.

As can be seen in Table 1 the hybrid has better performance. Figure 3 shows the differences between some given spectra and their predicted spectra. On low resolution spectra, again, the hybrid achieves better accuracy (see Table 1). After
adding noise the accuracy reached by both algorithms is almost the same (see Table 1).

Because the domains of each stellar population parameter are different: $F_2$ and $F_3 \in [0, 4]$, $p_1, p_2$ and $R \in [0, 1]$ and $z \in [0, 2]$. Then the meaning of the MAE’s of each parameter (see Table 1) depends on their respective domain.

In order to evaluate the efficiency of the proposed method we compare the rate of convergence of the objective function, $f$ (see Section 3.1). This function is implemented in the algorithms. The hybrid algorithm converge in less generations than the ES (see Figure 4). Table 2 shows how the convergence time is reduced for the hybrid when the stopping criterion is a given error.

![Fig. 3. Differences between spectra, formed from the spp of the Figure 2, and their predicted spectra. (a) and (b) correspond to high resolution spectra with and without noise respectively. (b) and (c) belong to low resolution spectra with and without noise respectively.](image)

5 Conclusions

We approached the problem of extracting stellar population parameters from photometric data as an optimization problem and we demonstrated that it can be solved with Evolution Strategies. We have also shown that the combination of ES with Locally Weighted Regression speeds up the convergence. This is due to the use of the solutions generated in each iteration of ES to form a linear model; this linear model is then used to predict a better solution. Our experiments
### Table 1. Mean absolute error of spp over 100 spectra (h=high, l=low, n=no, y=yes)

| Algorithm | Resolution/Noise | $F_2$ | $F_3$ | $p_1$ | $p_2$ | $p_3$ | $R$ | $Z$ |
|-----------|------------------|-------|-------|-------|-------|-------|-----|-----|
| ES        | h/n              | 0.67  | 1.03  | 0.0342| 0.048 | 0.0544| 0.0959| 0.0254|
| ES-LWLR   | h/n              | 0.12  | 0.28  | 0.0004| 0.0017| 0.0017| 0.0006| 0.0003|
| ES        | l/n              | 0.58  | 1.14  | 0.039 | 0.0479| 0.0524| 0.0758| 0.027 |
| ES-LWLR   | l/n              | 0.09  | 0.24  | 0.0005| 0.0007| 0.001 | 0.0005| 0.003 |
| ES        | h/y              | 0.62  | 1.19  | 0.0344| 0.0529| 0.587 | 0.1001| 0.0373|
| ES-LWLR   | h/y              | 0.89  | 0.85  | 0.0411| 0.0481| 0.0541| 0.0847| 0.389 |
| ES        | l/y              | 0.94  | 1.08  | 0.0696| 0.0936| 0.0833| 0.1641| 0.0731|
| ES-LWLR   | l/y              | 0.8   | 1.06  | 0.574 | 0.0888| 0.0949| 0.1461| 0.0654|

### Table 2. Mean absolute error of spp over 100 high resolution spectra. Stopping criterion error=1e-10

| Algorithm | $F_2$ | $F_3$ | $p_1$ | $p_2$ | $p_3$ | $R$    | $Z$   | $t$ (seg) |
|-----------|-------|-------|-------|-------|-------|--------|-------|-----------|
| ES        | 0.87  | 1.07  | 0.0366| 0.0450| 0.0560| 0.0922 | 0.0289| 2017      |
| ES-LWLR   | 0.1   | 0.45  | 0.0005| 0.003 | 0.0029| 0.0017 | 0.0006| 843       |

show that in noiseless data, the hybrid algorithm considerably reduces the time of computing with respect to ES when the stopping criterion is a given error (see Table 2 and Figure 4). The calculated MAE's of each stellar population parameter show that the hybrid algorithm has better accuracy than the ES, in
noiseless spectra, as was expected. On the other hand the performance of the hybrid algorithm after adding 10% Gaussian noise to the spectra is comparable to that of ES alone.

References

1. K. Abazajian, et al.: The first data release of the Sloan Digital Sky Survey. The Astronomical Journal. 126 (October 2003) 2081–2086
2. C.G. Atkinson, A.W. Moore, and S. Schaal.: Locally weighted learning. Artificial Intelligence Review. 11 (1997) 11–73
3. T. Bäck, F. Hoffmeister, H.P. Schwefel.: A survey of evolution strategies. Proceedings of the 4th International Conference on Genetic Algorithms, San Diego, CA. (1991) 2–9
4. A. Bressan, C. Chiosi, F. Fagotto.: Spectrophotometric evolution of elliptical galaxies. 1: Ultraviolet excess and color-magnitude-redshift relations. The Astrophysical Journal. 94 (1994) 63–115
5. P. Charbonneau. 1995.: Genetic algorithms in astronomy and astrophysics. Astrophysical Journal, Supplement Series. 101 (December 1995) 309–334
6. A.E. Firth, O. Lahav, R.S. Somerville.: Estimating photometric redshifts with artificial neural networks. Monthly Notices of the Royal Astronomical Society. 339 (March 2003) 1195-1202
7. O. Fuentes, T. Solorio.: Interferogram analysis using active instance-based learning. IASTED International Conference Artificial Intelligence and Applications, Benalmádena, Málaga, Spain. (2003) 386–390
8. J.F. Ramírez, O. Fuentes.: Spectral analysis using evolution strategies. IASTED, International Conference on Artificial Intelligence and Soft Computing, Banff, Alberta, Canada. (July 2002) 208–213
9. J.F.C. Santos, E. Bica.: Reddening and age for 11 Galactic open clusters from integrated spectra. Monthly Notices of the Royal Astronomical Society. 206 (1993) 915–924
10. M. Wahde.: Determination of orbital parameters of interacting galaxies using a genetic algorithm. Astronomy and Astrophysics Supplement Series. 132 (November 1998) 417–429
Using Evolution Strategies to Find a Dynamical Model of the M81 Triplet

Juan Carlos Gomez¹, Olac Fuentes¹, Lia Athanassoula², and Albert Bosma²

¹ Instituto Nacional de Astrofísica Óptica y Electrónica., Luis Enrique Erro # 1, Tonantzintla, Puebla, 72840, México
   {jcgc, fuentes}@inaoep.mx
² Observatoire de Marseille, 2 Place le Verrier, 13248, Cedex 4, Marseille, France

Abstract. In this work we present Evolution Strategies (ES) as an efficient optimization method for dynamic modelling of the main interacting group of three galaxies in M81. The M81 group is one of the nearest groups of galaxies; its biggest galaxy, M81, sits in the core of the group together with its two companions M82 and NGC3077. The interaction among these three galaxies is very well defined in an image taken in HI. In this first attempt we use non-self-gravitating simulations for modelling dynamically the group; even with this restriction our method reproduces the density distribution of the three galaxies with great precision. Results presented here show that ES is an excellent method to find an accurate model of groups of interacting galaxies, where a global search for a large number of real-valued parameters needs to be performed.

1 Introduction

It is now established that galaxies are not "island universes", but rather interact in pairs or in small or big groups. Interactions can form spirals, bars, warps, rings and bridges. Thus, observing the morphological and kinematic results of an interaction can give us crucial information about the interaction scenario.

Given the enormous time scales involved in galactic interactions, it is impossible to observe directly the evolution of a system. Instead, we use a combination of theoretical models, spectroscopy, and a single image, in order to determine the set of initial conditions and the interaction time that result in the current configuration.

The problem of finding the right parameters for modelling the interaction of a given system of galaxies can be posed as an optimization problem [5]. As input, we have an image of the interacting system and sometimes a velocity field, obtained from spectroscopy. To construct a dynamical model, we need to simulate the system of galaxies, giving a set of initial conditions and interaction time to a simulation program. These initial conditions are the basis to understand the dynamical nature of the system. Then the simulation program gives a projected surface density map and line-of-sight velocities. These can be compared to the corresponding observed quantities, and then the best model is the one
that minimizes the difference between maps. In the end, we have a set of initial conditions that given to a simulator program, can reproduce the interaction among galaxies.

In this work we use Evolution Strategies (ES) as the optimization algorithm to find the minimum difference between simulated and observational images. ES is a good method that works efficiently as a global search algorithm with continuous parameter spaces. Since most of the parameters of interacting systems of galaxies are continuous, this constitutes a clear incentive for trying out ES.

In this work we chose the M81 triplet as the interacting system to be studied. The M81 group is one of the nearest groups of galaxies. Its biggest galaxy, M81, sits in the core of the group together with its two nearby companions M82 (in the upper part of the image in Figure 3) and NGC3077 (in the lower part of the image in Figure 3). This group has a very well defined interaction scenario; the main galaxy has a spiral shape and forms clear tails with its two companions. Also, the interaction is only in the outer part of the galaxies, which facilitates the use of non-self-gravitating simulation. All of this made this group an ideal candidate to test ES.

The organization of the remainder of this paper is as follows: Section 2 contains a brief description of the method, the implementation is presented in Section 3, the results are given in Section 4, and conclusions and future work are presented in Section 5.

2 The Method

Evolution Strategies (ES) is a technique for finding the global minimum of a function with a large number of variables in a continuous space. We start by choosing K individuals, each characterized by an object parameter vector \( \mathbf{O} \) and a corresponding strategy parameter vector \( \mathbf{S} \):

\[
\mathbf{O}_i = \langle q_{1,i}, q_{2,i}, \cdots, q_{L,i} \rangle \quad i = 1, \cdots, K
\]

\[
\mathbf{S}_i = \langle \sigma_{1,i}, \sigma_{2,i}, \cdots, \sigma_{L,i} \rangle \quad i = 1, \cdots, K
\]

In the first generation, the elements of the \( \mathbf{O} \) and \( \mathbf{S} \) vectors, can be chosen either totally at random, or with help from previous knowledge about the problem to be solved. Each of the \( K \) individuals (set of parameters) must be evaluated according to a fitness function. The fitness function is what we need to minimize.

The next step is to produce a new population applying the genetic operators cross-over and mutation. For cross-over, two individuals (parents) are chosen at random, and then we create two new individuals (offspring) by combining the parameters of the two parents. Mutation is applied to the individuals resulting from the cross-over operation; each element of the new individual is calculated from the old individual using the simple equation:

\[
q_{j,\text{mut}} = q_j + \mathcal{N}(0, \sigma_j)
\]
where $N(0, \sigma_j)$ is a random number obtained from a normal distribution with zero mean and standard deviation $\sigma_j$, which is given from the strategy parameter vector. The process of cross-over and mutation is repeated until the population converges to a suitable solution.

3 Modelling the Interacting Group M81

In the beginning we had one image of the group of galaxies taken in HI (neutral hydrogen) by Yun [6], in addition, we had some physical information about the group, also from Yun. We used the HI image to approximate the model because in that image the interaction among galaxies is clearly defined. The HI image was translated and resized in such a way that M81, the main galaxy, was centered in the image. Then, we calculated the distances (in pixels) between the central points of each pair of galaxies.

The next step was to define the density map of the image. It was done simply by putting a grid over the image and counting the pixels that have a value greater than 0 in each cell of the grid. Then we established a relation between the total amount of pixels in the image and the number of particles we are going to use in the simulations, in such a way that we have a portion of the total mass in each cell. With this we have a density matrix that we can use to measure the fitness of the simulations, as was established by [5]. In this case we have used a 48x48 grid. As an example, Figure 1 shows a 6x6 mass density matrix of an artificial image that represents two interacting galaxies. The values in each cell represent the number of particles.

![Fig. 1. Artificial image and its density matrix](image)

Figure 2 shows a diagram of the solution process. First we create the individuals with ES, then each individual is used as an input for the simulator program, the program returns a mass distribution that is used to be compared with the
observational data; if the fitness is good enough or the maximum number of iterations has been reached, the process stops, otherwise we create a new population using genetic operators and return to step 1.

Fig. 2. Block diagram of the solution process

Applying ES to approximate a model for the group of galaxies M81 was done as follows: we use a population with 30 individuals per iteration \((K = 30)\), where each individual has the following structure:

\[
\mathbf{O}_j = \langle z_2, z_3, V_{x1}, V_{y1}, V_{x2}, V_{y2}, V_{x3}, V_{y3}, i_1, PA_1, i_2, PA_2, i_3, PA_3, \ldots, m_1, m_2, m_3, t \rangle, j = 1, 2, \ldots, K
\]

\[
\mathbf{S}_j = \langle \sigma_{i,1}, \sigma_{i,2}, \sigma_{i,3}, \sigma_{i,4}, \sigma_{i,5}, \sigma_{i,6}, \sigma_{i,7}, \sigma_{i,8}, \sigma_{i,9}, \sigma_{i,10}, \sigma_{i,11}, \sigma_{i,12}, \sigma_{i,13}, \ldots, \sigma_{i,14}, \sigma_{i,15}, \sigma_{i,16}, \sigma_{i,17}, \sigma_{i,18} \rangle, j = 1, 2, \ldots, K
\]

where \(z_2\) and \(z_3\) represent the distances in the line of sight between the main galaxy in the center of the image and its companions; \(V_x\) and \(V_y\) the velocities in the image plane; \(i\)'s the inclination angles (in the \(x\) axis); \(PA\)'s the position angles (in the \(z\) axis); \(m\)'s are the masses of the galaxies and \(t\) represents the total time of interaction. Subindex 1 is for M81 galaxy, 2 for M82 and 3 for NGC3077.

The first generation is created using random values, but with heuristic physical information as reference[6]: first, the masses and the time can not be negative (for physical reasons); since the main galaxy is perturbed only in the outer part, and the two companions are compact, then the main galaxy must have a predominant mass; separations in the line of sight must be small enough to allow perturbations in the galaxies; velocities in the image plane must be also on a range that allows the galaxy discs to be perturbed.

Each individual in the population, in this case each set of initial conditions, is used as input for the simulator program. With the simulation we obtain a surface
density map, and multiply it by a mass scale factor. Then the result of each simulation is evaluated with a fitness function. The fitness function compares the density maps of the simulated and original images using the Kullback-Leibler distance [2]:

$$F_d = \left[ \sum \left\{ \left( m_{i,j}^{obs} + m_\epsilon \right) \ln \left( \frac{m_{i,j}^{obs} + m_\epsilon}{m_{i,j}^{sim} + m_\epsilon} \right) \right\} \right]^{-1}$$  \hspace{1cm} (4)

where \( m_{i,j}^{sim} \) is the total mass in the cell under consideration for the simulation, \( m_{i,j}^{obs} \) is the same quantity for the observations, \( m_\epsilon \) is a very small quantity to avoid problems in regions with zero density, and the sum is carried over all the cells.

Once we have evaluated the first generation, it is necessary to create a new population using the genetic operators, as described in the previous section. Following the ideas described in [1], we use dynamic mutation and first mutate the strategy parameter vector using the equation:

$$\sigma_{mut} = \sigma \exp \left[ \frac{C_v \delta R_1}{\sqrt{2K}} + \frac{C_v \delta R_2}{\sqrt{2K}} \right]$$  \hspace{1cm} (5)

where \( \sigma \) is the value before mutation, \( R_1 \) and \( R_2 \) are random numbers obtained from a normal distribution with zero mean and a standard deviation equal to the \( \sigma \) before mutation, and \( C_v \) and \( \delta \) are numerical constants. We found the values \( C_v = 0.9 \) and \( \delta = 0.1 \) to be the most appropriate for this particular problem.

The crossover operator is uniform: two individuals are randomly selected from the original population in such a way that each individual is used exactly once, and each parameter in the two individuals has the same probability to be selected to form two new individuals.

The mutation operator is performed by following the classical process: adding a random number obtained from a normal distribution with zero mean and standard deviation \( \sigma \) (taken from strategy parameter vector \( S \)).

To form the children population we first apply crossover to the whole population, and then mutation to the resulting population. Then we merge both populations (parent and children), select the \( K \) best individuals from this merged population, and use that set as input for the next iteration.

For the simulation we use the test particle approach [5]. In this approximation, the mass of each galaxy is assumed to be concentrated in a single point in its center, while the disc, which responds most to the interaction, is represented by test particles, initially in co-planar circular orbits around the center of the galaxy. This approach is very fast and thus allows us to run the very large number of simulations necessary for tackling this problem. Furthermore, in our case, the galaxies are not inter-penetrating and thus they are perturbed only in their outer parts, making the test particle approach fairly adequate. We used a total of 4000 particles, 2000 for the main galaxy and 1000 for each companion.
4 Results

After 200 generations of evolution, yielding a total of 6000 simulations, we obtained a good artificial image, that matches very well the original image. The images in Figure 3 show the best simulation reached and the original HI image. Table 1 shows the corresponding parameters for the best simulation. The total time required to reach that model was 12 hours in a Sun Ultra 10 workstation.

In Table 1 index 1 corresponds to M81, 2 to M82 and 3 to NGC3077. The quantities are given in program units and we used the following scale to convert to physical units: the unit of length is taken to be 1kpc, the unit of time $1 \times 10^6$ yr, the unit of mass $1 \times 10^{10} \text{M}_\odot$ and the angles are given in degrees.

Table 1. Parameters to produce the simulation in Figure 1

| $z_2$ | $z_3$ | $V_{z1}$ | $V_{z2}$ | $V_{z3}$ | $V_{x1}$ | $V_{x2}$ | $V_{x3}$ | $PA_1$ | $PA_2$ | $PA_3$ | $m_1$ | $m_2$ | $m_3$ | $t$ |
|-------|-------|----------|----------|----------|----------|----------|----------|--------|--------|--------|-------|-------|-------|------|
| 62.05 | 11.14 | 3.17     | 1.59     | 61.91    | 41.78    | -168.75  | -0.90    | 44.90  | 113.37 | 38.62  | 32.58  | 53.83  | 232.87 | 19.47 | 1.04 | 1.06 | 812 |

In the simulated image we can see the spiral shape of the main galaxy M81, the tail that joins M81 with NGC3077 in the lower part of the image, the tail in the upper part of NGC3077, and part of the mass concentration in the upper part of M81.

The density was fairly well reproduced in this way, we reached a maximum of 0.45 with the fitness function (1 corresponds to a perfect match). Obviously, reaching a perfect match between images is not possible because of the limited number of particles used in simulations.

In this first attempt we used only data from the density map, without considering the velocity distribution, so the resulting velocities were inaccurate. To test the velocity match, we took some real velocity data in the line of sight from certain parts of the system, but in a comparison with the simulated velocity field, it does not match very well. So, in future work we are planning to introduce that velocity information to improve the velocity estimation.

5 Conclusions and Future Work

In this work we presented an efficient method, based on ES, to approximate a dynamical model for the M81 triplet.

Even with the several simplifying assumptions done in simulations, searching with ES has demonstrated to be an excellent method for optimization problems where a global exploration of continuous parameters spaces is needed. ES could find a set of parameters that results in a very good match to the HI density distribution in this problem. We are planning to extend the application of ES to the study of other interacting systems.

On the other hand, and in order to improve the method, the possibility of implementing a parallelization of the ES could be considered with the purpose of reducing the computing time required for simulations. We are planning to include
physical information about velocity, trying to reproduce the velocity field of the image. Also, methods based on self-gravitating N-body simulations can be used to improve the match between simulations and the HI density distribution.

Also we will implement new hybrid algorithms to accelerate the convergence. Two main algorithms are under consideration: using ES in combination with a traditional optimization algorithm, such as Newton’s method, and combining ES with Locally Weighted Linear Regression. Another possibility is trying to solve this problem with a different optimization algorithm, such as Simulated Annealing.

References

1. Back, T., Hoffmeister, F., Schwefel, H. P., A Survey of Evolution Strategies. Proceedings of the Fourth International Conference on Genetic Algorithms. (1991) 2–9
2. Kullback, S., Leibler, On Information and Sufficiency. R. A., Ann. of Math. St. (1951) 79–86
3. Rechenberg, I., Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Stuttgart: Fromman-Holzboog. (1973)
4. Wahde, M., Donner, K. J., Determination of the Orbital Parameters of the M51 System Using a Genetic Algorithm. Astronomy and Astrophysics. (2001) 115–124
5. Wahde, M., Determination of Orbital Parameters of Interacting Galaxies Using a Genetic Algorithm. Astronomy & Astrophysics Supplement Series. (1998) 417–429
6. Yun, M., Tidal Interactions in M81 Group. IAU Symp. 186, Galaxy Interactions at Low and High Redshift. (1999) 81

Fig. 3. Simulated and HI images for M81 group
Automated Classification of Galaxy Images

Jorge de la Calleja and Olac Fuentes

Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro 1, Tonantzintla 72840, Puebla, México
jorge@ccc.inaoep.mx, fuentes@inaoep.mx

Abstract. In this paper we present an experimental study of the performance of three machine learning algorithms applied to the difficult problem of galaxy classification. We use the Naive Bayes classifier, the rule-induction algorithm C4.5 and a recently introduced classifier named random forest (RF). We first employ image processing to standardize the images, eliminating the effects of orientation and scale, then perform principal component analysis to reduce the dimensionality of the data, and finally, classify the galaxy images. Our experiments show that RF obtains the best results considering three, five and seven galaxy types.

1 Introduction

The morphology of galaxies is generally an important issue in the large scale study of the Universe. Galaxy classification is the first step towards a greater understanding of the origin and formation process of galaxies, and the evolution processes of the Universe [10]. Galaxy classification is important for two main reasons. First, to produce large catalogues for statistical and observational programs, and second for discovering underlying physics [7].

In recent years, with numerous digital sky surveys across a wide range of wavelengths, astronomy has become an immensely data-rich field. For example, the Sloan Digital Sky Survey [1] will produce more than 50,000,000 images of galaxies in the near future. This overload creates a need for techniques to automate the difficult problem of classification. Several methods have been used to solve this problem, such as neural networks [4, 5, 6, 9, 10, 13], oblique decision trees [11], ensembles of classifiers [2, 4], and instance-based methods [4].

We propose an approach to perform galaxy classification that first generates a representation that is independent of scale and orientation, then generates a more compact and manageable representation using principal component analysis, and finally classifies the galaxy images using machine learning algorithms. In previous work [4], we used locally-weighted regression and neural networks to perform galaxy classification, and now, we investigate the performance of three other learning algorithms: the Naive Bayes classifier, the rule-induction algorithm C4.5 and the random forest (RF) predictor. We also use ensembles of these algorithms to classify the images.
The paper is organized as follows: Section 2 gives a brief introduction of the Hubble tuning fork scheme for galaxy classification. In Section 3 we describe the general architecture of the method, including the image analysis, data compression and learning stages. In Section 4 we show experimental results and finally in Section 5 conclusions and future work are presented.

2 The Hubble Tuning Fork Scheme

Galaxies are large systems of stars and clouds of gas and dust, all held together by gravity. Galaxies have many different characteristics, but the easiest way to classify them is by their shape; Edwin Hubble devised a basic method for classifying them in this way. In his classification scheme, there are three main types of galaxies: Spirals, Ellipticals, and Irregulars (Figure 1).

Elliptical galaxies (E0, E4 and E7 in Figure 1) have the shape of an ellipsoid. Spiral galaxies are divided in ordinary and barred; ordinary spirals have an approximately spherical nucleus, while barred spirals have a elongated nucleus that looks like a bar. Spirals are classified as Sa, Sb, or Sc; barred spirals are labeled as SBa, SBB, or SBC. The subclassification (a, b or c) refers both to the size of the nucleus and the tightness of the spiral arms. An Sa galaxy has a bigger nucleus than an Sc galaxy, and the arms of the Sc are wrapped more loosely. S0 are spiral galaxies without any conspicuous structure in their disks. Irregular galaxies do not have an obvious elliptical or spiral shape.

3 The Classification Method

The method that we developed for galaxy classification is divided in three stages: image analysis, data compression, and machine learning (see Figure 2). The method works as follows: It takes as input the galaxy images, which are then rotated, centered, and cropped in the image analysis stage. Next, using principal component analysis, the dimensionality of the data is reduced and we find a set of features. The projection of the images onto the principal components will be
the input parameters for the machine learning stage. At the end, we will have the classification of the galaxies. The following three subsections describe each part in detail.

![Galaxy images](image)

**Fig. 2.** The stages of the classification method

### 3.1 Image Analysis

Galaxy images generally are of different sizes and color formats, and most of the time the galaxy contained in the image is not at the center. So, the aim of this stage is to create images invariant to color, position, orientation and size, all in a fully automatic manner. First, we find the galaxy contained in the image applying a threshold; that is, from the original image $I$, we generate a binary image $B$, such that

$$B(i, j) = \begin{cases} 1 & \text{if } I(i, j) > \text{threshold;} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Then we obtain $\bar{i}$ and $\bar{j}$, the center row and column of the galaxy in the image, given by

$$\bar{i} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} iB(i, j) \quad (2)$$

$$\bar{j} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} jB(i, j) \quad (3)$$

where $m$ and $n$ are the number of rows and columns, respectively, in the image. Then we obtain the covariance matrix of the points in the galaxy image

$$C = \sum_{i=1}^{m} \sum_{j=1}^{n} B(i, j)[i - \bar{i}, j - \bar{j}]^T[i - \bar{i}, j - \bar{j}] \quad (4)$$

The galaxy’s main axis is given by the first eigenvector (the eigenvector with the largest corresponding eigenvalue) of $C$, the covariance matrix. We then rotate the image so that the main axis is horizontal. The angle is given by

$$\alpha = \arctan(p1(1)/p1(2)) \quad (5)$$
where \( p1(1) \) and \( p1(2) \) are the \( x \) and \( y \) values of the first principal component. Then we use an image warping technique to rotate the image (see Figure 3). After that, we crop the image, eliminating the columns that contain only background (black) pixels. Finally, we stretch and standardize the images to a size of 128x128 pixels. Figure 4 shows examples of the image processing stage for an elliptical galaxy, a spiral galaxy, and an irregular galaxy.

![Figure 3](image)

**Fig. 3.** Left: The first principal component (p1) is used to rotate the galaxy image. Right: Rotated galaxy

### 3.2 Data Compression

Principal component analysis (PCA) is a statistical method that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components (PCs). PCA is generally used to reduce the dimensionality of a data set while retaining as much information as possible. Instead of using all the principal components of the covariance matrix, we may represent the data in terms of only a few basis vectors. We used 8, 13 and 25 PCs to perform the classification because they represent about 75%, 80% and 85% of the information, respectively, in the data set. More details about this technique can be found in [14].

### 3.3 Machine Learning

**Naive Bayes Classifier.** The Naive Bayes classifier [8] is a probabilistic algorithm based on the assumption that the attribute values are conditionally independent given the target values. The Naive Bayes classifier applies to learning tasks where each instance \( x \) can be described as a tuple of attribute values \( a_1, a_2, \ldots, a_n \) and the target function \( f(x) \) can take on any value from a finite set \( V \). When a new instance \( x \) is presented, the Naive Bayes classifier assigns to it the most probable target value by applying the rule:

\[
f(x) = \underset{v_j \in V}{\text{argmax}} P(v_j) \Pi_i P(a_i \mid v_j)
\]

To summarize, the learning task of the Naive Bayes is to build a hypothesis by estimating the different \( P(v_i) \) and \( P(a_i \mid v_j) \) terms based on their frequencies over the training data.
Fig. 4. Examples: a) Original images, b) Rotated images, and c) Cropped images

C4.5. This method operates by recursively splitting a training set based on feature values to produce a tree such that each example can end up in only one leaf. An initial feature is chosen as the root of the tree, and the examples are split among branches based on the feature value for each example. If the values are continuous, then each branch takes a certain range of values. Then a new feature is chosen, and the process is repeated for the remaining examples. Then the tree is converted to an equivalent rule set, which is pruned. For a deeper introduction of this method we refer the reader to [8] and [12].

Random Forest Predictor. A random forest (RF) is a classifier consisting of a collection of individual tree classifiers. Basically, random forest does the following:

1. Select \( n_{\text{tree}} \), the number of trees to grow, and \( m_{\text{try}} \), a number no larger than the number of variables.
2. For \( i = 1 \) to \( n_{\text{tree}} \):
3. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
4. Grow a "random" tree, where at each node, the best split is chosen among \( m_{\text{try}} \) randomly selected variables. The tree is grown to maximum size and not pruned back.
5. Use the tree to predict out-of-bag data.
6. In the end, use the predictions on out-of-bag data to form majority votes.
7. Prediction of test data is done by majority votes from predictions from the ensemble of trees.

Details about RF can be found in [3].

Ensemble Method. An ensemble consists of a set of classifiers whose individual decisions are combined in some way, normally by voting, to classify new
examples. The ensemble method used here is bagging \cite{5}. It was chosen because this method almost always improves the accuracies obtained by individual classifiers. The idea in this ensemble is to generate randomly \( n \) training sets with the examples from the original training set, and to use each of these subsets for creating a classifier. Each subset is obtained by sampling, with replacement, from the original training set, thus some of the examples will appear more than once, while others will not appear at all.

4 Experimental Results

We test our method with 292 galaxy images. Most of them were taken from the NGC catalog on the web page of the Astronomical Society of the Pacific, and their classification was taken from the interactive NGC online catalog. For our purpose we consider three (E, S, Irr), five (E, S0, Sa+Sb, Sc+Sd, Irr) and seven (E, S0, Sa, Sb, Sc, Sd, Irr) galaxy types.

We used the Naive Bayes classifier, J48 (a particular C4.5 implementation) and the random forest classifier that are implemented in WEKA, and also the bagging ensemble method. We used 10-fold cross-validation for doing all the experiments. For C4.5 we used pruning and a confidence factor of 0.25. In the case of RF, 13 trees were used for creating the forest for all the experiments, however, we select different random features, i.e. five for the three-class case and two for five and seven classes.

| Table 1. Accuracy for individual classifiers and ensembles |
|-----------------|-----------------|-----------------|-----------------|
|                 | Naive Bayes     |                 |                 |
|                 | 3 classes       | 5 classes       | 7 classes       |
| PCs             | Ind std Ens std | Ind std Ens std | Ind std Ens std |
| 8               | 83.23 0.7819    | 86.77 0.7893    | 46.00 0.7819    |
| 13              | 80.68 0.7101    | 85.61 0.7283    | 44.53 1.1984    |
| 25              | 75.88 0.5156    | 82.73 1.7241    | 40.33 1.1987    |

|                 | C4.5            |                 |                 |
|                 | 3 classes       | 5 classes       | 7 classes       |
| PCs             | Ind std Ens std | Ind std Ens std | Ind std Ens std |
| 8               | 88.35 0.6388    | 86.77 0.6983    | 46.00 0.7819    |
| 13              | 87.39 0.2893    | 91.09 1.1865    | 46.77 1.9618    |
| 25              | 86.84 1.0167    | 91.02 0.6582    | 45.81 1.2463    |

|                 | Random Forest   |                 |                 |
|                 | 3 classes       | 5 classes       | 7 classes       |
| PCs             | Ind std Ens std | Ind std Ens std | Ind std Ens std |
| 8               | 90.39 1.0208    | 91.22 0.3086    | 47.18 1.4414    |
| 13              | 91.29 0.6077    | 91.64 0.1917    | 49.72 2.0179    |
| 25              | 91.29 0.5180    | 91.64 0.5737    | 47.87 1.9263    |

Table 1 shows the accuracy for each of the individual classifiers, and for the ensembles, and we also show the standard deviation. The accuracies were

1 www.apsky.org/ngc/ngc.html
2 www.seds.org/~spider/ngc/ngc.html
3 WEKA is a software package that can be found at www.cs.waikato.ac.nz/ml/weka
obtained by averaging the results of 5 runs of 10-fold cross validation for each method. The columns Ind, Ens and std denote individual classifier, ensemble of classifiers and standard deviation, respectively.

Analyzing the results, we can observe that RF obtained the best accuracy for all the galaxy classes, i.e. 91.64% accuracy for the three-class case, 54.72% accuracy for the five-class case, and 48.62% accuracy for the seven-class case; and the standard deviations were almost always the smallest. Only in the seven-class case, Naive Bayes obtained a smaller standard deviation than RF with 0.5725, but its accuracy was of 43.62%. We can also note that in all cases ensembles obtained better results than individual classifier. Examining the results considering the number of PCs, we can say that 13 are enough to perform the classification, obtaining good results. This way we can reduce computation by using few attributes.

5 Conclusions

We presented a method that performs morphological galaxy classification in a fully automatic manner producing good results. The use of standardized images helps to improve the accuracy of the learning algorithms. We have shown experimentally that a small number of principal components is enough to classify the galaxies. Also, the ensemble permits to improve the classification accuracy.

Future work includes testing this method for other types of astronomical objects, such as nebulas and clusters, and extending the system to deal with wide-field images, containing multiple objects.

References

1. Ball, N. Morphological Classification of Galaxies Using Artificial Neural Networks. Master's thesis, University of Sussex, 2002
2. Bazell, D., Aha, D.W. Ensembles of Classifiers for Morphological Galaxy Classification. The Astrophysical Journal, 548:219-233, 2001
3. Breiman, L. Random Forests, Machine Learning, 45(1), 5-32, 2001
4. De la Calleja, J., Fuentes, O. Machine learning and image analysis for morphological galaxy classification, Monthly Notices of the Royal Astronomical Society, 349:87-93, 2004
5. Dietterich, T.G. Machine Learning Research: Four Current Directions. AI Magazine, 18(4):97-136, 1997
6. Goderya, S. N., Lolling, S.M.. Morphological Classification of Galaxies using Computer Vision and ANNs. Astrophysics and Space Science, 279(377), 2002
7. Lahav O. Artificial neural networks as a tool for galaxy classification, in Data Analysis in Astronomy, Erice, Italy, 1996
8. Mitchell, T. Machine Learning. McGraw Hill, 1997
9. Madgwick, D.S. Correlating galaxy morphologies and spectra in the 2dF Galaxy Redshift Survey. Monthly Notices of the Royal Astronomical Society, 338:197-207, 2003
10. Naim, A., Lahav O., Sodré, L. Jr., Storrie-Lombardi M.C. Automated morphological classification of APM galaxies by supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society, 275(567), 1995

11. Owens, E.A., Griffiths, R.E., Ratnatunga K.U. Using Oblique Decision Trees for the Morphological Classification of Galaxies. Monthly Notices of the Royal Astronomical Society, 281(153), 1996

12. Quinlan, J.R. Induction of decision trees. Machine Learning, 1(1):81-106, 1986

13. Storrie-Lombardi, M.C., Lahav, O., Sodré, L., Storrie-Lombardi, L.J. Morphological Classification of Galaxies by Artificial Neural Networks. Monthly Notices of the Royal Astronomical Society, 259(8), 1992

14. Turk, M.A., Pentland, A.P. Face Recognition Using Eigenfaces, in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 586-591, 1991
Automatic Solar Flare Tracking

Ming Qu\textsuperscript{1}, Frank Shih\textsuperscript{1}, Ju Jing\textsuperscript{2}, Haimin Wang\textsuperscript{2}, and David Rees\textsuperscript{3}

\textsuperscript{1} College of Computing Sciences, New Jersey Institute of Technology
Newark, NJ 07102
\textsuperscript{2} Center for Solar-Terrestrial Research, New Jersey Institute of Technology
Newark, NJ 07102
Big Bear Solar Observatory, New Jersey Institute of Technology
40386 North Shore Lane, Big Bear City, CA 92314
\textsuperscript{3} School of Mathematics and Statistics,
University of Sydney, NSW 2006 Australia

Abstract. In October 2003 we began a four year project whose goal is to build a real-time space weather monitoring and forecasting system. A major component of the project is the use of image processing and pattern recognition techniques to detect and characterize three important solar activities in real-time: filament eruptions, flares, and emerging flux regions. In this paper we describe a method for automatic tracking of the apparent separation motion of two-ribbon flares using a support vector machine for flare detection and image segmentation for flare characterization.

1 Introduction

Solar activity is closely related to the near earth environment – summarized descriptively as space weather. Changes in space weather have adverse effects on many aspects of life and systems on earth and in space, such as safety of spacecrafts and astronauts, interruption of communication and navigation systems, damage to power grids and trans-continental pipelines due to grounding electric currents caused by geomagnetic storms, and hazardous conditions to polar routed flights. Real-time, high-quality data and data processing would be a key element to forecast space weather promptly and accurately.

In October 2003 we obtained funding from the US National Science Foundation for a project entitled “Innovative Information Technology for Space Weather Research”. The goal of this four year project is to build a real-time space weather monitoring and forecasting system (see Figure 1 for an overview). We plan:

1. To use image processing and pattern recognition techniques to detect and characterize three important solar activities in real-time: filament eruptions, flares, and emerging flux regions.
2. To use state-of-art parallel computing and phase diversity speckle imaging techniques to yield near real-time diffraction limited images with a cadence of approximately 10 sec.
3. To develop Web based software tools to post our processed data, events and forecasting in real-time, and integration of these with current solar activity and space weather prediction Web pages at Big Bear Solar Observatory (BBSO).

The focus of the present paper is on the first aspect of the project, image processing and pattern recognition. In particular we describe a method for automatically tracking the ribbon separation motion of two-ribbon flares. The apparent motion of flare ribbons reflects the progress of magnetic reconnection. The measured properties of flare evolution can be compared with associated events such as coronal mass ejections (CMEs) (Yurchyshyn et al [1]). Statistical properties of flares can only be derived by the study of a large number of events. Therefore, we build a computer system to achieve such an objective.

The rest of the paper is organized as follows. Section 2 describes the pattern recognition used for automatic flare detection. Section 3 outlines the image processing used for automatic flare characterization culminating in automatic motion tracking of a two ribbon flare. Experimental results are given in Section 4, and the conclusion and future directions for research in Section 5.

2 Automatic Flare Detection

Qu et al [2] compared the multi-layer perceptron, radial basis function, and support vector machine (SVM) trained with nine global features of Hα images from BBSO on the task of flare detection. The SVM performs best, achieving a classification success rate of more than 95%. If we focus on large flares, the classification rate can be further improved. Using our classification program, we can detect the beginning and ending of a flare. After a solar flare is detected by SVM, we obtain the position of a flare using the key pixel, i.e. the one with the maximum gray level difference between current and previous images. An image may have several key pixels if the image contains several flares.
3 Automatic Flare Characterization

For automatic flare characterization the sequence of image processing steps consists of three major phases: preprocessing; segmentation; and motion tracking (for a detailed account see Qu et al [3]).

In the preprocessing phase we use image enhancement and filtering techniques to obtain high quality images. A median filter is used to remove additive noises, and a recursive soft morphological filter is applied to be less sensitive to additive noises and small variations (Shih and Puttagunta [4]). Then we detect the center of the solar disk in each image using an IDL (Interactive Data Language by Research Systems, Inc.) program, called fit limb, to align images using their centers. A solar flare is only a small feature on the full-disk solar image which at BBSO is 2048×2048 pixels. Based on our understanding of typical flare sizes, we pick a 400×400 window centered at the key pixel of the flare and use an empirical formula to normalize the mean brightness in this window. In the event that there are two flares at the same time on the solar image, we choose two 400×400 windows, one for each flare.

In the segmentation phase we combine both region based and adaptive boundary based methods (Castleman [5]), in addition to morphological and hole filling techniques, to obtain the accurate boundary and the structure of flares. This involves

- Region growing using the maximum graylevel in the window as the seed (see Figure 2b)
- Edge detection using Sobel and Laplacian operators. The flare boundary is difficult to find using a single global threshold. An adaptive approach for handling such a situation is to divide the original image into subimages and then utilize a different threshold for each subimage. Non-flare features that remain are removed by an improved adaptive thresholding method (see Figures 2c, d and e).
- Region growing using pixels on the bright side to expand the bright side area (see Figure 2f). We add the result of region growing to the result of the improved adaptive boundary method to have the best possible result (see Figure 2g).
- Morphological closing on Figure 2g is used to erase gaps and to smooth contour (see Figure 2h)
- To remove a small part, we count each individual component in the binary image. If the number of pixels in a component is less than a threshold, we remove it. Small holes inside the flare are filled by an empirically determined thresholding method (see Figure 2i).

In the motion tracking phase there are two major techniques to estimate the motion field: differential techniques and matching techniques (Trucco and Verri, [6]). We label the components of the flare, and build a flare model for each image and calculate differences between consecutive images using the previous segmented image results. After having the current model, the previous
Fig. 2. Automatic procedure to compute flare region and flare motion Image segmentation results for a flare peaked at 21:40:21 on 21 May 2002. a. Original image, b. result of region growing, c. result of global threshold boundary method, d. result of adaptive threshold boundary method, e. result of improved adaptive threshold method, f. result of boundary growing method, g. result of region growing plus result of boundary growing method, h. result of morphological closing, i. result of small part removing and hole filling.

model, and differences between the current and previous models, we calculate the expanding orientation and distance for flare using the following steps:

– Labeling the flare components according to the size and distances between every component. We also check each component with the previous model, and make sure they are overlapped (see Figure 3a).

Fig. 3. a. Result of component labeling, b. result of the final model, c. result of differences between the current and previous images
Building the first model using the main body of the flare, and reconstruct the new model referred by the previous model. For the first flare model, morphological closing with a very large structure element is used to erase big gaps, then recover the outer edge and keep the connections among the objects. For the following flare models, we adjust the old model according to the outer edge of the new flare components (see Figure 3.b).

- Obtaining differences on Figure 3c between consecutive images by subtracting the current model with the previous model.
- Based on the differences image, splitting the pixels into two parts: pixels on the current model and pixels on the previous model. Then we correspond the pixels on the current model to the pixels on the previous model using the shortest distance matching. We obtain the motion orientation and separation distance for all pairs of the pixels. Finally, we calculate the flare motion in each direction, and use orientation which we can obtain the maximum pixel pairs.

4 Experimental Results

We developed the programs in IDL. The programs run on a DELL Dimension L733r with CPU time 733 Mhz and memory of 256 Mbytes under Windows 2000. The process of image segmentation takes less than 20 seconds for each image, and the process of motion tracking less than 10 seconds. This computation time is acceptable for near real-time detection and characterization. To test the system we selected H\(\alpha\) two-ribbon flare images observed on May 21, 2002. Data were obtained at BBSO around 21:40:21UT.

In Figure 4a we compare the two-ribbon flare’s expanding distance obtained by hand using Qiu et al’s method and our method. The results are quite similar, however ours are obtained by an automatic procedure in near real-time.

We measure the magnetic reconnection rate in term of electric field \(E\) in reconnection current sheet. Methods and uncertainties of these measurements are extensively discussed by Qiu et al. Using our methods, we can measure the expansion velocity \(V_r\) of flare ribbons, then align the ribbon with the magnetic fields \(B_n\) they sweep through. Because our method can obtain all the moving pixels, we can obtain accurate \(B_n\) by using the average \(B\) of moving pixels. Then we can obtain \(E\) by

\[ E = V_r \times B_n \]  

The comparison of \(E\) field between Qiu et al’s method and our method is shown in Figure 4b. It shows our method can obtain a better result since the peak of \(E\) field from our result is closer to the peak of light curve of the flare observed in radio spectrum.

5 Conclusion and Future Directions

We have presented a method of automatic solar flare detection and characterization. The experimental results show that we can obtain accurate results, and the
Fig. 4. a. Comparison of the two-ribbon separation distance as a function of time for the flare happened on May 21, 2002. Red curve denotes the result of previous method, green curve denotes the result of our method. b. Comparison of derived electric field of the above two-ribbon flare. The light curve of 8.2 GHz radio emission is overplotted to indicate onset of the flare.

Processes of image segmentation and motion tracking take less than 30 seconds for each image. Our automatic process is valuable for forecasting and studies of solar flares since this process dramatically improves efficiency and accuracy via the automatic procedure. The method will allow us to study evolution properties of a large number of solar flares efficiently, which will help space weather forecasting.

Currently we are exploring two new directions for this research. Firstly we note that over the last decade there has been considerable progress in image processing based on nonlinear partial differential equations (PDEs) (see reviews by Pollak [8] and Tschumperle and Deriche [9]). The work of Weeratunga and Kamath [10] indicates that PDE-based processing could be a powerful tool for solar feature segmentation. Moreover PDEs have recently been used for object tracking in video, and it would be interesting to apply PDE tracking to a flaring region. Secondly we note that in this paper forecasting is related to what happens
after an event on the solar surface has occurred. It would be useful to be able to predict the onset of such events. One promising approach is to study the spatial complexity in the image which is clearly linked to the complexity of the magnetic structure in an active region. The temporal evolution of images texture measures, such as fractal dimension, could be a useful proxy for solar flare prediction (Gallagher et al [11]).

Acknowledgements

This work is supported by the National Science Foundation (NSF) under grants IIS-0324816, ATM 0233931 and ATM 0313591.

References

1. Yurchyshyn, V. B., Wang, H., Qiu, J., Goode, P. R. and Abramenko, V. I.: Magnetic topology in 1998 November 5 two-ribbon flare as inferred from ground–based observations and linear force–free field modeling. ApJ 540 (2000) 1143–1149.
2. Qu, M., Shih, F.Y., Jing, J., and Wang, H.: Automatic solar flare detection using MLP, RBF and SVM. Solar Phys. 217 (2003) 157–172.
3. Qu, M., Shih, F.Y., Jing, J., and Wang, H.: Automatic solar flare tracking using image processing techniques. Solar Phys. submitted (2004).
4. Shih, F.Y. and Puttagunta, P.: Recursive soft morphological filters. IEEE Trans. Image Processing 4(7) (1995) 1027-1032.
5. Castleman, K. R.: Digital Image Processing, Prentice Hall (1996).
6. Trucco, E. and Verri, A.: Introductory Techniques for 3-D Computer Vision, Prentice Hall (1989).
7. Qiu, J., Wang, H., Cheng, C.Z. and Gary, D.E.: Magnetic reconnection and mass acceleration in flare - coronal mass ejection events. ApJ (2003) in press.
8. Pollak, I.: Segmentation and restoration via nonlinear multiscale filtering. IEEE Signal Processing Magazine, Sept (2002), 26-36.
9. Tschumperle, D., and Deriche, R.: Diffusion PDEs on vector valued images. IEEE Signal Processing Magazine, Sept (2002), 16-22.
10. Weeratunga, S.K. and Kamath, C.: An investigation of implicit active contours for scientific image segmentation. Proc. Visual Communications and Image Processing Conference, IS&T SPIE Symposium on Electronic Imaging, San Jose, CA, (2004).
11. Gallagher, P. T., McAteer, R.T.J. and Ireland, J.: The fractal dimension of solar active regions. ApJ (2004) in press.
Source Separation Techniques Applied to Astrophysical Maps

E. Salerno\textsuperscript{1}, A. Tonazzini\textsuperscript{1}, E. E. Kuruoğlu\textsuperscript{1}, L. Bedini\textsuperscript{1}, D. Herranz\textsuperscript{1}, and C. Baccigalupi \textsuperscript{2}

\textsuperscript{1}ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
\textsuperscript{2}SISSA-ISAS, Via Beirut 4, 34014 Trieste, Italy

Abstract. This paper summarises our research on the separation of astrophysical source maps from multichannel observations, utilising techniques ranging from fully blind source separation to Bayesian estimation. Each observed map is a mix of various source processes. Separating the individual sources from a set of observed maps is of great importance to astrophysicists. We first tested classical fully blind methods and then developed our approach by adopting generic source models and prior information about the mixing operator. We also exploited a Bayesian formulation to incorporate further prior information into the problem. Our test data sets simulate the ones expected by the forthcoming ESA’s mission \textit{Planck Surveyor Satellite} mission.

1 Introduction

To recover the different components from sets of observed maps is an important problem in astrophysics. A radiometric map observed at any frequency band is a combination of emissions received from different sources, whose radiative properties, which affect the coefficients of the combination, are only partially known. Among the various components, the cosmic microwave background (CMB) is of utmost importance since its anisotropies give information about cosmological parameters which would allow a choice among competing cosmological theories to be made. The other components, or foregrounds, can be of galactic or extragalactic origin, and each of them has its own interest. Thus, rather than just filtering out the foregrounds, our aim is to extract each individual source. Finding efficient separation methods is an important issue, since an increasingly vast amount of radiometric observations is being made available by current or planned observational missions. Any totally blind source separation (BSS) problem cannot have a unique solution from the observed data alone, since both the coefficients and the sources are to be estimated, and this leads to an unsolvable ambiguity. To eliminate it, one should rely on known source properties. One such approach is independent component analysis (ICA) \cite{6}, which assumes mutual independence between the source signals. Even though in principle the astrophysical sources are not mutually independent, we attempted ICA techniques to assess their tolerance to an imperfect data model. We also tested an independent factor analysis method (IFA) \cite{1}, which allowed us to treat space-varying noise.
In this paper, we summarize and comment our experiments in astrophysical map separation, with particular reference to the data that will be made available by the ESA’s Planck Surveyor Satellite [11].

2 Data and Source Models

A common assumption in astrophysical source separation [2] is that each radiation process has a spatial pattern \( s_j(\xi, \eta) \) that is independent of its frequency spectrum \( F_j(\nu) \), where \( \xi \) and \( \eta \) are angular coordinates on the celestial sphere, and \( \nu \) is frequency. The total radiation observed in a certain direction at a certain frequency is given by the sum of a number \( N \) of signals of the type above. Ignoring the effects of the different telescope beams at different frequencies, the observed signal at \( M \) distinct channels can be modelled as

\[
x(\xi, \eta) = A s(\xi, \eta) + n(\xi, \eta)
\]  

where \( x=\{x_i, \ i=1, \ldots, M\} \) is the \( M \)-vector of the observations, \( i \) being a channel index, \( A \) is an \( M \times N \) matrix whose entries, \( A_{ij} \), are related to the spectra \( F_j(\nu) \) of the radiation sources and the frequency responses of the measuring instruments on the different channels, \( s=\{s_j, \ j=1, \ldots, N\} \) is the \( N \)-vector of the individual source processes and \( n=\{n_i, \ i=1, \ldots, M\} \) is the \( M \)-vector of instrumental noise. This noise is normally Gaussian and space-varying. A strictly blind source separation approach assumes \( A \) totally unknown and aims at obtaining \( A_{ij} \) and \( s_j \) from the observations \( x_i \) alone. In our application, however, each column of \( A \) has a known relationship to one of the source spectra \( F_j(\nu) \), which either depends on a single unknown spectral index or is completely known (this is the case for CMB and the Sunyaev-Zeldovich radiation from clusters of galaxies). Each element of \( A \) has thus the form \( A_{ij} = c_j g_j(\nu_i; \beta_j) \), where \( c_j \) is independent of frequency, \( g_j \) is a known function of frequency and of a possibly unknown spectral index \( \beta_j \), \( \nu_i \) is the center frequency of the \( i \)-th channel. Since our problem can be solved up to a scaling ambiguity [6], we can assume a modified matrix whose generic element is

\[
\tilde{A}_{ij} = \frac{A_{ij}}{A_{1j}} = \frac{g_j(\nu_i; \beta_j)}{g_j(\nu_1; \beta_j)}
\]

The data we used to test all the methods revised below have been either simulated or obtained by extrapolating existing data sets to the frequency range and angular resolution expected for the forthcoming Planck data.

3 Fully Blind Approaches and Independent Component Analysis

As the first step of our exploration, we assumed a noiseless model and adopted the fully blind ICA approach, proposing the first BSS technique to solve the
The separation problem in astrophysics [2], although in an highly idealised setting. The separation device was a feed-forward neural network that updates the entries of an $N \times M$ separation matrix $W$ at each received sample of the observed signal. The learning algorithm was a uniform gradient search aimed to minimize the Kullback-Leibler divergence between the probability density $p_u(u)$ of the output vector $u(\xi, \eta) = Wx(\xi, \eta)$ and a function $q(u)$, that should represent the factorized joint probability density of the true sources. Since the true source densities are usually unknown, function $q$ is simply chosen according to the supposed sub-Gaussianity (i.e., negative kurtosis) or super-Gaussianity (positive kurtosis) of the sources. The results obtained by this method were quite promising in terms of source estimation accuracy. The accuracies in the estimation of the columns of $A$ were comparable to the ones obtained for the related sources. The robustness against noise of this technique is not high.

In a successive work [10], we included noise in our model and investigated the performance of the noisy FastICA algorithm [5]. The telescope beam was still assumed frequency independent, and the Gaussian instrumental noise was assumed space-invariant, but at the mean nominal levels for Planck. In terms of accuracy and computational speed, this approach was a considerable improvement over the neural algorithm described above.

An alternative method to deal with noisy mixtures is independent factor analysis (IFA) [1]. IFA employs an analytic source model where the source distributions are mixtures of Gaussians, whose parameters are to be estimated jointly with the mixing matrix. The mixing model also contains the noise covariance matrix, which can be estimated as well. IFA is performed in two steps: in the first one (learning), the mixing matrix, the noise covariance and the source density parameters are estimated via an EM algorithm. In the second step (separation), the sources are estimated by using the densities obtained in the first step. In [7], we developed an extension of the original algorithm that assumes a known and space-dependent noise covariance matrix, and updates the source parameters pixel by pixel to consider the different noise variances. We performed the learning step by simulated annealing instead of expectation-maximization (as was done in [1]). This also made the algorithm flexible enough to introduce prior knowledge about the matrix. Experiments with fixed model parameters gave better results than expectation-maximization, even with low SNRs (e.g. 14 dB), yielding a good convergence to the correct mixing matrix. For low SNRs, some of the mixture-of-Gaussian model parameters were estimated rather poorly. Nevertheless, the maximum-likelihood source estimates were better than the ones obtained by FastICA.

Another approach based on generic source models exploits Markov Random Fields models for describing the local autocorrelation of the individual sources, and relies on Bayesian estimation. Though implemented with no assumption on the mixing matrix, and with fixed hyperparameters for the Markov distribution, this method has already given better results than FastICA, in terms of robustness against, possibly nonstationary, noise [12][9].
To profit from the whole richness of prior information we have available, we have extended our formulation to a full Bayesian one that enabled us to assign priors to the source model parameters, which were instead fixed in IFA. This richness in formulation comes with the price of analytical intractability, which we tackled with numerical techniques, namely, Markov Chain Monte Carlo. This formulation enabled us to obtain posterior densities for the variables of the mixing system and to make inferences about all of the source statistics. In comparison with FastICA, we obtained significantly better results [8], at the price of a much higher computational cost.

We also attempted a Bayesian formulation that extends Kalman filtering to non-Gaussian time series. This approach is called particle filtering, and is the first to explicitly address the non-stationarity of the data in a problem of astrophysical map separation. This approach potentially provides the most elaborate formulation to the problem, and our initial experiments have already given very promising results [4].

4 Semi-blind Approaches and Dependent Component Analysis

All the blind source separation approaches proposed in the literature assume mutual independence between sources. This assumption is often unphysical. In the case we are examining here, for example, significant cross-correlations between the galactic foregrounds are expected. On the other hand, if we exploit the parametrisation of the mixing matrix described in Section 2, the independence assumption may become unnecessary. Parametrising the mixing matrix also allows us to just use second-order statistics to find a unique solution to both the learning and the separation problems. This is a novelty with respect to the partially or totally blind separation techniques proposed so far in astrophysical data analysis, since parametrisation reduces the number of unknowns and allows some of the correlation coefficients to be estimated as well.

Assuming to know which pairs of sources are correlated to each other, we derived a very fast model learning algorithm, based on matching the theoretical zero-shift data covariance matrix to the corresponding empirical matrix [3]. The unknowns of our problem are the set of parameters specifying matrix $A$ (see Section 2), plus all the nonzero elements of $C_s$. From eq. (1), the covariance matrix of the observed data at zero shift is:

$$C_x(0,0) = \langle [x(\xi,\eta) - \mu_x] [x(\xi,\eta) - \mu_x]^T \rangle = AC_s(0,0)A^T + C_n. \quad (3)$$

where the angle brackets mean expectation, and $C_n$ is the noise covariance matrix, which is known and diagonal.

Let us now define the matrix

$$H = C_x(0,0) - C_n = AC_s(0,0)A^T. \quad (4)$$
An estimate of $H$ can be derived from matrix $C_n$ and the sample average of matrices $[x - \mu_x][x - \mu_x]^T$ (see eq. 3). Matrices $A$ and $C_s$ can then be estimated by minimizing the following form over all the unknowns

$$||H - AC_s(0,0)A^T||_F$$

where subscript $F$ denotes the Frobenius norm. Of course, the existence of a convenient minimum to (5) is determined by the content of information of the available data set and by the total number of unknowns. In our experiments (see Figures 1 and 2), we learned the model efficiently with four observed maps on a subset of the Planck channels, and a $4 \times 3$ mixing matrix. Our unknowns were two spectral indices specifying $A$ and four nonzero elements of the source covariance matrix.

The individual source maps can then be roughly recovered by multiplying the data vectors by the generalised inverse of the mixing matrix estimated. The output maps are corrupted by amplified Gaussian noise with known statistics from which, by deconvolution, we can also recover the source densities. The output noise statistics can also be exploited to improve the accuracy in the estimated angular power spectra of the individual sources, which are of great interest to astrophysicists.

This strategy resulted much more immune from erratic data than the ones based on higher-order statistics. By also taking into account nonzero-shift covariance matrices, it will be possible to estimate the source covariance matrices as functions of the shift. This extension of the method is currently being developed.

In Figure 1, we show how knowledge of the output noise statistics can help in obtaining an accurate power spectrum estimation. In this case we have the spherical harmonic CMB power spectrum for multipoles $l$ from 1 to 2000 (the significant range for the Planck observations), obtained by the second-order-statistics method described above. The original data were very noisy (the CMB-to-noise ratio was 0 dB at 100 GHz). It can be seen that by simply subtracting the theoretical noise spectrum from the estimated plot, we can obtain a corrected version of the spectrum. In Figure 2 we see the original and reconstructed CMB
and galactic synchrotron maps on a sky patch centered on the galactic plane, where the separation of CMB is normally very difficult. The CMB-to-noise ratio was still 0 dB at 100 GHz. As can be seen, the reconstructed maps are almost indistinguishable from the originals. The results we obtained by this algorithm were the best in terms of robustness against noise, and we assessed them by extensive Monte-Carlo trials [3]. It is also to note that some of the foreground maps in many cases were strongly correlated to each other. This did not impair the learning procedure, from which it was also possible to accurately estimate the cross-correlation coefficients at zero shift.

5 Conclusions

In this paper, we have presented and discussed several blind and semi-blind methods for the separation of components in astrophysical images. Some of them are novel approaches to BSS in imaging, or original extensions of previous methods. In particular, we developed strategies for handling non-stationary noise and for introducing available a priori information into the problem, related to the autocorrelation properties of the individual sources and to the relationships among the mixing coefficients. This brief presentation reflects our path through this challenging problem: starting from the methods based on the pure ICA paradigm, i.e., mutually independent sources and fully blind estimation, we have now achieved significant results in dealing with auto and cross-correlated sources, stationary and non-stationary noise, and in exploiting efficiently the a priori knowledge coming from the physics of the problem.
Acknowledgements

D. Herranz is supported by the EC Human Potential Programme under contract HPRN-CT-2000-00124 CMBNET. We are indebted to our colleagues from the Planck technical working group 2.1 (diffuse components separation) for fruitful discussions and exchange of results and data. Some of the data processing needed for this work has exploited the HEALPix spherical pixelization scheme (http://www.eso.org/science/healpix), by K. M. Górski et al.

References

1. Attias, H., 1999, Independent factor analysis, Neural Computation, 11, 803.
2. Baccigalupi, C. et al., 2000, Neural networks and the separation of cosmic microwave background and astrophysical signals in sky maps, M.N. Royal Astronomical Society, 318, 769.
3. Bedini, L. et al., 2003, A semi-blind second-order approach for statistical source separation in astrophysical maps, ISTI-CNR, Pisa, Technical Report ISTI-2003-TR-35.
4. Costagli, M., Kuruoğlu, E.E., Ahmed, A., 2003, Bayesian separation of independent components in astrophysical images using particle filters, ISTI-CNR, Pisa, Technical Report, ISTI-2003-TR-54.
5. Hyvärinen, A., Oja, E., 1997, A fast fixed-point algorithm for independent component analysis, Neural Computation, 9, 1483.
6. Hyvärinen, A., Karhunen, J., Oja, E., 2001, Independent Component Analysis, Wiley, New York.
7. Kuruoğlu, E.E. et al., 2003, Source separation in astrophysical maps using independent factor analysis, Neural Networks, 16, 479.
8. Kuruoğlu, E.E., Milani Comparetti, P., 2003, Bayesian source separation of astrophysical images using Markov Chain Monte Carlo, Proc. PHYSTAT, Stanford, 8-11 September 2003.
9. Kuruoğlu, E.E., Tonazzini, A., Bianchi, L., 2004, Source separation in astrophysical images modelled by Markov random fields, submitted to ICIP’04.
10. Maino, D. et al., 2002, All-sky astrophysical component separation with fast independent component analysis (FastICA), M.N. Royal Astronomical Society, 334, 53.
11. http://astro.estec.esa.nl/planck/
12. Tonazzini, A. et al., 2003, Blind separation of auto-correlated images from noisy images using MRF models, Fourth International Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, 675.
Counting Magnetic Bipoles on the Sun by Polarity Inversion

Harrison P. Jones

NASA’s Goddard Space Flight Center, Laboratory for Astronomy and Solar Physics, Southwest Solar Station, c/o National Solar Observatory*, PO Box 26732, Tucson, AZ 85726 hjones@nso.edu

Abstract. This paper presents a simple and efficient algorithm for deriving images of polarity inversion from NSO/Kitt Peak magnetograms without use of contouring routines and shows by example how these maps depend upon the spatial scale for filtering the raw data. Smaller filtering scales produce many localized closed contours in mixed polarity regions while supergranular and larger filtering scales produce more global patterns. The apparent continuity of an inversion line depends on how the spatial filtering is accomplished, but its shape depends only on scale. The total length of the magnetic polarity inversion contours varies as a power law of the filter scale with fractal dimension of order 1.9. The amplitude but not the exponent of this power-law relation varies with solar activity. The results are compared to similar analyses of areal distributions of bipolar magnetic regions.

1 Introduction

Loci of polarity inversion for the Sun’s line-of-sight magnetic field (commonly but misleadingly termed “neutral lines”) have long been a central feature of both observational and theoretical studies of solar activity. Original interest was based on the idea that magnetic reconnection and rapid energy release necessary for impulsive phenomena such as solar flares could only occur in certain polarity configurations near a true null in magnetic field (see, for example, Sturrock, 1968). More recently such regions have been termed separators (Pontin, Priest, and Longcope, 2003). They mark three-dimensional discontinuities in the connectivity of magnetic field lines and only accidentally coincide with polarity inversion. Nonetheless, polarity inversion is still a useful tool for classifying the topography of the Sun’s magnetic field and is of considerable interest in, for example, demarcation of coronal structures (McIntosh, 1994), filaments and filament channels (Martin, Bilimoria, and Tracadas, 1993; Chae et al., 2001), and empirical prediction of coronal mass ejections (Falconer, Moore, and Gary, 2002).

* The National Solar Observatory is operated by the Association of Universities for Research in Astronomy under contract with the National Science Foundation.
Inversion lines are often drawn by hand, left to the eye of the viewer of a designated magnetogram, or are determined by contouring algorithms with implicitly specified resolution. A more objective, cartographic perspective is taken for this paper; i.e., focus is on algorithms for determining polarity inversion from magnetograms with pre-specified, explicit resolution scales and on how the character of polarity inversion depends upon this scale.

2 Method

The observational data for this paper are magnetograms from the National Solar Observatory/Kitt Peak Vacuum Telescope (NSO/KPVT) (Livingston et al., 1976) which made daily full-disk images of the photospheric line-of-sight field from 1973 to 2003. Incident solar light was passed through a circular polarization modulator and imaged onto a spectrograph entrance slit which subtended a 512 arc-second segment of the solar image. The magnetograph was placed at the exit plane of the astigmatic spectrograph and used the Zeeman wavelength modulation of a spectrum line in circularly polarized light to infer the line-of-sight field at every position along the entrance slit. The solar image was scanned across the entrance slit in a four-swath pattern to form a two-dimensional, full-disk magnetogram.

A recipe for extracting polarity inversion lines from magnetograms is not difficult to devise. The following procedure was especially tuned for application to NSO/KPVT magnetograms.

- **Zero-Point Correction:** To correct the poorly determined zero point for magnetograms obtained prior to 1992 with the original 512-channel Diode Array Magnetograph, a linear offset was applied for each position of the entrance slit. To compensate for temporal variations of instrument response during the spatial scanning process, the offset amount was computed as the median of all fields at each spatial pixel along the slit within ±10 Gauss. (Pixels with these field strengths are representative of the noise distribution of the instrument and should ideally be centered around zero.) Failure to perform this step resulted in polarity inversion contours which close around regions of one polarity.

- **Block Averaging:** $L \times L$-pixel block averages of the original magnetograms were formed for better continuity of the polarity inversion lines and for enhanced efficiency of spatial filtering. The details of the block averaging enter into the determination of the resolution scale and are discussed in more detail below.

- **Smoothing:** The block averaged images were further smoothed with a Gaussian filter with standard deviation of $\sigma$ pixels. Again, the details affect the resolution scale. For large images, performing the block average before the Gaussian filter greatly enhances efficiency.

- **Algebraic Sign:** Images were formed of the algebraic sign of the magnetograms after the above two-stage spatial filtering.
- **Edge Detection:** A change of algebraic sign was detected by comparing the summed absolute values of the four differences between nearest and diagonally nearest pixels to a threshold.

Many reasonable variants of this recipe can be constructed which may differ cosmetically, as discussed below, but which will produce the same polarity inversion contours for the same resolution scale. The two-stage smoothing routine used here was developed by trial and error and has two advantages. First, it is efficient since it avoids full Gaussian convolution with large-scale kernels. Second, as discussed below, it enables a simple pixel-counting procedure for measuring the length of the polarity inversion loci without reference to contouring routines.

One can obtain a measure of the scale, $s(L, \sigma)$, by requiring $s$ to be inversely proportional to the the half-width at half-maximum of the Fourier transform of the two-stage block-average/Gaussian smoother (the product of a Gaussian and a sinc function). The constant of proportionality is set by demanding that $s(1,0) \equiv 1$ pixel. Numerical evaluation of this function shows that, within a couple of percent,

$$s(L, \sigma) \approx L(1 + 10.35\sigma^2)^{1/2}. \quad (1)$$

Figure 1 shows inversion line images superposed on original magnetograms for $s = 16$ and $s = 64$ arc-seconds for two days, one near solar minimum (1984 Sep 15), the other near solar maximum (1992 Dec 15). As the scale decreases, the maps become more complex and outline smaller magnetic features. The maps at a scale of 64 arc-seconds probably best match inversion lines which most researchers would draw.

A measure of this complexity is the total length of the inversion loci. To facilitate measurement of inversion line length, define $n(s,L)$ as the number of inversion-line pixels in an $L \times L$ block-averaged image. If the “width” of the inversion loci were unity, $nL$ would measure the length in pixels of the original image. Empirically,

$$Ln(s,L) \approx constant \equiv n(s,1) \quad (2)$$

for a subrange of $L$ at a given $s$ but falls below this relation as $L$ approaches unity where the inversion lines lose continuity. Thus one may determine the “length” of the inversion loci in pixel units of the original image by extrapolation to unit pixel scale and averaging over the subrange:

$$\ell(s) \equiv n(s,1) \approx \langle Ln(s,L) \rangle. \quad (3)$$

### 3 Analysis

Inspection of logarithmic plots of $\ell(s,L)$ shows a power-law (fractal) dependence with respect to scale over nearly two orders of magnitude of scale. Since this behavior is reminiscent of the areal distributions reported for bipolar magnetic
regions by Harvey and Zwaan (1993), $\ell(s, L)$ has been computed for their sample of 233 active-Sun and 259 quiet-Sun magnetograms. The results are shown in Figure 2. Solid lines connect the average data, dashed lines are the linear least-squares fits to the log-log plots, and “error” bars show $\pm$ one standard deviation (not standard error of the mean) over the respective samples. The fractal dimensions of the linear fits are $-1.873 \pm 0.005$ and $-1.891 \pm 0.008$ for quiet- and active-sun magnetograms, respectively. Uncertainties here are $\pm$ 3 standard errors of the mean of individual fits over the entire sample. The log amplitudes at 0.033 solar radii are $2.220 \pm 0.003$ and $2.129 \pm 0.005$ respectively.

Note from Fig. 2 that the amplitude of the variation varies with solar cycle, but the exponent of the power law does not, a behavior similar to that found
Fig. 2. Inversion line length vs resolution scale for the active- and quiet- magnetogram samples of Harvey and Zwann (1993). Solid lines connect average data points with sample standard deviations indicated by error bars. Dashed lines show linear fits to the averages.

by Harvey and Zwaan (1993). To compare with their results in more detail, one may characterize their results as

\[ n(t, A) \sim f(t)g(A). \]

One finds from their Figure 4 that \( f(\text{active})/f(\text{quiet}) \approx 5 \) and \( g(A) \propto A^{-1.8} \). From the inversion line results, \( \ell_{\text{quiet}}(s)/\ell_{\text{active}}(s) \approx 1.23 \). By expressing inversion line length as a function of resolution area rather than scale, one can combine these results to infer an empirical relation:

\[ \ell(s) \sim \text{const} \ast [g(s^2)/\sqrt{f(t)}]^{0.25}. \]

Thus, at least empirically, length of the polarity inversion loci may be used to “count” bipolar magnetic regions.
4 Discussion

This paper outlines simple procedures for efficiently smoothing magnetograms with a well-defined and pre-specified resolution scale and deriving loci of polarity inversion from the smoothed magnetograms. Unlike subjective methods or packaged contouring routines, the resolution scale is both objective and fixed. A reasonable measure of inversion-line length varies as a power-law of resolution scale over about two orders of magnitude with a fractal dimension of about 1.9, irrespective of solar activity; that is, inversion lines are nearly space-filling. Finally, by comparing the results to the statistics for bipolar regions derived by Harvey and Zwaan (1993) an approximate relation is inferred which suggests that polarity inversion is an economical way to measure the areal distribution of bipolar magnetic regions.

Future work includes extending the analysis to all available magnetograms from the NSO/KPVT and comparing the results with simulations and physical models of solar magnetic fields. Is this result a constraining feature of such models or does it inevitably apply to a wide class of models?

Acknowledgements

NSO/KPVT data used in this paper were produced cooperatively by NSF/AURA, NASA’s GSFC, and NOAA/SEC. This work was supported by NASA Office of Space Sciences grants 344-12-51-14 and 344-12-51-19.

References

Chae, J., Martin, S. F., Yun, H. S., Kim, J., Lee, S., Goode, P. R., Spirock, T., and Wang, H.: Small Magnetic Bipoles Emerging in a Filament Channel. Astrophys. J. 548 (2001) 497-507.
Falconer, D. A., Moore, R. L., Gary, G. A.: Correlation of the Coronal Mass Ejection Productivity of Solar Active Regions with Measures of Their Global Nonpotentiality from Vector Magnetograms: Baseline Results. Astrophys. J. 569 (2002) 1016-1025.
Harvey, K. L. and Zwaan, C.: Properties and Emergence of Bipolar Active Regions. Solar Phys. 148 (1993) 85-118.
Livingston, W. C., Harvey, J., Pierce, A. K., Schrage, D., Gillespie, B., Simmons, J., Slaughter, C.: The Kitt Peak 60-cm Vacuum Telescope. Appl. Opt. 15 (1976) 33-39.
Martin, S. F., Bilimoria, R., and Tracadas, P. W.: Magnetic Field Configurations Basic to Filament Channels and Filaments. Bull. Amer. Astron. Soc. 25 (1993) 1217.
McIntosh, P. S.: YOHKOH X-Ray Image Interpretation with Overlays of Hα Neutral Lines. in X-ray Solar Physics from Yohkoh, Y. Uchida, T. Watanabe, K. Shibata, and H. S. Hudson (eds.) (1994) 271.
Pontin, D. I., Priest, E. R., and Longcope, D. W.: A Framework for Understanding the Topology of Complex Coronal Structures. Solar Phys. 212 (2003) 319-342.
Sturrock, P. A.: A Model of Solar Flares. in Structure and Development of Solar Active Regions, K. O. Kiepenhuer (ed.), IAU Symposium 35 (1968) 471-479.
Correlation of the He I 1083 nm Line Width and Intensity as a Coronal Hole Identifier

Olena Malanushenko\textsuperscript{1} and Harrison P. Jones\textsuperscript{2}

\textsuperscript{1} National Solar Observatory, Tucson, AZ 85719, U.S.A.
\textsuperscript{2} NASA’s Goddard Space Flight Center, Laboratory for Astronomy and Solar Physics, Southwest Solar Station, c/o National Solar Observatory, PO Box 26732, Tucson, AZ 85719
\{elena,hjones\}@noao.edu

Abstract. The locations of coronal holes are usually based on equivalent width images in the He I 1083 nm line. However, it is difficult to differentiate coronal holes from the centers of quiet chromospheric network without complementary data and the skill of an experienced observer. Analysis of imaging spectroscopy shows that line width and central intensity are oppositely correlated in coronal holes and quiet Sun. This fact can be used to form images of linear combinations of these quantities in which coronal holes are easily identified.

1 Introduction

Coronal Holes (CHs) are large-scale phenomena in the solar atmosphere ranging in height from the upper chromosphere to the corona. CHs overlie large-scale unipolar magnetic fields and are identified as sources of high speed solar wind, the driving mechanism of which is still under investigation.

On coronal X-ray and UV images CHs are seen as large dark areas outlining depressions of hot coronal radiation. Because of the He atomic structure, the intensity of the He I 1083 nm line strongly depends on photoionizing radiation from the upper transition region and corona (see (Andretta and Jones 1997); (Avrett, Fontenla and Loeser 1994)). The deficiency of coronal radiation in CHs thus induces a weakening of the He line which can be seen as bright areas on 1083 nm equivalent width images displayed with negative contrast (Harvey, J. et al. 1975). However, the centers of supergranular cells are also comparably bright and are easily confused with CHs.

Another distinctive feature of CHs in He line observations is a diffuse, low-contrast internal structure which visually resembles chromospheric network in quiet areas (Harvey, J. et al. 1975). This phenomenon has been named as a “depression of contrast of the chromospheric network” and is regularly used as part of the operational definition of a CH. Contrast of chromospheric network is a difficult parameter to compute since it should take into account not only variation of intensity but also the character of non-regular, small scale structures.
Sometimes the difference between CHs and quiet regions on He intensity or equivalent width images is clear enough to overplot CH manually. In other cases CHs cannot be recognized without complementary data and the skill of an experienced observer. Harvey, K. and Recely (2002) used composite criteria for CH recognition based on: He equivalent width images (value, contrast of network, size of area), Fe I 868.8 nm magnetograms (predominately polarity of small scale magnetic fields), and the depression of intensity on X-ray images. Almost all CH maps collected at the Digital Library of National Solar Observatory (NSO) were prepared manually using this method.

Even where a location of CH is clear, it is difficult to create a computer algorithm for automatic recognition. One automatic procedure for CH recognition was applied on recent Kitt Peak Vacuum Telescope He synoptic observations (J.Harvey, private communication) The method is based on analysis of intensity and local variability for two sequential days on He images and cannot be used for a real time CH recognition.

Recently, we found another distinctive feature of CHs in He line observations: the broadening of the He line in spectra of CHs ((Malanushenko and Jones 2002), 2004). Here, we explore whether the broadening of the He line in CHs can be a useful CH diagnostic, particularly for distinguishing CHs from centers of chromospheric network.

2 Observations

Imaging spectroscopy in the He line was obtained with the NSO Kitt Peak Vacuum telescope on 2000 April 17 using the NASA/NSO spectromagnetograph ((Jones et al. 1992)). A CCD camera recorded spectra along a spectrograph entrance slit and a scanning system stepped the solar image perpendicular to the slit to produce an image. The spectral dispersion of the observations was 0.083Å/pix, and the spectra covered a range of 1.7 nm, roughly centered on 1083 nm. Length of the slit was 480′′, and spatial resolution along entrance slit was 1.1′′, and the scanning step was 1.3″. To improve the signal-to-noise ratio, these spectra were averaged to a spatial scale of 2.6″. The observation subtended a region of about 470″ × 900″ on the solar disk, and about 3 × 10⁵ spectra were recorded for analysis.

An intensity image at the central wavelength position of He 1083.0 nm line is shown on Fig. 1a. One can see a large CH in the center of image as the area with low absorption and depressed internal structure contrast. Figs. 1.b,c,d show the same image at different contour levels which were unsuccessful attempts to outline the CH area. Fig.(b) shows a contour 2% above average intensity in the quiet Sun, and this marks too small a central area for the CH. Contours at the 1% level (c) correspond better to our visual concept of CH area on this image, but it delineates not only the CH but also small scale bright centers of chromospheric network, and the separation is inadequate. Contours in (d) show average levels for the quiet Sun but and show complicated structure of chromospheric network.
Correlation of the He I 1083 nm Line Width and Intensity

and a large area around the CH. The solid line on the Figure 1.a shows where spectra were selected for further discussion below.

Fig. 1. (a) An image at the center of He 1083.0 nm, observed with the KPVT/SPM on 17 April 2000; (b) - the same image with superposed central intensity contours 2% above average; (c) contours 1% above average; (d) contours of average central intensity

3 Data Reduction

Data reduction includes dark and flat-field correction of the CCD images, alignment of spectra to solar lines, normalization of the spectra to a spectral standard to compensate for non-linear continuum, and de-blending of nearby solar and telluric lines. The data reduction technique is described in detail by Jones (2003) and by Malanushenko and Jones (2004). A summary explanation follows.

For dark and flat corrections were applied using a special technique, developed for synoptic observations at the NSO/KPVT (Jones 2003). The telescope is moved limb-to-limb from south to north, parallel to the entrance slit, as signals are summed. Each pixel of the CCD camera accumulates the same solar illumination during this observation and least-squares fitting procedures are used to separate spectral lines from defects such as dust, non-uniform illumination, and pixel-by-pixel gain variations.

At every spatial position of the entrance slit, the spectra along the slit are shifted and stretched to force the positions of solar lines at Si I 1082.71 nm and Na I 1083.49 nm to correspond to their position in reference atlas. This
compensates for non-uniform spectrograph dispersion along the entrance slit and for local line-of-sight motions on the Sun. Gaussian profiles are fit to the cores of both solar lines in every spectrum and wavelength interpolation forces the match.

Correct determination of the continuum level in He I 1083 nm data is crucial for studying coronal holes since the line is weak and small errors in the continuum are amplified when compared to the central depth of the line profile. The continuum in this spectral range is difficult to determine since spectral blends mask the appearance of the continuum for several Å around line center, and instrumental spectral response, after flat-fielding is nonlinear. These effects are treated using our method based on comparison of the observations with a well calibrated spectral standard in a variation of a technique originally proposed by Malanushenko, V. et al. (1992).

To separate the He I 1083 nm lines from unwanted spectral blends, an average spectrum, with He removed, is subtracted from the normalize spectrum. The He component which is removed from the average spectrum is determined from a multi-profile least-squares fit. This procedure is reasonably accurate for our observation since only the He I lines show noticeable variation over the field of view.

4 Data Analysis

Gaussian profiles are fit to the central part of main spectral component of the He multiplet and central intensity (\(I\)), half width at half maximum (\(W\)) and line shift relative to the average spectrum (\(V\)) are computed from the parameters determined by the fit. Figure 2 presents plots of how the parameters depend on disk position for one row of data. Solid lines correspond to values of the average spectrum. The CH location is determined from central intensity as an area of more than 2-3 network cells where intensity is 1-2% larger than the average level and contrast of internal structure is depressed. Two vertical strips show one CH and one network cell of locations (Q).

As discussed earlier, it is difficult to separate the CH from the network cell center based on central intensity alone. However, inspection of Fig. 2 reveals that the line width in CH is higher than the average value and qualitatively correlates positively with spatial variation of intensity. The situation is different in the network cell, where line width is smaller and correlates negatively with intensity. This difference in the correlation of line width and intensity allows one to clearly distinguish CHs from network cells when both plots are shown. The different correlative properties between intensity and width in CH and in quiet Sun network cells suggests that cell-like structures in the two regions are physically different.

From plots of radial velocities one sees a negative (blue) shift of the He line in the CH and in the network cell center, corresponding to outward flows; there is also a red shift in the borders of the network cell. Line shifts correlate positively with intensity in both CH and cell network, and no principal differences between them are seen in on velocity data.
Fig. 2. Intensity ($I$), half width at half maximum ($W$) and relative line shift ($V$) of the He line on as a function of position along entrance slit for a single row of data. The horizontal solid lines correspond to values for the average spectrum. Shaded vertical regions show coronal-hole (CH) and one network-cell (Q) locations.

One can note some differences between the behavior of intensity on and line shift. On passing from the quiet Sun to inside the CH, the intensity increases to a plateau value very quickly, but the width and line shift approach their maximal values more gradually.

The different correlative properties allow one to form images of linear combinations of width and intensity in which the CH is easily distinguished from quiet-Sun network. To make the units and variability of the two quantities comparable, $I$ and $W$ are statistically standardized by subtracting their respective quiet-sun means and dividing the residuals by the corresponding standard deviations ($\sigma$). The resulting parameters have dimensionless $\sigma$ units (e.g., $i = (I - <I>)/\sigma_I$ and $w = (W - <W>)/\sigma_W$).

Figure 3 shows standardized images of $i$ (a), $w$ (b), with a contrast range of $\pm 5.0 \sigma$ and their sum $i + w$ (c) and difference $i - w$ (d) with a contrast range of $\pm 10.0 \sigma$. Both $i$ and $w$ show similar high values in CH and opposite contrast.
444 O. Malanushenko and H.P. Jones

Fig. 3. Normalized images (dimensionless units; see text). (a) intensity $i$; (b) half-width $w$; (c) $i-w$; (d) $i+w$. Grey scale is linear over (-5,5) for a,b plates and (-10,10) for c,d in chromospheric network. Neither images are sufficient to outline a CH without confusion with quiet-Sun network. However, the sum $i+w$(c) doubles the contrast in the CH and suppresses the contrast of the network so that contour at 2.0 $\sigma$ objectively outlines the CH area (see fig. 3.c). On the other hand, the difference of normalized images $i-w$(d) cancels contrast in CH and increases the contrast in quiet-Sun network. This result is consistent with the above discussion of one row of spectra.

5 Conclusions

We apply a new method for locating CHs in high resolution imaging spectroscopy using He I 1083 nm data obtained at the NSO/KPVT with the NASA/NSO spectromagnetograph and find the following results.

1. Intensity and line width data are both individually insufficient for CH recognition. One cannot distinguish the centers of network from a CH based on the intensity data, and one cannot separate a CH from the borders of network based on the line width alone.

2. Intensity and line width in CHs are spatially correlated but are negatively correlated in the quiet Sun. The sum of standardized intensity and line-width images shows increased contrast in a CH and depressed contrast in the network. This can be used as an objective coronal hole diagnostic.
Acknowledgments. This research was partially supported by NASA Supporting Research and Technology task 344-12-52-14 and 344-12-52-19. NSO/Kitt Peak data used here were produced cooperatively by AURA/NSO, NASA/GSFC, and NOAA/SEC.

References

Andretta, V., Jones, H.P.: On the Role of the Solar Corona and Transition Region in the Excitation of the Spectrum of Neutral Helium. Astrophys. J. 489 (1997) 375–394
Avrett, E.N., Fontenla, J.M., Loeser, R.: Formation of the solar 10830 A line. IAU Symp. 154, Infrared Solar Physics, D.M Rabin et al. ed., (1994) 35–47
Harvey, J. W., Krieger, A. S., Davis, J. M., Timothy, A. F., Vaiana, G. S.: Comparison of Skylab X-Ray and Ground-Based Helium Observations. Bull. Americ. Astronom. Soc 7 (1975) 358
Harvey, K.L., Recely, F.: Polar Coronal Holes During Cycles 22 and 23. Solar Phys. 211 (2002) 31–52
Jones, H. P.: Data Calibration and Analysis for He I 1083 nm Imaging Spectroscopy. Solar Phys. 218 (2003) 1–16
Jones, H. P., Duvall, T. L. Jr., Harvey, J. W., Mahaffey, C. T., Schwitters, J. D., Simmons, J. E.: The NASA/NSO spectromagnetograph. Solar Phys. 139 (1992) 211–232
Malanushenko, O.V., Jones, H.P: New Analysis of He I 1083 nm Imaging Spectroscopy. Bull. Americ. Astronom. Soc 34 (2002) 700
Malanushenko, O.V., Jones, H.P.: Analysis of He I 1083 nm Imaging Spectroscopy Using a Spectral Standard. Solar Phys. in press
Malanushenko, V.P., Polosukhina, N.S., Weiss, W.W.: Spectrum variations of HD 215441 (Babcock’s star) Astron. Astrophys. 259(1992) 567–573
Automated Recognition of Sunspots on the SOHO/MDI White Light Solar Images

S. Zharkov, V. Zharkova, S. Ipson, and A. Benkhalil
Department of Cybernetics, University of Bradford, BD7 1DP, UK
S.Zharkov@brad.ac.uk

Abstract. A new technique is presented for automatic identification of sunspots on the full disk solar images allowing robust detection of sunspots on images obtained from space and ground observations, which may be distorted by weather conditions and instrumental artefacts. The technique applies image cleaning procedures for elimination of limb darkening, intensity noise and non-circular image shape. Sobel edge-detection is applied to find sunspot candidates. Morphological operations are then used to filter out noise and define a local neighbourhood background via thresholding, with threshold levels defined as a function of the quiet sun intensity and local statistical properties. The technique was tested on one year (2002) of full disk SOHO/MDI white light (WL) images. The detection results are in very good agreement with the Meudon manual synoptic maps as well as with the Locarno Observatory Sunspot manual drawings. The detection results from WL observations are cross-referenced with the SOHO/MDI magnetogram data for verification purposes.

1 Introduction

With a substantial increase in the size of solar image data archives, the automated detection and verification of various features of interest is becoming increasingly important for, among other applications, data mining and the reliable forecast of the solar activity and space weather. However, this raises the accuracy and reliability required of the detection techniques used for automated recognition which have to be significantly improved in comparison with the existing manual ones in order to create a fully automated Solar Feature Catalogue. Manual sunspot catalogues in various formats are produced in various locations all over the world such as the Meudon Observatory (France), the Locarno Solar Observatory (Switzerland), the Mount Wilson Observatory (US) and many others. Sunspot studies play an essential part in the modelling of the total solar irradiance and in determining variations of sunspot properties with latitude and/or the solar cycle phase.

The compilation of the Zurich relative sunspot numbers, or since 1981 the Sunspot Index Data (SIDC), is one of the most commonly used measures of solar activity (Hoyt & Schatten[1] and Temmer, Veronig, and Hanslmeier[2]). As integral component of Solar Active Regions, sunspot behaviour is also used in the study of Active Region evolution and in the forecast of solar flare activity (see Steinegger et al [4]).
A sunspot is a dark cooler part of the Sun’s surface photosphere and is characterised by a strong magnetic field, formed below the photosphere, which extends out into the solar atmosphere and corona. Sunspots are best observed in the visible spectrum also known as ‘white light’. Sunspots can also be observed in CaII K1 absorption line images. Sunspots generally consist of the two parts: a darker, roughly circular central disk called the umbra, and a lighter outer area called the penumbra.

2 The Existing Methods of Detection

From the point of view of digital imaging the sunspots (as represented on white light, CaII k1, CaII k3 and H-alpha line images) can be generally characterised by the following two properties: they are considerably darker than the surrounding photosphere and they have well-defined borders, i.e. the intensity change occurs over reasonably short distance from photospheric value to the spot value. The existing techniques for sunspot detection can be divided into the three basic classes. A number of existing methods [1-9], called thresholding methods but also including region-growing techniques, rely on sunspots lower intensity variations. There are also methods, called border methods developed by Győr [14], Pettauer and Brandt [8], making use of the intensity gradient of the sunspot image. In addition, substantial work has been carried out on Active Region detection and labelling by means of Bayesian Pattern Recognition methods by Turmon, Pap and Mukhtar [13] that also incorporates sunspot detection (penumbra only).

The above mentioned methods, with the exception of Bayesian ones, can be described as semi-automatic techniques since they require a user participation of some kind (for verification, threshold choice, choice of input image for instance). At the same time, all these methods are data specific in the sense that they were developed for specific data sets and, hence, make a number of assumptions about the data photometric properties, image resolution and presence of image artefacts.

3 Data Description and Preprocessing

It can be observed that, in general, the results of sunspot detection on digital images depend on the following: seeing conditions (for ground-based observatories); wavelength (more spots in white light than Ca II k1); instrumental accuracy and artefacts; image resolution (smaller sunspots/pores may not be seen in smaller resolution images).

MDI images taken aboard the SOHO satellite with continuous (4 synoptic images per day) coverage since May 1995 are also characterised by extensively descriptive and precise header information and absence of seeing (weather) condition effects. This makes this dataset very attractive for our purposes, notwithstanding its relatively low resolution of 2 arc seconds per pixel. To improve the data coverage and provide catalogue continuity we have extended our methods to sunspot detection on the Meudon Observatory daily Ca II k1 line images. Both data sets cover the time period spanning April 1, 2002 to April, 30, 2002. The SOHO/MDI data for the entire year 2002 have been also processed.
The SOHO/MDI instrument provides almost continuous observations of the Sun in the white light continuum in the vicinity of the Ni I 676.7 nm line with resolution comparable to the ground-based telescopes. For the SOHO/MDI intensity spectrograms the pre-processing stage was as follows. For the pixels outside of the solar disk, as specified by the FITS file header information, the intensity values were set to zero, thus taking image dynamic intensity range into the non-negative integers set. Parameters such as disk centre, resolution, date of observation, disk radius were extracted from the FITS file header and solar disk intensity was then flattened by compensating for the limb darkening curve, see Zharkova et al. [11] for details, thus producing a flat white light image, on which sunspot detection is run. For a flat image, the Quiet Sun Intensity value, \( I_{QSun} \), is defined as the most populated non-zero intensity (i.e. as the intensity with the highest pixel count, see Figure 1).

While our detection method relies mainly on intensity properties of the white-light image, SOHO/MDI magnetogram data were used for verification purposes. In most cases, it is possible to locate a magnetogram virtually simultaneous (observations made within 30 seconds of each other) with the continuum observations. Since time difference between these observations is always under 2 hours, we synchronise both observations to a ‘continuum point of view’ by rotating the magnetogram data to the ‘continuum’ time of observation using SolarSoft procedures.

Fig. 1. Determination of the Quiet Sun Intensity as the highest pixel count from the histogram of a “flattened” photospheric image
4  The Sunspot Detection Method for the SOHO/MDI Images

Following the discussion in the Introduction, in order to extract as much information as possible from the chosen data we combine the thresholding and border methods. In order to avoid dependency on the choice of a global intensity threshold, edge detection is used instead. By examining the features with well defined borders we are then able to apply thresholding methods locally. Good results are achieved for noisy ground-based images using either the morphological gradient or Sobel operators.

The detection code is applied to a SOHO MDI continuum “flattened“ full disk image, $\Delta$ (Figure 2, top left image), with determined quiet Sun intensity, $I_{QSun}$ (Figure 1). image size, solar disk center pixel coordinates, disk radius, date of observation, and resolution (in arc seconds per pixel). A SOHO/MDI magnetogram, $M$, is synchronised to the continuum image via a (temporal) rotation and a spatial displacement to obtain the same point of view as the continuum.

The detection algorithm is described below in the pseudo-code.

1. Apply histogram equalization to increase a contrast (if required, optional)
2. Apply Gaussian smoothing with sliding window 5x5 followed by a Sobel operator to a copy of $\Delta$;
3. Using the initial threshold value, $T_0$, threshold the edge map and apply the median filter to the result. Count the number of connected components (Feature Candidates, Figure 2, top right). If it is too large, increase $T_0$ and repeat step 3 from the beginning.
4. Remove the edge corresponding to the limb from Candidate Map and fill the possible gaps in the feature outlines using IDL’s morphological closure and watershed operators to define a candidate feature, $F_i$, as a set of pixels representing a connected component on the resulting binary image, $B_\Delta$ (Figure 2, second row, left).
5. Create an empty Sunspot Candidate Map – a byte mask for the original input image indicating detection results with pixels belonging to umbra marked as 2, penumbra as 1.
6. For every $F_i$ extract a cropped image containing $F_i$ and define the
   i. if $F_i <= 5$ pixels assign the thresholds:
      for penumbra $T_s = 0.91 \ I_{QSun}$; for umbra $T_u = 0.6 \ I_{QSun}$
   ii. if $F_i > 5$ pixels assign the thresholds:
      for penumbra: $T_s = 0.93 \ I_{QSun}$;
      for umbra: $T_u = \max \{ 0.55 \ I_{QSun} ; (\langle P \rangle - \Delta P) \}$,
   where $\langle P \rangle$ is mean intensity value and $P$ a standard deviation for $F_i$
7. Threshold a cropped image at this value to define the candidate umbral and penumbral pixels and insert the results back into \( \Delta B \) (Figure 2, second row, right).

8. To verify the detection results, cross check \( \Delta B \) with \( \mathcal{M} \), as follows:

   for every connected component \( S_i \) of \( \Delta B \) extract

   \[
   B_{\text{max}}(S_i) = \max(\mathcal{M}(p) \mid p \in \mathcal{M})
   \]
   \[
   B_{\text{min}}(S_i) = \min(\mathcal{M}(p) \mid p \in \mathcal{M})
   \]

9. If \( \max(\text{abs}(B_{\text{max}}(S_i)), \text{abs}(B_{\text{min}}(S_i))) < 100 \) then disregard \( S_i \) as noise.

10. For each \( S_i \) extract and store the following parameters: gravity center coordinates (Carrington and projective), area, diameter, umbra size, number of umbras detected, maximum-minimum-mean photometric intensity (as related to flattened image), maximum-minimum magnetic flux, total magnetic flux and total umbral flux.

5 The Results and Conclusion

The results of sunspot detection on the image taken on the 2nd April are presented in Figure 2, (second row, right) with a closer view of a particular sunspot group presented in the third and fourth rows of Figure 2. The technique has been used to process one year of SOHO/MDI data with results stored and shared over the Internet. The ability to verify the detection results by checking the magnetic data helped to increase the detection accuracy (both in terms of False Acceptance Rates and False Rejection Rates) on the SOHO/MDI images and to detect the presence of the smaller pores, which are normally detectable on the images with a higher spatial resolution.

Since daily Wolf Numbers are primary indicators of sunspot activity extracted manually, in order to be able to statistically compare our detection results with manual catalogues we have to develop our method further to classify detected sunspots into sunspot groups, thus generating Wolf numbers.

A manual comparison of the sunspots detected on the SOHO/MDI images using the above technique with the sunspot drawings for June-July, 2002, produced in Locarno Solar Observatory, revealed an excellent (about 98%) agreement between the data sources and detected features with minor differences naturally present due to the differences in spatial resolution, observation time and seeing conditions for the ground-based observatory.

The accuracy of the technique developed for the sunspot detection on the Meudon Observatory Ca II k1 images, with more noisy backgrounds, was tested by a comparison with the manual synoptic maps for sunspots generated at the Meudon Observatory (Zharkov at al., 2003). Comparison was based on the data for April ’02 with the false rejection and acceptance rates, FRRs and FARs, calculated for daily observations that did not exceed in average 0.8 and 1.3, respectively. The discrepancies occurred because the Meudon images have more noise from the atmospheric conditions.
Fig. 2. An example of sunspot detection on a SOHO/MDI white light continuum image taken on 2nd April 2002
as well as because of the usage of extra knowledge in their manual technique, i.e. placing the sunspots on the map when they are not actually seen on the image. The SOHO/MDI data was concluded to be a preferable data source for the same period of observations since it provides higher accuracy and a much better coverage while the features visible in Ca II k1 are a subset of those seen in ‘white light’.

In summary, the new technique for automated sunspot detection on full disk white light SOHO/MDI images achieved the detection of sunspots with an excellent accuracy, the extraction of sunspot locations, umbral/penumbral areas, diameters, irradiance and their correlation with magnetic field variations.

This technique is used for building the Sunspot Feature Catalogue for the European Grid of Solar Observations (EGSO) funded by the European Commission within the IST Fifth Framework, grant IST-2001-32409.

References

1. Hoyt, D.V., & Schatten, K.H. Solar Physics, 1998b; 181: 491-512
2. Temmer, M., Veronig, A., Hanslmeier, A. Hemispheric Astronomy and Astrophysics. 2002; 390, 707-715
3. Steinegger, M., Vazquez, M., Bonet, J.A. and Brandt, P.N. Astrophysical Journal. 1996; 461: 478.
4. Chapman, G. A. and Groisman, G. Solar Physics, 1984, 91: 45
5. Steinegger, M., Brandt, P. N., Pap, J., and Schmidt, W. Astrophysics and Space Science. 1990; 170: 127-133
6. Brandt, P. N., Schmidt, W., and Steinegger, M., Solar Physics 1990; 129: 191
7. Beck, J.G. Chapman, G.A. A study of the contrast of sunspots from photometric images. 1993, Solar Physics, 146: 49
8. Pettauer, T., Brandt, P.N. Solar Physics, 1997, 175: 197
9. Steinegger, M., Bonet J., A., Vazquez, M.. Solar Physics, 1997; 171: 303
10. Preminger, D. G., Walton, S.R., Chapman, G.A. Solar Physics, 2001, 202: 53
11. Zharkova, V.V., Ipson, S. S., Zharkov and 3 other authors Solar Physics, 2003; 214: 89-105
12. Győrő, L., Solar Physics, 1998; 180:109-130
13. Turmon, M., Pap, J. M., Mukhtar S., ApJ, 2002; 568:396-407
14. Zharkov, S., Zharkova V.V., Ipson, S.S., Benkhalil A., Proc. of the AISB’03 Symposium on Biologically-inspired Machine Vision, Theory and Application, University of Wales, Aberystwyth, 7th-11th April 2003, pp. 74-84, ISBN 1-902956-33-1, 2003
A Procedure for the Automated Detection of Magnetic Field Inversion in SOHO MDI Magnetograms

S.S. Ipson, V.V. Zharkova, S.I. Zharkov, and A. Benkhalil

Department of Cybernetics, University of Bradford, BD7 1DP, UK
s.s.ipson@bradford.ac.uk

Abstract. Magnetic inversion lines are generally constructed from line-of-sight magnetic field data which have been smoothed to reduce the resolution scale of the data. This eliminates the fine details of the magnetic inversion lines found in regions of strong magnetic field. The paper presents a new approach to constructing magnetic neutral lines, based on a distance transform, which aims to construct neutral lines retaining fine detail in regions of strong magnetic field while reducing the detail elsewhere. The method of implementation is described and results obtained are compared with those obtained by applying Gaussian smoothing to solar magnetograms and with filaments visible in an H image for 2002 July.

1 Introduction

Solar phenomena such as sunspots, filaments, flares and active regions vary with an 11 (22) year cycle, which means they are connected with the behaviour of the solar magnetic field [1], [2]. The solar magnetic field measured with the line-of-sight (LOS) ground-based or space-based magnetographs ([3], [4]), or recent full vector magnetographs ([5], [6]) provides an important data resource for understanding and predicting solar activity. The importance of magnetic field has been emphasized in many theoretical works including: heating of the solar atmosphere; formation and support of multi-levelled magnetic loops and filaments resting on top of them; activation of energy releases in solar flares and coronal mass ejections; and in many other small and large scale events occurring in the solar atmosphere and interplanetary space. It is thought that different scales of magnetic field magnitudes account for solar events of different scales and that event scenarios result mostly from magnetic configurations of loops with opposite magnetic polarities. For example, solar filaments often appear on the boundaries of coronal holes ([7], [8], [9]) or above a middle line between two-ribbon flares ([10], [11]) that usually is very close to the location of magnetic inversion lines, or magnetic neutral lines (MNLs).

This paper is concerned with the construction of magnetic neutral lines from line-of-sight magnetograms and four different approaches have been identified in the literature. In the vicinity of active regions contour methods have been used. An example of this approach is the work of Falconer et al. [12], investigating neutral-line magnetic shear and enhanced coronal heating in solar active regions. They defined the MNL to be the region between the -25 G and +25 G contours. In the regions of interest MNLs
were found to be 2 pixels wide or less. However, away from active regions, the regions between the -25 G and +25 G contours become very wide. Several authors have presented methods for identifying magnetic field boundaries between large regions of predominantly positive or negative polarity. For example, Bornmann et al. [13], used morphological operators to extend the strongest positive and negative regions which were then added to the original magnetograms and smoothed. Two thresholds were used to identify the limits to the neutral line locations. Durrant [14] remapped magnetograms to a latitude-longitude grid applied smoothing over circles of diameters 60, 120 and 180 arc seconds and then transformed back to the original geometry. Smoothed magnetograms were displayed using 256 shades of grey with positive magnetic fields using the upper half of the range with increasing field magnitude displayed with increasing darkness and negative magnetic fields use the lower half of the grey range with increasing field magnitude corresponding to increasing darkness. The zero-field contour is marked by the discontinuity in the greyscale between mid-grey and peak white. Ulrich et al [15], reported a method of averaging magnetogram data over multiple observations at different times, after taking differential rotation at different latitudes into account, and extracting the east-west and meridional components of the slowly evolving large-scale magnetic field to get improved magnetic and neutral line synoptic maps.

This paper presents an approach to determining magnetic neutral lines which aims to retain the local detail of MNLs in regions of high field while simultaneously constructing boundaries between larger scale positive and negative magnetic regions.

2 Method

Solar magnetograms are produced by a number of the ground and space based observatories. However, this paper is concerned with magnetograms produced by the Michelson Doppler Instrument (MDI) installed on the SOHO spacecraft. The magnetograms produced by this instrument provide an array of 1024 by 1024 pixel measurements of the line of sight component (the dominant component of the magnetic vector field in the central portion of the disk) of the solar magnetic field in the photosphere. The radius of the solar disc in these magnetograms is about 489 pixel (corresponding to a pixel resolution of about 2 arc sec) and the magnetograms are produced at a rate of one per minute except during periods of occasional down-time. The magnetogram pixel values are stored using 16 bits, with a signal range from about -1500 G to +1500 G and a noise level of about ±20 G.

Several studies, (e.g. [16], [17], and [18]) have shown that the general magnetic field of the sun consists of small (1-2 arc sec) but intense (~ 1200 G) flux knots of mixed polarity. In some regions the flux knots are fairly evenly distributed while in others one polarity dominates and at low resolution observation masks those of the other polarity to give what is often described as a unipolar region. In a typical example of an MDI magnetogram from SOHO the boundary between adjacent positive and negative magnetic regions is well defined only at comparatively few locations between local concentrations of opposite field associated with sun spots. Elsewhere, positive (and negative) regions are indicated by the presence of mainly positive (or negative) fluctuations from the ±20 G, near zero, range of magnetic field intensity defined by
the noise level. The boundaries between positive and negative regions of an MDI magnetogram generally have a great deal of fine structure, as is hinted at by the magnetograms shown at the top left corner in Figs. 1 and 2 which have the minimum smoothing in the two sets of magnetograms. This fine detail can be reduced by local averaging of individual magnetograms (e.g. [14]) or by averaging, after differential rotation, all available magnetograms for a Carrington rotation (e.g. [15]). This paper examines another approach which is applied to a single magnetogram and is capable of retaining the structure of neutral line boundaries in high field regions, while reducing the fine detail elsewhere.

The principal idea for the automatic construction of boundary lines separating adjacent collections of small positive magnetic features from adjacent collections of small negative magnetic features is to grow the features isotropically (using a distance transform) and to mark the points where the positive and negative regions make contact. The resulting points are used as estimates of points along inversion lines. The initial seed regions are obtained by applying positive and negative thresholds of equal magnitude which defines the solar disk into three regions: positive regions, negative regions and neutral regions within which the neutral lines are found. The specific steps used to construct the inversions lines in this illustration are as follows.

The magnetogram data are first read from the fits file and stored in a floating point array. The data is then segmented into the three types of magnetic regions by applying a two level magnetic field threshold \( \pm T \), where \( T \) could be set at the noise level 20 Gsay. A fast Euclidean distance transform, based on the approach of Cuisenaire and Macq [19], is applied to determine the distances of the pixels in the neutral region from the nearest segmented boundary pixel. This is implemented as indicated in the following pseudo code. First three array buffers, \( x, y \) and \( z \), with dimensions the same as the magnetogram are initialised. The elements of \( x, y \) and \( z \), store the horizontal and vertical spatial offsets and the distances from the nearest object respectively. The array elements with positions corresponding to the positive and negative regions defined by the threshold are initialised to zero and the remaining elements are set to values larger than any which will be encountered during the computation.

A first pass is made through the data, from left to right and bottom to top, updating the buffer elements in sequence. During this pass, at the end of each row a reverse scan is made of each row, updating the buffers.

Initialise buffers \( x[N, N], y[N, N] \) and \( z[N, N] \)
Set row index \( I = 1 \)
Set column index \( J = 1 \)
If current position has \( x[J, I] \) or \( y[J, I] \) greater than 0
  If \( J > 1 \) Compute \( dx = x[J-1, I] + 1, dy = y[J-1, I], d = dx^2 + dy^2 \)
  If \( d < z[J, I] \)
    \( z[J, I] = d, x[J, I] = dx, y[J, I] = dy \)
  if \( I > 1 \) Compute \( dx = x[J, I-1], dy = y[J, I-1] + 1, d = dx^2 + dy^2 \)
  If \( d < z[J, I] \)
    \( z[J, I] = d, x[J, I] = dx, y[J, I] = dy \)
Increment \( J \) and, if \( J \leq N \), repeat
Set column index \( J = N-1 \)
If current position has $x[J, I]$ or $y[J, I]$ greater than 0
Compute $dx = x[J + 1, I] - 1$, $dy = y[J + 1, I]$, $d = dx^2 + dy^2$
If $d < z[J, I]$
$z[J, I] = d$, $x[J, I] = dx$, $y[J, I] = dy$
Decrement $J$ and, if $J >= 1$, repeat
Increment $I$ and, if $I <= N$, repeat

To complete the transform, a second pass in the opposite direction from right to left and top to bottom is made through the data up. The pseudo code for this second pass, shown below, is similar to the previous code apart from the change of direction.
Set row index $I = N - 1$
Set column index $J = N - 1$
If current position has $x[J, I]$ or $y[J, I]$ greater than 0
If $J < N$ Compute $dx = x[J+1, I] - 1$, $dy = y[J+1, I]$, $d = dx^2 + dy^2$
If $d < z[J, I]$
$z[J, I] = d$, $x[J, I] = dx$, $y[J, I] = dy$
if $I < N$ Compute $dx = x[J, I+1]$, $dy = y[J, I+1] - 1$, $d = dx^2 + dy^2$
If $d < z[J, I]$
$z[J, I] = d$, $x[J, I] = dx$, $y[J, I] = dy$
Decrement $J$ and, if $J >= 1$, repeat
Set column index $J = 1$
If current position has $x[J, I]$ or $y[J, I]$ greater than 0
Compute $dx = x[J - 1, I] + 1$, $dy = y[J - 1, I]$, $d = dx^2 + dy^2$
If $d < z[J, I]$
$z[J, I] = d$, $x[J, I] = dx$, $y[J, I] = dy$
Increment $J$ and, if $J <= N$, repeat
Decrement $I$ and, if $I >= 1$, repeat

On completion of the second pass, the horizontal and vertical offsets of each pixel in the neutral region from the nearest object have been computed. These offsets are used together with the segmented array to identify the polarities of the nearest objects, which then replace the neutral pixel values with the appropriate polarity labels. The resulting boundaries between the positive and negative labelled regions (forming a Voronoi tessellation) indicate the estimated positions of the neutral lines. The neutral lines are marked explicitly by applying a morphological edge detector, to the segmented array.

3 Discussion

The general affect that the value of threshold $T$ has on the resulting neutral lines is illustrated by the sequence of images shown in Fig. 1. From left to right along the top row and then from left to right along the bottom row, the threshold varies in steps of 10 G from 20 to 90 G. The result is a reduction in the number of regions and a simplification of the boundaries of the resulting regions except where the neutral line separates two regions of strong magnetic field. These latter sections of neutral lines can be identified within the images with larger threshold values as regions where the neutral
A Procedure for the Automated Detection of Magnetic Field Inversion

line possesses fine structure. For comparison, a similar sequence of magnetograms, but with increasing amounts of Gaussian smoothing, is shown in Fig. 2. No attempt has been made to make the smoothing applied to the individual magnetograms in Fig. 2 correspond to the reduction in detail obtained by increasing the threshold value applied to the magnetograms in Fig. 1. Nevertheless, over the central two thirds of the solar magnetograms, in the low field regions, the reduction of detail as the smoothing or threshold value is increased produces similar results. Near the limb, the results are distorted by foreshortening and by boundary effects. Durrant’s method [14] reduces this effect by mapping the observed magnetograms to a latitude-longitude grid, applying smoothing and then mapping the result back to the solar disk and this approach could be applied with the distance transform too.

**Fig. 1.** Magnetic neutral lines indicated by the boundary between dark (negative) and light (positive) regions for thresholds varying from 20 G to 90 G from top left bottom right. MDI July 23 2002

Examples of magnetic neutral lines, constructed using the distance transform method and using Gaussian smoothing, superimposed on an H-alpha image which has been rescaled and registered to the size and position of the solar disc in the original magnetogram are shown in Fig. 3. It is evident from this figure that in both cases there is a strong correlation between the locations of most of the filaments and inversion lines. The biggest discrepancy is for the large and prominent filament near the top of the image. It lies in a region of positive magnetic polarity and only for lower smoothing/threshold values do its foot points appear close to neutral lines.
Fig. 2. Neutral lines indicated by the boundary between dark (negative) and light (positive) regions for smoothing radii varying from 5 to 40 from top left bottom right. MDI July 23 2002. In each case the full width of the Gaussian kernel corresponds to 5 standard deviations

Fig. 3. The composite images show magnetic neutral lines superimposed on an H-alpha image (July 23 2002) from the Meudon observatory. The left hand composite image shows magnetic neutral lines using T = 70 G superimposed on the original magnetogram while the composite image on the right shows magnetic neutral lines using a Gaussian smoothing with a radius of 20 pixel

4 Conclusions

Two quite different methods of estimating the positions of magnetic neutral lines have been compared. The results found are similar although as the magnetograms are simplified, in one case by increasing smoothing and in the other by increasing threshold,
the distance transform method retains the fine structure in the strong magnetic field regions. In the cases examined most of the filaments are found to be close to neutral lines as suggested in previous literature. This research has been done for the European Grid of Solar Observations (EGSO) funded by the European Commission within the IST Fifth Framework, grant IST-2001-32409.

References

[1] Priest E. R, “Solar magneto-hydrodynamics”, Geophysics and Astrophysical Monographs, Dordrecht: Reidel, 1984.
[2] Priest E. and Forbes T, “Book Rev.: Magnetic reconnection” Cambridge U Press, 2000.
[3] Babcock H. V. and Babcock H. D, “The Sun's Magnetic Field, 1952-1954”, Astrophysical Journal, vol. 121, pp.349-366, 1955.
[4] Scherrer P. H. et al., “Annual Review”, Astr. Astrophys., Vol. 2, 363, 1995.
[5] Wang H, Denker C, Spirock T, and 7 other authors, “New Digital Magnetograph At Big Bear Solar Observatory” Solar Physics, vol. 183, Issue 1, p. 1-13, 1998.
[6] Ulrich R. K, “In Cool Stars, Stellar Systems and the Sun”, edited by M. S. Giampapa and J. A. Bookbinder, Astron. Soc. of the Pacific, San Francisco, Calif., p. 265, 1992.
[7] Kippenhahn, and Schluter “Eine Theorie der solaren Filamente”, Zeitschrift für Astrophysik, vol. 43, p.36-62, 1957.
[8] Kuperus M. and Raadu M. A, “The Support of Prominences Formed in Neutral Sheets”, Astronomy and Astrophysics, Vol. 31, pp. 189-193, 1974.
[9] Lerche I. and Low B. C, “Cylindrical prominences and the magnetic influence of the photospheric boundary2, Solar Physics, vol. 66, pp. 285-303, 1980.
[10] Somov B. V, “Cosmic Plasma Physics”, Astrophysics and Space Science Library, v. 252. Boston, Mass.:Kluwer Academic Publishing, 2000.
[11] Sturrock P. A. and Jardin M, “Book Review: Plasma physics” Cambridge U Press, 1994.
[12] Falconer D. A, Moore R. L, Porter J. G, and Gary G. A, “Neutral-line magnetic shear and enhanced coronal heating in solar active regions”, Astrophysical Journal, Vol. 482, pp. 519-534, 1997.
[13] Bornmann P. L, Winkelman J. R, Cook D, Speich D, “Automated solar image processing for flare forecasting”, Solar Terrestrial Workshop, Hitachi: Japan, pp. 23-27. 1996.
[14] Durrant C. J, “Polar magnetic fields – filaments and the zero-flux contour”, Solar Phys., vol. 211, pp. 83-102, 2002.
[15] Ulrich R. K, Evens S, Boydjen J. E. and Webster L, “Mount Wilson synoptic magnetic fields: improved instrumentation, calibration and analysis applied to the 2000 July 14 flare and to the evolution of the dipole field”, Astrophysical Journal Supplement Series, vol. 139, No. 1, pp. 259-279, 2002.
[16] Severny A, “Vistas Astron” Vol. 13, p 135, 1972.
[17] Howard R. F, “Annual Review”, Astron. Astrophys. Vol. 15, p 153, 1977.
[18] Zwaan C, “Annual Review”, Astron. Astrophys. Vol. 25, p 89, 1987.
[19] Cuisenaire O. and Macq B, “Fast and exact signed Euclidean distance transformation with linear complexity”, In Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP99), vol. 6, pp. 3293-3296, Phoenix (AZ), March 1999.
Automatic Detection of Active Regions on Solar Images

A. Benkhalil, V. Zharkova, S. Ipson, and S. Zharkov
Department of Cybernetics, University of Bradford, BD7 1DP, UK
a.k.benkhalil@brad.ac.uk

Abstract. In this paper techniques are described for the automated detection of solar Active Regions (ARs). AR detection is achieved using intensity thresholds and a region growing procedure. These procedures have been tested on full-disk solar images from the Meudon observatory for the months of April and July 2002 and compared with their manually generated synoptic maps. Comparisons were also made with AR data published by the National Oceanic and Atmospheric Administration observatory (NOAA) and very good correspondence was found.

1 Introduction

There are a growing number of archives of digitized images of the Sun taken from ground-based and space-based instruments in various wavelengths. These archives are available from different locations and are to be unified by the European Grid of Solar Observations (EGSO) project [1].

There are three different approaches identified in the literature for the automatic identification of bright ARs (plages). The first is based on the selection of a threshold to separate the object from a background and is straightforward if the intensity histogram is bimodal, but otherwise can be difficult [2]. The second approach is based on region growing techniques segmenting images into bright and dark regions [3, 4] and is applied to solar images in various wavelengths, including H, from a number of sources. Finally, the third approach uses the Bayesian inference method for automatically identifying various surface structures on the Sun [5]. All these approaches can give a reasonable accuracy of detection with suitable images, but the Bayesian based methods are the most computationally expensive of the three. The intensity threshold-based methods are simple and fast, but are relatively sensitive to noise which affects the reliability of the segmentation results obtained.

In order to replace the existing manual detection methods, the current paper presents techniques combining elements of the first two approaches above for the automated detection of ARs (plages) at different heights in the solar atmosphere, which are revealed at different wavelengths. The types of solar images which have been used are H and CaIIK3 line spectroheliograms from the Meudon observatory. The two types of image are processed using the same region growing technique but with different intensity thresholds in the initial processing stage, to find seed locations, and in the
subsequent AR segmentation stage. The methods used for AR detection and comparison of the resulting identified regions are discussed in Section II. Conclusions are given in Section III.

2 The Techniques for Active Regions Recognition

There are two basic assumptions made about the solar images when applying the developed techniques. The first is that the input images are standardized with size 1024 pixel×1024 pixel, solar disk radius 420 pixel, solar centre 511.5 pixel × 511.5 pixel, and free of a radial limb darkening. In order to comply with this assumption, the techniques are applied to full-disk high-resolution solar images which have been standardized using procedures [6] for limb fitting, shape correction and limb darkening removal. At the same time the images are also transferred to the required format. The second assumption is related to the properties of ARs, and is simply that they are the brightest features on the solar disk. This means that the intensity values inside the detected regions of interest are greater than the intensity values of the local background.

2.1 The Initial Segmentation

In order to define a suitable local threshold all images were first remapped into polar coordinates with origin at the solar disc centre. After remapping, localized intensity thresholds, with values derived as explained below, are used for an initial segmentation of the bright plages. Pixels whose intensity values are over this intensity threshold have their values set to 1 and all other pixels have their values set to zero. The choice of these initial intensity threshold values is very important because a value that is too high may lead to real features being missed, whereas a value that is too low may lead to noisier binary images and, hence, spurious features. The optimum global threshold value also varies with the image brightness levels and the non-radial large scale intensity variations, which are present in some Meudon images, are a particular problem. To overcome these problems optimized local intensity threshold values \( T \) are calculated for quarter-sized regions of an image as follows:

\[
T_i = \mu_i + (1 + \Delta_i) \times \sigma_i
\]

where \( \mu_i \) is the mean intensity value for the region \( i \), \( \sigma_i \) is the standard deviation of the intensity for the same region and \( \Delta_i \) is a constant that was set to 0.4 after investigating more than 30 images.

The main stages of this technique are illustrated in Fig. 1 for H\( \alpha \), and in Fig. 2 for CaIIK3 full-disk images, respectively. Subfigures (a) present the cleaned initial images; subfigures (b) show the results of their remapping into the polar coordinates. The results of the initial segmentation based on equation (1) are presented in subfigures (c). Subfigures (d) and (e) show initial and final segmentation results discussed below.
2.2 Noise Reduction and Region Labeling

The initial segmentation will generally include noise and unwanted small features caused by the over-segmentation. Over-segmentation is preferable to under-segmentation as the former can be remedied using Median filtering and Morphological operations whereas the latter could lose significant information. Firstly, in order to remove small features a 7×7 Median filter is used. The size was chosen through experimentation. Then Morphological opening and closing operations are applied using a structure element of size 8×8. This smoothes the features and fills in holes. Figures 1(d), 2(d) and 3(d) show the detected regions after applying Median and Morphological processing and transformation back to Cartesian coordinates. As can be seen, the noise and over-segmentation problems have been remedied.

The result of this initial segmentation is a set of the segments, each of which corresponds to an AR present on the solar disk. Every segment is labeled, and its centroid is calculated for use as a seed in a region growing procedure. Prior to this, because the shape may be complex, the location of the seed is checked in order to ensure that it is inside the region and, if not, its position is adjusted. In this case a new seed is selected by investigating pixel values in the eight nearest neighbor directions, until a new seed is found inside the region from which the region growing procedure can start.

2.3 The Region Growing Technique

The region growing procedure starts with a set of seed pixels and aims to grow a uniform and connected region from each seed. A pixel is added to a growing region if and only if:
Automatic Detection of Active Regions on Solar Images

- It has not been assigned to another region
- It is an 8-neighbour of the growing region
- The extended region created by the addition of a new pixel is still uniform.

The region growing algorithm takes as input a standardized image and a corresponding set of seed points obtained by the procedure described in section 2.2. The algorithm begins at each seed pixel and scans the neighboring 8 pixels in a circular fashion, to determine a membership of the region around the central pixel that complies with the rules above and the following constraints. Two forms of constraint have been considered. The first uses a fixed threshold range (with an upper and lower pixel value) and second uses a variable threshold range set to a multiple factor of the standard deviation of the pixel values in the current region. After experimentation the fixed threshold range was chosen as it was found to give more accurate control in defining the outer boundaries of regions while also reducing the occurrence of holes in the regions.

The upper and lower threshold values within initially detected ARs are determined by exploiting the statistical properties of the locally homogeneous background regions. The lower threshold value is defined as $\mu - 0.3 \sigma$ (where $\mu$ is the mean and $\sigma$ is the standard deviation of that region) and the upper threshold is set to the maximum intensity value of the image. As pixels are added, the process repeats with the newly added pixels as the central pixels. A list of the tested pixels is maintained, in order to avoid unnecessary duplication of the tests. In this way, the region is constructed by using the membership criteria already discussed. If in a binary image more than one seed pixel has been obtained from an AR, the region growing method will merge the detected pixels to form a single contiguous area of AR. Figs. 1(e) and 2(e) show the final results of applying the region growing procedure in the Hα and CaIIK3 images, respectively.

### 2.4 AR Verification Using Magnetograms

It is not possible to confirm or deny the identification of ARs on the basis of CaIIK3 data alone and for this reason we started to look at magnetogram data also. Series of FITS file format Michelson Doppler Imager (MDI) images are obtained from the MDI instrument aboard the SOHO satellite. These images are one-minute cadence, full-disk magnetograms. The CaIIK3 images from ground observations at Meudon are acquired once a day. As The CaIIK3 and MDI images are taken at different time and from different locations (and have different sizes) the images need to be synchronized.

Firstly we select and download the MDI image that is closest in time to the Meudon CaIIK3 observation. This is generally straightforward as we have MDI image every minute. Secondly, to correct MDI-image sizes and to convert them to earth view, the MDI images are converted to Interactive Data Language (IDL) map objects using the index2map.pro IDL function. The Map object is a structure that contains 2D image data with accompanying pixel coordinate and spatial scale information. The map is converted from SOHO-view map to Earth-view using the map2earth.pro function. To make the dimensions and pixel spacing of a SOHO-MDI map object identical to the CaIIK3 images (i.e. size 1024 pixel×1024 pixel, radius 420 pixel and solar centre
511.5 pixel ×511.5 pixel) we use the grid_map.pro function. Finally the map objects are converted back to index images using the map2index.pro function. The final results are MDI images synchronized to earth view and standardized in shape and size to the CaIIK3 Meudon images.

Once an AR is detected the system will automatically crop two images, with identical sizes and locations, one from the CaIIK3 image and one from the MDI image. The latter data is then checked to identify positive and negative polarity regions which, if found, are used to confirm the detection of an AR and also to classify the ARs into newborn, highly active and decaying categories.

Fig. 3 shows the results of tracking and investigating an AR (NOAA 9893) over three days 8-10/04/2002. The left column contains the CaIIK3 cropped images, the middle column contains the MDI cropped images, where the black areas indicate positive polarity and the white area indicate negative polarity, and the right column shows the magnetic field superimposed on the CaIIK3 images.

![Fig. 3. Cropped images containing AR NOAA 9893. Left column CaIIK3, middle column magnetogram and right column showing magnetic field polarity superimposed on CaIIK3 image](image)

2.4 The Accuracy of the Technique

The procedures have been tested on synoptic image sequences of full-disk solar images from the Meudon observatory for the months of April and July 2002. For further testing the results obtained from the Meudon images were compared with those of the NOAA observatory as illustrated in Fig. 4.

A quantitative comparison of the results obtained using the present technique, with those done manually at the Meudon observatory and at NOAA observatory was done for the two month (April and July 2002). In comparison with the other results, those from Meudon detect about 50% more ARs on most days. For example, on the
30/07/2002, and as shown in Fig. 4, there were 24 ARs included in the Meudon results while our procedure detected only 11 ARs and the NOAA observatory showed only 12 ARs. In order to quantify these differences the False Acceptance Rate (FAR) (where we detect an AR and they do not) and the False Rejection Rate (FRR) (where they detect an AR and we do not) were calculated for every day. In most cases there are a higher number of ARs detected by us than by NOAA with an average FAR of 1.8 per day in April and only 1 in July. The FRR was very low at about 0.2 in both months, with only 5 days in each month when we failed to detect a region detected by NOAA. In some cases we detect an AR while NOAA splits it into two regions. This does affect the quantitative comparison.

Fig. 4. A comparison of AR detection results. The present results for H (a), CaIIK3 solar images (b) showing 12 ARs. Results from the NOAA observatory (c) showing 12 ARs. Results from the Meudon observatory (d) showing a map of 24 ARs.

We believe the reason for these different results is due to a difference in definitions of ARs. At Meudon all bright regions (plages) are detected and these are defined as the regions in the chromosphere which are brighter than the normal “quiet” Sun background. At NOAA a detected AR is defined as a bright area on the Sun with a large
concentration of magnetic field, often containing sunspots. However, not all plages contain a strong magnetic field as they might be decaying ARs with a weakening magnetic field [7]. Fig. 4 clearly illustrates this case by showing the results of ARs detection at NOAA, Meudon and using our techniques with Hα and CaIIK3 images on the same day (30/07/02). In Fig. 4(e) the Meudon map shows 24 ARs (all the white area (plages) are counted) resulting in double the number detected by us and by NOAA. In general, the agreement with NOAA is good, considering that NOAA bases its decisions on more information about magnetic field than we do at this stage.

3 Conclusions

In this paper an efficient procedure for the automated detection of solar ARs is presented. Comparisons with other recognition results derived manually show that the procedures developed have achieved a satisfactory accuracy in the automated detection and segmentation of ARs on full disk Hα and CaIIK3 solar images from Meudon. The automated solar AR detection and characterization is one component of a larger project concerned with establishing a Solar Feature Catalogue. This research is a part of the EGSO project funded by the European Commission within the IST Framework 5.

References

[1] Bentley R. D, “EGSO - the next step in data analysis,” Proceedings of the Second Solar Cycle and Space Weather Euro-conference, 24 - 29 September 2001, Vico Equense, Italy, Edited by Huguette Sawaya-Lacoste, ESA Publication SP-477, 2002.

[2] Steinegger M, and Brandt P. N, “On the determination of the quiet Sun centre-to-limb variation in Ca K spectroheliograms,” Solar Physic Journal, vol. 177, pp.287-294, 1998.

[3] Hill M, Castelli V, Chung-Sheng L, Yuan-Chi C, Bergman L, and Thompson B, “Solarspere: querying temporal solar imagery by content,” International Conference on Image Processing, Thessaloniki, Greece, 7-10 Oct. 2001, vol. 1, pp. 834–837, 2001.

[4] Veronig A, Steinegger M, Otruba W, Hanslmeier A, Messerotti M, and Temmer M, “Automatic Image Processing In The Frame Of A Solar Flare Alerting System,” HOBUD7, vol. 24, no. 1, pp.195-200, 2001.

[5] Turmon M, Pap J. M, and Mukhtar S, “Automatically Finding Solar Active Regions Using Soho/Mdi Photograms And Magnetograms,” Proc. SoHO 6/GONG 98 Workshop, Structure and Dynamics of the Interior of the Sun and Sun-like Stars, Boston, 1998.

[6] Zharkova V. V, Ipson S. S, Zharkov S. I, Benkhalil A. K and Aboudarham J, “A full disk image standardisation of the synoptic solar observations at the Meudon observatory,” Solar Phys., vol. 214/1, pp. 89-105, 2003.

[7] Driel-Gesztelyi L. V, “Emergence and Loss of Magnetic Flux on the Solar Surface,” Proc. SOLMAG the Magnetic Coupling of the Solar Atmosphere : Euroconference and IAU Colloquium 188, Santorini,Greece, 11-15 June 2002, pp.113-116, 2002.
Automatic Detection of Solar Filaments Versus Manual Digitization

N. Fuller and J. Aboudarham

LESIA - Observatoire de Paris-Meudon, 5 place Jules Janssen, 92190 Meudon, France
{Fuller, Aboudarham}@obspm.fr

Abstract. The technique which is presented here is based on careful attention to image cleaning in order to achieve robust automatic filament detection and avoid confusion between filaments and image defects, due to particular shapes of filaments. The main part of the detection process is based on seed selection and region growing. The procedures developed have been tested on four months of full-disk images from the Meudon Observatory. The results are compared with those manually generated in Meudon, for several hundred filaments. A very good correspondence is found, showing the robustness of the method described.

1 Introduction

Full disk solar images (spectroheliograms) are obtained at the Paris-Meudon observatory every day, weather permitting, (approximately 300 days per year) for several wavelengths. The spectroheliograms in which the filaments are best seen are taken in the H transition of Hydrogen (656.3 nm). Filaments appear as dark features on the solar disk, often very elongated and sometimes more compact (see figure 1). The monitoring of filaments is important in Solar physics because they are indicators of the Solar activity and also can be precursors of very energetic phenomena such as flares or coronal mass ejections which are very important in the frame of Solar-Terrestrial relationship.

Solar filaments have been manually detected for dozens of years from these daily observations. The increasing power of computers and the evolution of detection techniques now offer an opportunity to develop automatic detection codes. This has been done in the frame of the European Grid of Solar Observations (EGSO) [2], a project funded by the European Commission within the Information Society Technologies thematic priority of the Fifth Framework Program (IST-2001-32409).

Such automated detection methods have already been developed in order to track filaments in time. Gao, Wang and Zhou [4] used thresholding and region-based techniques, like region growing, to detect filament disappearances. Collin and Nesme-Ribes [3] used similar techniques to investigate the rotation of the Sun’s surface. More recently, Shih and Kowalski [7] proposed a method based on directional filtering to identify thin and elongated objects and Wagstaff et al. [9] a region based method to detect both filaments and sigmoids. All these methods give good results, but some-
times neglect the smallest, weakest or blob-like filaments. Our goal is to extract the most complete possible set of filaments from one observation, and we thus can’t make any postulate about the features shapes. As region growing has proved to be a reliable means to investigate such features, we have based our method on it, improving the way seeds and thresholds are chosen.

Prior to any feature recognition, the solar images have to be pre-processed in order to correct them for geometrical or photometric distortions. These preliminary steps are described in Zharkova et al. [10]. Correction of ellipticity, centering the disk in the image, removing center-to-limb darkening, etc. all make the images more suitable for further processing. These standardized images are further enhanced by removing defects specific to the measurement conditions such as non-geometrical darkening due to clouds during the observation or dark lines across the disk due to dust particles. We present here techniques based on median filtering and on the Hough transform for removing them.

![Image of a spectroheliogram showing several filaments](image)

**Fig. 1.** Meudon spectroheliogram showing several filaments (dark features)

We finally compare our results to those manually obtained and show that the automated techniques developed give very good results on a large set of images, even if subjectivity is not included in our work.
2 Image Cleaning

2.1 Intensity Normalization

The first cleaning process we use consists of normalizing the intensity over the disk. As we use a region growing method (see section 3) to segment the filaments, and because some of them can be very long, the variations of the background intensity ('Quiet Sun') should be as small as possible.

To identify the large scale variations of the intensity we first use median filtering with a large window. The median filter will eliminate the highest and lowest values (corresponding to bright plages and dark filaments) to give a first approximation of the background fluctuations which can then be subtracted from the original image. From this first normalized image we can more efficiently locate pixels corresponding to bright and dark regions with two suitable thresholds. These thresholds are obtained from the new image histogram, and the pixels values are replaced by their corresponding values in the background map. By applying median filtering again (with a smaller window) we then get a more reliable estimate of the large scale variations. The following pseudo code contains the details of the algorithm and figure 2 provides an example of the resulting intermediate images:

```
begin
  Rescale I to a smaller size: I_{small} // saves computer time
  B_{small} = median(I_{small}, W_{size}) // W_{size} is the filter size
  I'_{small} = I_{small} - B_{small} + mean(B_{small}) //Subtract and get back to original intensity level
  Hist = histogram(I'_{small})
  H_{M} = max(Hist)
  Let V_{M} be the intensity value corresponding to H_{M}
  Let V_{1} be the intensity value corresponding to H_{M}/a_{1} (V_{1}<V_{M})
  Let V_{2} be the intensity value corresponding to H_{M}/a_{2} (V_{2}>V_{M}) //a_{1}, a_{2} are constants
  Let be S the set of pixels in I'_{small} lower than V_{1} and greater than V_{2}
  I'_{small}[S] = B_{small}[S]
  B_{small} = median(I'_{small}, W_{size}/2)
  Rescale B_{small} to original size: B
  I_{n} = I - B' + mean(B') // I_{n}: normalized image
end
```

2.2 Dust Lines Removal

As spectroheliograms are generated by scanning the disk horizontally, any dust grain on the entrance slit leads to a dark straight line on the image. These lines are particularly difficult to differentiate from the filaments, especially if they overlap, and we thus need to remove them.
A binary image is first computed from the normalized image and its cumulative histogram used to control the ratio between the non-zero and zero pixels. A morphological thinning operator is then applied to thin the features with contiguous non-zero pixels to a thickness of one pixel. The Hough transform [1] of the thinned image is then thresholded at a value corresponding to the minimum number of points needed to identify a straight line. Finally the parameters of the identified straight lines give the locations of the lines of pixels whose values can now be corrected (see figure 3).
2.3 Sharpness Enhancement

In order to better define filaments contours and to detect the thinnest parts of the filaments more efficiently (see figure 4) we enhance the image sharpness using the Laplacian method [6].

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & -8 & 1 \\
1 & 1 & 1 \\
\end{array}
\]

Laplacian mask Filament before and after enhancement

**Fig. 4.** Laplacian mask and the enhancement effect on filament’s contrast

This sharpening is done by subtracting a Laplacian filtered version of the image from the original image. The Laplacian of an image is obtained by computing its second order derivative. The digital equivalent is to convolve the original image with a suitable mask (such a mask is shown in figure 4). Then, if A is the original image and B the result, the formulation is:

\[
B(i,j) = A(i,j) - A(i-1,j-1) - A(i,j-1) - A(i+1,j-1) - A(i-1,j) - A(i+1,j) - A(i-1,j+1) - A(i,j+1) - A(i+1,j+1) + 8\times A(i,j)
\]

3 Filament Detection

Once the image is fully cleaned, we can investigate the filaments using a region growing method based on the image grey level properties. The principle is to group pixels into larger regions if these pixels fall into a predefined intensity range. The procedure is started from a pixel or small region called a seed. This method is much more effective than applying a basic automatic threshold as it associates a grey level condition with a connectivity condition. The efficiency of the method will thus depend on the seed selection procedure and on the intensity range definition.

3.1 Seeds Selection

As we said above, the seed selection is a major step in the procedure. Getting too many of them might lead to false detections. Alternatively, missing some would
imply missing filaments. To select the seeds as precisely as possible we use a windowed threshold. In the first stage, pixels whose values are distant from the mean are discarded from the calculation of a second mean $M_w$, which better reflects the ‘Quiet Sun’ intensity within each window (i.e. mean without brightest and darkest regions). Then the standard deviation $\sigma_w$ is computed from the same set of pixels and the window threshold is given by:

$$T_w = M_w - \alpha_1 \times \sigma_w$$

[3] where $\alpha_1$ is a constant obtained after testing a large set of images. All pixels with values lower than $T_w$ are set to 1 and the rest are set to 0.

### 3.2 Region Growing

The next major step consists in growing the seed sub-regions into larger ones. Once again we have to be very careful when computing the intensity criteria. As we did for seeds, we use windows again to be sure to get the most reliable local statistics. These windows are centred on the seeds and their sizes depend on the seed dimensions in both Ox and Oy directions. The lowest value of the intensity range is set to 0 and the highest $T_{\text{max}}$ is obtained in the same way that we computed it for seeds:

$$T_{\text{max}} = M_w - \alpha_2 \times \sigma_w \quad \text{with} \quad \alpha_2 < \alpha_1$$

![Original cleaned image](image1)

![Detected filaments superimposed](image2)

**Fig. 5.** Filament detection result

Each pixel connected to a seed is then investigated. If its value is within the range $[0,T_{\text{max}}]$ it is appended to the seed. Figure 5 gives an example of the final segmentation for one observation. Note that a minimum region size is also defined to remove regions that did not grow large enough.
4 Comparison with Manual Detection

4.1 Shape Description

In order to compare our results with what has been obtained manually we need to describe the filaments in a similar way. The manual digitization involves choosing a few points along the path of the filament and computing the full path by linking the points together. The closest way to represent our filaments is to compute their pruned skeletons. Such skeletons are the results of a thinning process based on the Hit-Or-Miss transform which is a combination of erosion and dilation operators that finds a match in foreground pixels according to a predefined structuring element [8]. The thinning of array $A$ by the structuring element $S$ can be written:

$$\text{Thin}(A, S) = A – \text{HitOrMiss}(A, S)$$

In the thinning process, every matched pixel is set to 0 when this logical difference is computed ($X – Y = X \cap Y^c$). Given a set of structuring elements $S$, this operation is successively applied for each element $S_n$, and the whole process iteratively repeated until there are no more pixels to remove from the foreground regions (see figure 6).

![Original shape (1), full skeleton (2) and pruned skeleton (3)](image)

Fig. 6. Original shape (1), full skeleton (2) and pruned skeleton (3)

The algorithm used for pruning is based on the distance between node points in the skeleton tree and end of branches. From each node point, the closest end points are iteratively eliminated until there are only two end points left in the tree.

4.2 Comparison Results

The procedure for automatic detection of filaments has been applied to the period between April and July 2002, on more than a hundred observations, because the filaments have already been manually detected, and this corresponds to a period where many (1265 manually detected) filaments were visible on the Sun. Examples of the manual detection can be seen on the Bass2000 web site (http://bass2000.obspm.fr).
Automatically and manually detected filaments (with size greater than 5°) have been compared one by one for the whole month of April, giving the following results:

- 89% of automatically detected filaments match the manually detected ones
- 4% of the ‘automatic’ filaments don’t correspond to ‘manual’ ones. In fact, most of them have not been manually detected, but a careful inspection shows that they are real filaments but faint ones most of the time.
- 11% of ‘manual’ filaments haven’t been automatically detected. But the error in the total length of all filaments is only 7% (i.e. non detected filaments were small ones). Again, when carefully looking at these filaments, it appears that these are also faint ones in general.

5 Conclusions

The complete comparison between manual and automatic detection of filaments shows that most of the differences between the two methods come from the subjective interpretation of the original image. The knowledge that a filament is often located at the border of bright regions helps the user to guess the presence of a filament. Moreover the knowledge that a filament was present on a previous day indicates that one may search for even a very thin one at the place it was previously seen.

The technique used here is robust for standard filament detection. It doesn’t seem to be possible to improve it easily, taking into account the remarks above. Nevertheless the next step could be to teach the code how to guess whether a filament could be at some place or not. Neural networks could be well suited to this type of application.

References

1. Ballester P.: Hough transform for robust regression and automated detection, *Astron. Astrophys.* (1994) 286 1011.
2. Bentley R. D.: EGSO – the next step in data analysis, in Proceedings of the 2nd Solar Cycle and Space Weather Euro-Conference, Vico Equense, September, ESA SP-477, ESA Publications, (2001) 603.
3. Collin B. and Nesme-Ribes E.: Pattern recognition applied to Hα spectroheliograms, Solar rotation and meridional circulation, in Proceedings of an ESA workshop on Solar Physics and Astrophysics at interferometric resolution, Paris, 17-19 Feb., (1992) 145.
4. Gao J., Wang H. and Zhou M.: 2002, *Solar Physics*. (2002) 205 93.
5. Gonzalez R. C. and Woods R. E.: Digital Image Processing, Second Edition, Prentice-Hall, Inc., Upper Saddle River, New Jersey, (2002) 613.
6. Russ, J. C.: The image Processing Handbook, Fourth Edition, CRC Press, (2002) 219.
7. Shih F. Y. and Kowalski A. J.: *Solar Physics*. (2003) 218 99.
8. Sonka M., Hlavac V. and Boyle R.: Image Processing Analysis and Machine Vision, Second Edition, PWS Publishing, (1999) 578.
9. Wagstaff K., Rust D.M., LaBonte B.J. and Bernasconi P.N.: Automated Detection and Characterization of Solar Filaments and Sigmoid, Solar Image Recognition Workshop Brussels, October 23-24, 2003.

10. Zharkova V.V, Ipson S.S., Zharkov S.I., Benkhalil A.K., Aboudarham J. and Bentley R.D.: A full-disk image standardisation of the synoptic solar observations at the Meudon Observatory, *Solar Physics*. (2003) 214/1 89.
Adaptation of Shape of Dendritic Spines by Genetic Algorithm

A. Herzog\textsuperscript{1}, V. Spravedlyvyy\textsuperscript{1}, K. Kube\textsuperscript{1}, E. Korkotian\textsuperscript{3}, K. Braun\textsuperscript{2}, and B. Michaelis\textsuperscript{1}

\textsuperscript{1}Institute of Electronics, Signal Processing and Communications, Institute of Biology, Otto-von-Guericke University Magdeburg, P.O.Box 4120, 39016 Magdeburg, Germany
\textsuperscript{2}The Weizmann Institute, Department of Neurobiology, Rehovot 76100, Israel
\texttt{herzog@iesk.et.uni-magdeburg.de}

Abstract. The role of dendritic spines in information processing of a neuron is still not clear. But it is known that they change their shape and size during learning processes. These effects may be important for storing of information (memory). We analyze the influence of shape variations on the electrical signal propagation in a group of dendritic spines by biologically realistic electrical simulation. In order to show the potential of shape changes a genetic algorithm is used to adapt the geometric parameters of the spine group to specific timing of incoming spikes. We can show that such a group of spines can do information processing like coincidence detection just by adjustment of its geometry.

Introduction

The input region of artificial neurons is often simplified by a weighted superposition of incoming signals. But biological neurons have a nonlinear behavior in this input region with a greater variability to modulate and integrate input signals. On many types of neurons input synapses are located on dendritic spines with special geometry (see Fig. 1 left). Although spine existence was mentioned long time ago [1], the discussion about its role in nervous system has been continued so far [2,3,4]. Morphological studies showed that their density and spatial distribution but also their size and shape are variable and change during brain development and during learning processes [5,6]. The geometrical parameters of spines can be estimated by analyzing microscope images [7,8]. But the influence of geometry on the signal transmission is not clear yet due to shortcomings of electrical measuring techniques in these small dimensions.

By biologically realistic numeric simulation including relevant properties of membrane and ion channels [9], we can study the signal transmission inside a group of spines conditioned by its specific geometry to answer the questions: what influence have the observed changes in morphology on the information processing in dendritic tree and is it possible to use geometrical changes to learn a new behavior? Analyzing a
single spine on a dendrite, we show the fitting of electrical properties by changing geometric parameters with simulated annealing (SA) algorithm [10]. But to adapt the signal transmission of a spine group for a specific task, it is difficult to adjust fitting parameters of SA (cooling) because many combinations of geometric parameters with lot of local maxima exist. So we designed a modified genetic algorithm including the established simulated annealing as asexual reproduction.

2 Simulation Model

We use a simplified part of a dendrite with five spines (Fig. 1 right). The geometry of spine is simplified to four basic parameters: length $l_n$ and diameter $d_n$ of neck and length $l_h$ and diameter $d_h$ of head. The simulations are done by a compartment model [12], which is computed numerically. The spine is split into 10 compartments (3 for head, 7 for neck) and the dendrite in 20 compartments. The spine neck is modeled by a number of compartments and not by a single resistor, because we will include effects of its dynamics caused by the distribution of capacities along the spine neck. A special compartment is added to the dendrite to simulate proximal parts and soma and a glutamate synapse is integrated in first compartment of spine head to simulate the synaptic input.

![Fig. 1. Part of a dendrite with spines. Shadow projection from confocal microscope image (left), artificial model for electrical simulations (right)](image)

Electrical properties of membrane and environment are set to standard values ($R_m = 0.333 \ \Omega/m^2; \ Ra = 0.3 \ \Omega/m; \ C_m = 0.1 \ F/m^2, \ E_r = 0.07 \ V$). The concentration of active ionic (HH) channels (only in spine head; neck and dendrite are passive) is set to a subthreshold level. These values are kept constant during the simulation experiments. Investigations with varying values did not show qualitative changes of results as long as concentration of active channels is sub-threshold. Possible biochemical mechanisms of adapting the synapses (LTP, LTD) are not included because only morphological influence is of interest here. For numerical simulations of the compartment model the program GENESIS [12] has been used. Data provision and analysis has been done in C++ and MATLAB. All calculations run in parallel on a Beowulf PC-Cluster using Linux.
3 Geometric Adaptation by Genetic Algorithm

3.1 Setup of Genetic Algorithm and Fitness Definition

Each of the five spines has four geometric parameters, so overall 20 genes are used. But there are some geometric restrictions. The dimensions are limited to minimal and maximal values and the diameter of head must be larger than neck diameter. For optimizing geometric parameters we use 50 individuals including two elitist individuals, which do not change by recombination or mutation and are always the best ones. The fitness function is the result of our simulation experiment. The parents for recombination are chosen by roulette wheel. We evaluate the postsynaptic potential (PSP) of dendrite depending on different timings of input spikes. One of the tests is shown in Fig. 2. Here the spine geometry is to be optimized for coincidence detection. That means that the peak potential of synchronous spikes has to be higher than the peak potential of asynchronous spikes.

![Fig. 2. Result PSP on different incoming timing of spikes but same geometric parameters. Synchronous spikes (right) cause higher peak value than asynchronous (left)](image)

To obtain the fitness of a specific geometry, the simulation is carried out with these two different timed input spikes. The peak values are measured and compared. In other tests we improve the distinction of different spike timings and order. A specific order of spike can be separated from the reverse order or coincidence of two different spines can be detected. Test to analyze statistics of spike trains (more than one spike of a spine) are in preparation.
3.2 Combination of Genetic Algorithm with Simulated Annealing

Analyzing a single spine on a dendrite, we showed that the adaptation of electrical properties by changing geometric parameters by simulated annealing (SA) perform well [10]. SA is used here as fitting algorithm, because it utilizes the known twitching of spines [11], needs only simple feedback signals, but no inside memory and seems biologically realistic in this way. But to handle a group of spines is more complex because of the high number of parameters and its interactions which cause a lot of local maxima in fitness function have to be considered. SA may work theoretically as well, but it is difficult to adjust fitting parameters (cooling) to accomplish both: go carefully into a local maxima and analyze it and on the other hand jump to other local maxima. To overcome this we designed a modified genetic algorithm and include the established simulated annealing as asexual reproduction. The idea is to use the genetic algorithm to check several parameter combinations, recombining and mutating the best spines. But inside local areas in parameter space SA is used for optimization (see Fig. 3). Biologically this can be compared with stronger modifications of the dendritic tree by pruning or formation of spines.

Simulated annealing can be also considered as a genetic algorithm. Each individual performs an asexual reproduction including mutation changing gene position. If the child has a better fitness, it will be the source for the next generation. If not, it will be ignored in most cases. But some times it survives with a low probability, except for elite individuals. In our experiments we combine 10 steps of asexual with one step of sexual reproduction. Similar combinations are found in biologic systems (e.g. jellyfishes). As in biology, we use the benefit, that no information transfer between the individuals during asexual phases takes place. This means that each individual can be calculating parallel on different nodes of the pc-cluster without any communication.

3.3 Changing Gene Position

Each individual consists of five spines here. The spines are functional subunits. They must work together like a small community of (sub-) individuals with a collective
fitness function. The position of an individual spine in the group plays a role because of the time-delay of signal transmission to measure point and of the interaction between the spines. But it seems that a small probability of changing positions (transposons) is favorable to clone a good spine or to push out a bad one. This way we define an additional recombination option during mating and also during the asexual reproduction (SA). Normally each child gene is a combination of corresponding genes of its parents. But in our approach, a spine on position \( j \) can choose the genes not only from corresponding but also from another position \( i \) with a small probability depending on the distance \( |i-j| \) between points.

![Recombination of spines during mating](image)

**Fig. 4.** Recombination of spines during mating. To do recombination a child spine chooses a spine from each parent with a probability depending on distance between spine positions

## 4 Experiments and Results

For coincidence detection test, we use the input shown in Fig. 2. The fitness results from relative difference of measured peak membrane potentials at dendrite in case of synchronous \( V_{\text{synch}} \) and asynchronous \( V_{\text{asynch}} \) stimulation.

Starting with 50 individuals of random geometry in the specified range, fitness increases (see Fig. 5). The steps in upper curve in Fig. 5, show that most improvements of best individual are caused by sexual reproduction. But sometimes asexual reproduction leads also to some enhancement (50-70 generations). The asexual reproduction by simulated annealing increases the average fitness (middle curve) and prepares the individuals for the next recombination in this way. The lower curve (worst individual) shows that there is a range of individuals in one generation. Other tests (e.g. specific order of spikes compared with reverse order) performed with similar results, but can’t be shown here in detail because space limitation.

The calculation time of one generation is approximately one minute on a 3.06 GHz Dual-Xeon with 1 Gbyte memory in non parallel mode. Most part of this time takes the electrical simulation during fitness calculation. Parallel processing of fitness calculation and separation of individuals during asexual reproduction speed up the program nearly linearly with participating processors up to number of individuals.
Adaptation of Shape of Dendritic Spines by Genetic Algorithm

Fig. 5. Change of fitness over generations. Fitness of best individual (upper curve) increase mostly on sexual reproduction (every 10 generation), average (middle curve) is increased also on SA (preparing individuals for next generation), fitness of worst individuals (lower curve) indicates that not all individuals have the same parameters, images show best individuals of selected generations

5 Discussion and Conclusion

A group of dendritic spines is able to separate between different timings of incoming spikes by the value of peak membrane potential. We show the accuracy of discrimination can be optimized by adapting the geometric parameters of spines. A combination of genetic algorithm and simulated annealing is used to find a good parameter combination. Comparing the result geometries with geometry of real spine groups, may help to understand the role of dendritic spines better. Furthermore this experiments of biologic realistic modeling and simulation can help to improve artificial spiking neurons and find new learning algorithms

This work is supported by VolkswagenStiftung (grant I/76 692) and LSA 3431.

References

1. S. Ramón y Cajal: Estructura de los centros nervioso de las aves. Rev. Trim. Hitol. Pat., 1, 1 – 10 (1888)
2. 2 C. Koch, A. Zador: The function of dendritic spines: devices subserving biochemical rather than electrical compartmentalization. J Neurosci. 13(2), 413-422 (1993)
3. 3 M. Segal, P. Andersen: Dendritic spines shaped by synaptic activity. Cur. Op. in Neurobiology, 10, 582 – 586 (2000)
4. 4 I. Segev, W. Rall: Excitable dendrites and spines: earlier theoretical insight elucidate recent direct observation. Trends in Neuroscience, 21(11), 453 – 460 (1998)
5. C. Helmeke, G. Poeggel, K. Braun: Differential emotional experience induced elevated spine densities on basal dendrites of pyramidal neurons in the anterior cingulate cortex of octodon degus. Neuroscience, 104 (4), 927 – 931 (2001)
6. L. Tarelo-Acuña, E. Olvera Cortés, I. Gonzáles Burgos: Prenatal und Postnatal exposure to athanol induces changes in the shape of the dendritic spine from hippoampal CA1 pyramidal neurons of the rat. Neuroscience Letters 286, 13 – 16 (2000)
7. A. Herzog: Rekonstruktion dendritischer Spines aus dreidimensionalen Mikroskopbildern unter Verwendung geometrischer Modelle. Sharker (2002)
8. A. Herzog, G. Krell, B. Michaelis, K. Braun, J. Wang, W. Zuschratter: Restoration of Three-Dimensional Quasi-Binary Images from Confocal Microscopy and its Application to Dendritic Trees. SPIE 2984, pp. 146-157, (1997)
9. C. Koch: Biophysics of Computation: information processing in single neurons. Oxford University Press (1999)
10. A. Herzog, V. Spravedlyvyy, K. Kube, R. Schnabel, K. Braun, B. Michaelis: Learning by geometrical shape changes of dendritic spines. ESANN 385-390 (2004)
11. F. Crick: Do dendritic spines twitch? Trends in Neuroscience, 44-46 (1982)
12. 12 M. Brown, D. Beeman: The book of Genesis: exploring realistic neural models with the GEneral NEural SImulation System. Springer (1999)
Detection of Dynamical Transitions in Biomedical Signals Using Nonlinear Methods

Patrick E. McSharry

1 Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
2 Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK
3 Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK
patrick@mcsharry.net
www.mcsharry.net

Abstract. The ability to detect the existence of nonlinear dynamics may facilitate medical diagnostics for identifying, monitoring and predicting transitions from health to sickness. Detection of such transitions depends on the quality of the available biomedical signals and the relevance of the nonlinear statistics employed. The dangers of using irrelevant statistics are discussed. A simple intuitive nonlinear statistic, which evaluates changes in the distribution of points in state space is shown to be capable of detecting both linear and nonlinear dynamical transitions. This new technique, known as multi-dimensional probability evolution (MDPE), is illustrated using a synthetic signal obtained by mixing stochastic and a chaotic processes. Its utility in detecting transitions in biomedical data is demonstrated using a database of electrocardiograms collected from subjects who experienced partial epileptic seizures.

1 Introduction

There are many statistics that can be employed for detecting dynamical transitions from observed signals. These range from linear statistics such as the mean, variance, power spectrum, to nonlinear statistics such as the correlation dimension and maximum Lyapunov exponent. Nonlinear statistics rely on a state space reconstruction and are likely to vary when the distribution of points in this state space changes. In practice it is difficult to obtain accurate estimates for these statistics from noisy data sets generated by non-stationary processes. Rather than calculating these statistics with the belief that they are necessary to see nonlinear effects, changes in the distribution of points in the state space are measured.

A new technique called Multi-Dimensional Probability Evolution (MDPE) [1] which computes the Probability Density Function (PDF) within a multi-dimensional state space for each window of the time-varying signal is proposed. By defining a learning period it is then possible to calculate the probability that a transition (dynamics different to those in the learning period) has occurred.
in subsequent windows of the signal. MDPE is capable of capturing both linear and nonlinear changes in the underlying dynamics.

2 Methodology

A delay vector reconstruction of the signal \( s_i \) is defined by

\[
x_i = [s_{i-(m-1)\tau}, \ldots, s_{i-\tau}, s_i],
\]

where \( m \) is the reconstruction dimension and \( \tau \) is the time delay. A reference set \( \mathcal{A} \), representing normal activity, is constructed from the state vectors recorded during the learning period. Choosing \( N_c \) centres \( \xi_i \) \((i = 1, \ldots, N_c)\) in the state space yields a partition such that \( x \in \mathcal{A} \) is a member of partition \( B_i \) if \( ||x - \xi_i|| < \min_{j \neq i} ||x - \xi_j|| \). Counting the number of points \( n_i^0 \) in each of the \( N_c \) partitions yields a discrete distribution for the reference set \( \mathcal{A} \). Similarly, for any given window of the recording, it is possible to calculate its distribution \( n_i \) in the state space. A \( \chi^2 \)-test \[2\] may be used to compare the distribution \( n_i \) with that of the reference set \( n_i^0 \). Suppose that the total number of points in the reference set and the window are \( N^0 \) and \( N \) respectively. The \( \chi^2 \)-statistic is

\[
\chi^2 = \sum_{i=1}^{N} \left[ \frac{rn_i^0 - (1/r)n_i}{n_i^0 + n_i} \right]^2,
\]

where \( r = (N/N^0)^{1/2} \). Note that \( \chi^2 \) is zero if both distributions have equal numbers in each partition. In the case of both time series having different numbers of points, the number of degrees of freedom is \( \nu = N_c \). The \( \chi^2 \) probability distribution \( p = Q(\chi^2|\nu) \), an incomplete gamma function, gives the probability of observing values greater than \( \chi^2 \) under the null hypothesis that the data sets are drawn from the same distribution \[2\]. A small value of \( p \) indicates that there is a significant difference between the two distributions. \( \gamma = -\log_{10} p \) was calculated in order to avoid problems with computer precision. The reconstruction parameter values, \( m = 2 \), \( \tau = 1 \) and \( N_c = 100 \) were used in the following analyses.

3 Results

The utility of the MDPE technique is first illustrated using a synthetic signal obtained by mixing a stochastic process with a deterministic chaotic process. This analysis demonstrates why the MDPE is capable of detecting changes that may be invisible to classical linear measures. Furthermore, the synthetic signal enables an analysis of the ability of MDPE to distinguish between the stochastic and chaotic signals when faced with additive measurement noise and datasets of different lengths. Finally, MDPE is used to explore a dataset of ECG recordings from subjects who experienced partial epileptic seizures.
3.1 A Synthetic Signal

A synthetic signal using a single parameter to control the amount of nonlinearity in the underlying dynamics at any given instant in time is constructed. Consider a random process with linear temporal correlations such as the Auto-Regressive process of order one AR(1)

\[ x_{i+1} = \alpha x_i + \epsilon_i, \]

where \(-1 < \alpha < 1\) and \(\epsilon_i \sim N(0,1)\). The standard deviation of this process is \(\sigma_x = 1/(1-\alpha^2)^{1/2}\) and its auto-correlation \(\rho_k\) is given by \(\rho_k = \alpha^{|k|}\) where \(k\) is the time delay. The nonlinear deterministic system known as the skew-tent map given by

\[ y_{i+1} = \begin{cases} ay_i, & 0 \leq y_i \leq a \\ \frac{1-a}{1-a} y_i - 1, & a \leq y_i \leq 1 \end{cases}, \]

is a non-invertible transformation of the unit interval into itself with the parameter \(a\) chosen to satisfy \(0 < a < 1\) and its invariant measure is uniform on the unit interval. This dynamical system is chaotic with Lyapunov exponent \(\Lambda = -a \log_2(a) - (1-a) \log_2(1-a)\). If the parameter values of these two systems are chosen such that \(\alpha = 2a - 1\), then both systems will have identical power spectra. A measurement function, \(z_i = h(y_i)\), was used to transform the output of the skew-tent map, \(y_i\), so that \(z_i \sim N(0,\sigma_x)\) as in the case of the AR(1) process \(x_i\). The required measurement function is \(z_i = \sqrt{2/\sigma_x} \Phi[2(y_i - 1/2)]\) where \(\Phi\) is the inverse error function [2]. Figure 1 contrasts the statistical properties of \(x_i\) and \(z_i\) for \(a = 0.95\). While these two time series appear different, their PDFs and auto-correlation functions are identical in the limit of an infinite amount of data [1]. The difference in the underlying dynamical equations may be seen from their return maps (Fig. 1d & Fig. 1h), which may be detected by the MDPE technique.

A synthetic signal \(s_i\) is defined as the linear combination of the random process \(x_i\) and the nonlinear deterministic process \(z_i\):

\[ s_i = \sqrt{\beta_i} z_i + \sqrt{1-\beta_i} x_i, \]

where \(\beta_i\) controls the amount of nonlinear deterministic structure in \(s_i\).

The dynamical changes induced by blending in the nonlinear deterministic process in the synthetic signal are invisible to the linear statistics (Fig. 2). Figures 2a and 2b show the synthetic signal \(s_i\) and the associated control parameter \(\beta_i\). A moving window analysis of the mean, variance, power spectrum, and auto-correlation function is shown in Figures 2c to 2f; none of these linear statistics are capable of detecting the changes due to the nonlinear correlations. Low probabilities (large \(\gamma\)) indicate excursions into new or rarely visited regions of state space, reflecting abnormal behaviour with respect to the learning data set (first 100 seconds) between -200 and 200 seconds (Fig. 2g). \(\gamma\) clearly reveals the dynamical changes introduced by the nonlinear deterministic process when \(\beta_i \neq 0\) (Fig. 2b & 2g).
Fig. 1. Comparison of AR(1) process $x_i$ with nonlinear deterministic process $z_i$ for $a = 0.95$: (a) $x_i$, (b) PDF of $x_i$, (c) auto-correlation function of $x_i$, (d) return map of $x_i$, (e) $z_i$, (f) PDF of $z_i$, (g) auto-correlation function of $z_i$ and (h) return map of $z_i$

Fig. 2. Synthetic signal (a) $s_i$, (b) control parameter $\beta$, non-overlapping 20 second window (c) mean $\mu$, (d) variance $\sigma$, (e) power spectral density (arbitrary units), (f) auto-correlation function and (g) MDPE statistic $\gamma$
In a clinical setting, the observed biomedical signal will contain measurement errors and the length of the datasets available for analysis may often be small if stationary segments are required. For this reason it is interesting to examine the robustness of the MDPE technique in distinguishing between a stochastic and a chaotic process. Consider two data sets of length $N$ from the AR(1) process and the skew-tent map with observational uncertainty given by adding normally distributed measurement errors with standard deviation $\sigma_{\text{noise}}$. As $\sigma_{\text{noise}}$ increases the ability of MDPE to distinguish between the two data sets decreases (Fig. 3). In contrast, as the length of these data sets, $N$, increases, the performance of MDPE improves. These results highlight the fact that while many biomedical signals may have underlying nonlinear dynamics, the quality of the data and the length of available stationary sections may prohibit the detection or measurement of this nonlinearity.

![Fig. 3. Ability of MDPE to distinguish between data from the chaotic skew-tent map and the stochastic AR(1) process for different amounts of additive measurement errors with standard deviation $\sigma_{\text{noise}}$. The lengths of the datasets are $N = 1000$ (circle), 5000 (square), and 10000 (triangle). The error bars reflect one standard deviation above and below the mean of 40 realisations](image)

It may be unwise to use nonlinear measures to quantify the state of health of a subject from a biomedical signal. Consider, for example, the use of the Lyapunov exponent to detect epileptic seizures from electroencephalogram (EEG) recordings. The Lyapunov exponent summarises an entire section of EEG using one single number and therefore many different datasets will be mapped onto a single value. The disadvantage of this many-to-one mapping is illustrated using the skew-tent map for which the Lyapunov exponent is known analytically for any parameter value (Fig. 4). Consider the time series (Fig. 4b) generated by the skew-tent map (4) with a drifting parameter value, $a$, shown in Fig. 4. MDPE is applied to the entire time series using the first 500 points for learning the reference dynamics. The values of $\gamma$ (Fig. 4d) obtained clearly reflect the changes in the dynamics given by $a$ (Fig. 4e). In contrast, the Lyapunov exponent, $\Lambda$, represents many different parameter values by the same value of...
Fig. 4. MDPE analysis of data from the skew-tent map with drifting parameter value: (a) five different state spaces of the skew-tent map for $a = 0.1, 0.25, 0.5, 0.75, 0.9$ respectively (b) the time series $x_i$, (c) the parameter value $a$ versus time $i$, (d) the MDPE $\gamma$ and (e) the Lyapunov exponent $\Lambda$.

Fig. 5. MDPE analysis of the partial epilepsy database showing the beat-to-beat heart rate (grey line), the values of $\gamma$ obtained (black line) before, during and after the seizures. The seizures are indicated by shaded vertical bars.
\( \Lambda \), resulting from the symmetry, \( \Lambda(a) = \Lambda(1-a) \). For example the structure of the skew-tent map for \( a = 0.1 \) and \( a = 0.9 \) (Fig. 4a) both have \( \Lambda = 0.4690 \). While these two configurations of the skew-tent map have the same properties in terms of the average rate of divergence of nearby trajectories, there is no reason to believe that biomedical signals with similar values of \( \Lambda \) should reflect similar states of health.

3.2 Post-ictal Heart Rate Oscillations in Partial Epilepsy

The database of single lead ECG recordings demonstrates post-ictal heart rate oscillations in a heterogeneous group of patients with partial epilepsy [4]. These oscillations are characterised by the appearance of transient but prominent low-frequency heart rate oscillations (0.01 - 0.1 Hz) immediately following five of 11 seizures recorded in five patients.

The MDPE technique was applied to the heart rate time series in order to visualise these transitions (Fig. 5). The state space was reconstructed using \( m = 2 \) and \( d = 1 \) and \( N_c = 30 \) centres were used. The first ten minutes of each time series was used as a reference data set and was taken as representing non-seizure activity. A non-overlapping window of 100 seconds was then used to compute the value of \( \gamma \) throughout the entire time series. MDPE detects all the seizures apart from the two in the sixth recording (Fig. 5).

4 Conclusion

A technique, known as MDPE, for detecting both linear and nonlinear dynamical changes has been presented. Its performance was illustrated using a synthetic signal obtained by mixing stochastic and chaotic processes. MDPE was successful in identifying the quantity of the chaotic process in the signal. An analysis of the robustness of MDPE to both measurement noise and the length of the data sets was also provided. This emphasises that the detection of nonlinearity depends on the quality of the recorded signal and the amount of stationary data available. The MDPE analysis of heart rate obtained from ECG recordings of subjects with partial epilepsy demonstrated that it was possible to detect most of the seizures. The analysis of biomedical data with nonlinear measures should be accompanied with relevant significance tests [5]. This will be the focus of future investigations.

References

1. McSharry, P.E., He, T., Smith, L.A., Tarassenko, L.: Linear and nonlinear methods for automatic seizure detection in scalp electroencephalogram recordings. Medical & Biological Engineering & Computing 40 (2002) 447–461
2. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C. 2nd edn. CUP, Cambridge (1992)
3. Iasemidis, L.D., Sackellares, J.C.: Chaos theory and epilepsy. The Neuroscientist 2 (1996) 118–126
4. : (http://www.physionet.org)
5. McSharry, P.E., Smith, L.A., Tarassenko, L.: Prediction of epileptic seizures: are nonlinear methods relevant? Nature Medicine 9 (2003) 241–242
On Retrieval of Lost Functions for Feedforward Neural Networks Using Re-Learning

Naotake Kamiura¹, Teijiro Isokawa¹, Kazuharu Yamato², and Nobuyuki Matsui¹

¹Graduate School of Engineering, Himeji Institute of Technology, University of Hyogo, 2167 Shosha, Himeji, 671-2201, Japan
{kamiura, isokawa, matsui}@eng.u-hyogo.ac.jp
²Dept. of Economics & Information Science, Hyogo University, 2301 Shinzaike Hiraoka, Kakogawa, 675-0101, Japan

Abstract. This paper proposes the re-learning for feedforward neural networks where weight faults would occur. The sequences of target outputs are encoded by means of single-parity-check codes so that a single-bit error caused by the faults can be on-line detected at the output layer. The re-learning is made every time a network produces the error, and its lost function is retrieved. The proposed scheme can easily achieve high MTTF (Mean Time To Failure).

1 Introduction

The hardware implementation of feedforward neural networks (hereafter NN’s) has been actively pursued[1], [2]. Since it is inevitable that faults occur in the implemented NN, fault-tolerant strategies for the NN have been devised. In [3]-[8], the standard backpropagation algorithm (hereafter STBP) is modified to increase the possibility that normal outputs emanate from an NN with faults. However, once some faults cause errors (i.e., false outputs) while such an NN is at work, they are neglected and a function of the NN is lost. Though re-learning schemes[9] are available for retrieving the lost function, they have difficulties either in on-line detecting a fault or in adjusting the appropriate interval between invocations of the re-learning.

This paper proposes the re-learning made synchronously with on-line detecting an error, to retrieve the lost function. A sequence of target outputs is encoded by means of the single-parity-check code, and hence a parity check finds an error at the output layer. Since more than one error per sequence with actual outputs invalidates the check, fault-tolerant approaches in [7] and [8] are incorporated into the learning. Experimental results on a character recognition problem show that the proposed re-learning is superior to others in length exclusive of the total re-learning time.

2 Preliminaries

This paper focuses on an NN with the input layer, single hidden layer and output layer. The output \( u_i \) of the \( i \)-th neuron in some layer is expressed by \( u_i = f(a x_i) = (1 + \exp(ax_i))^{-1} \), where \( a (a>0) \) is a gradient of a sigmoid activation function \( f \), and \( x_i \) is the potential of the \( i \)-th neuron. \( x_i \) equals the well-known weighted sum of the
outputs of all neurons in the preceding layer. Let $w_{ij}$ denote a weight between the $j$-th neuron in the preceding layer and the $i$-th neuron. Assuming that STBP is adopted, the learning rule for $w_{ij}$ is as follows.

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}.$$  

In this paper, $\eta = 0.15$, and the momentum factor is set to 0.04 for the momentum term. In the classification task, the sigmoid activation function in the output layer is replaced with the following threshold activation function after the learning: if $x_i \geq 0$, $o_i = 1$; otherwise, $o_i = 0$ where $o_i$ is the actual output of the $i$-th neuron.

Figure 1 illustrates the fault-synchronous re-learning (re-learning for short) and periodic re-learning (re-learning for short) proposed in [9]. Even if stuck-at-0 faults of links occur in an NN being at work and they disrupt a function given to the NN by means of the initial learning, the above re-learning updates the weights where non-faulty links are connected, and can often retrieve the lost function. The re-learning is invoked every time a fault occurs, and the re-learning is invoked at regular intervals, $\Delta t$'s. Then, STBP is adopted to update the weights. Although the re-learning achieves high MTTF easily, arbitrary faults must be detected while the NN is at work. There is no discussion about such on-line detection in [9]. The re-learning achieves as high MTTF as re-learning, provided that $\Delta t$ is very short. The very short $\Delta t$ however results in a problematic situation that the re-learning is invoked perpetually.

3 Error-Synchronous Re-learning

In the following, weight faults are discussed as fault models. When the value of some weight is stuck at $F$ belonging to an interval denoted by $[L_F, U_F]$, it is referred to as a stuck-at-$F$ fault of a weight (s-a-$F$ fault for short). It covers stuck-at-0 faults of links[3], [5]-[7]. In addition, while an NN is at work after the learning, it is said that an error appears when a false actual output emanates from some neuron in the output layer. This “error” does not mean a “square error” related to the learning.
3.1 Re-learning Reacting to Error Detection

Faults sometimes result in no errors in the output layer of an NN being at work. The re-learning requires on-line detecting such faults, and this detection is very hard to execute by observing only the actual outputs. This paper, therefore, proposes the following error-synchronous re-learning (re-learning for short).

Let us suppose that an \( N_I - N_H - N_O \) NN, which has \( N_I \), \( N_H \), and \( N_O \) neurons in the input, hidden and output layers respectively, is applied to classification tasks such as character recognition. A sequence of target outputs is referred to as a target sequence. The target sequences are encoded by means of single-parity-check codes, and hence each of them is a binary \( N_O \)-bit codeword. The single-bit errors are then easily found by parity checks while the NN is at work. If the Hamming weight of an actual sequence (i.e., a sequence of actual outputs) is an odd number for some input pattern in spite of setting that of any target sequence to an even number, or vice versa, a single-bit error arises. Accordingly, the re-learning is immediately invoked. Figure 2 illustrates the behavior of an NN based on this re-learning. Let \( t_i^p \) (or \( o_i^p \)) denote the target (or actual) output of the \( i \)-th neuron in the output layer, when the \( p \)-th learning pattern is fed. The learning-completion condition is as follows:

\[
\max_{p,i} (t_i^p - o_i^p)/2 < 0.005, \tag{2}
\]

provided that maximum number of epochs equals 3000. If the number of epochs exceeds 3000, the behavior reaches “End” and the NN is regarded as being down.

![Fig. 2. Behaviour of NN based on re-learning](image)
3.2 Approaches Allowing On-Line Parity Check

The case where two or more errors appear per actual sequence is unfavorable for the re-learning\(^*\). Processes A and B shown in Fig. 3 make it possible to reduce the probability of such case arising. The reasons why they are employed are detailed in [7]. Let \( E_0 \) and \( E_{-1} \) denote square errors (i.e., \( E = \sum p_i \sum t_i^p - o_i^p \cdot l^2 / 2 \)) calculated at the present epoch and the preceding epoch, respectively. Two sigmoid activation functions \( f_{a1}(x_i) \) with a gradient \( a_1 = 1.0 \) and \( f_{a2}(x_i) \) with \( a_2 = 0.5 \) are prepared for the output layer. Process A is made as follows, assuming that \( E_c = 5.0 \times 10^{-4} \), and the gradient \( a \) of the sigmoid activation function \( f \) in the hidden layer is set to 0.5.

\(<\text{Process A}>\)

\[\text{[Step 1] Calculate } E_{-1} - E_0.\]
\[\text{[Step 2] If } E_{-1} - E_0 > E_c, \text{ employ } f_{a2}(x_i); \text{ otherwise, employ } f_{a1}(x_i). \text{ Continue the learning. If the learning converges, go to Process B; otherwise, go to Step 1.}\]

In Process B, the gradient \( a \) of \( f \) in the hidden layer is manipulated as follows.

\(<\text{Process B}>\)

\[\text{[Step 1] Add } \Delta a \text{ to } a, \text{ where } \Delta a = 0.1 \text{ in this paper.}\]
\[\text{[Step 2] Apply any learning pattern to NN. If there exists at least one pattern not to be recognized, subtract } \Delta a \text{ from } a \text{ and stop. If not so, go to Step 3.}\]
\[\text{[Step 3] If a decrease in minimum among absolute values of potentials is observed in the output layer, subtract } \Delta a \text{ from } a \text{ and stop. If not so, go to Step 1.}\]

In addition to the above, modifying the potential calculation in [8] is made for any neuron as another fault-tolerant approach.

\[\text{Fig. 3. Manipulation of activation function incorporated into learning}\]
4 Experimental Results

The re-learning, re-learning and re-leaning implemented in language C were applied to the character recognition with 16 characters from A to P, each of which is depicted on a 7×7-binary-image plane, as learning patterns. Arbitrary black (1) pixels and white (0) pixels are fed to a 49-NH-NO NN, where NH=20, 30. The initial weights randomly take values uniformly distributed inside [-1, 1]. For the re-learning and re-learning, a gradient of the sigmoid function is set to 1.0.

4.1 Occurrence of Errors in Output Layer

Errors emanating from the NN being at work are examined. Since 16 characters are employed, the parity check is available in the cases of NO=16 and NO=5. If NO=16 (or NO=5), the target sequences are encode so that the Hamming weight of each sequence will be one (or an even number). After approaches in Subsect.3.2 are complete, pf percent of all of the weights are randomly chosen, and s-a-F faults are injected into the chosen weights. For this recognition problem, subject to applying STBP, the values of weights in the 49-NH-NO NN have belonged to the interval [-3.5, 3.5]. So Fs are chosen randomly from this interval for each injected fault. Under pf=5 (or 10), 1000 fault distributions are generated. After applying any leaning pattern to the NN with one of the distributions, the number of errors is checked for each actual sequence. The percentage of the number of fault distributions that result in a single-bit error at most per actual sequence, compared to the total number of fault distributions, is assessed as a metric. Let Rs denote this metric. The above evaluation is made for 50 random start positions (initial weights). Table 1 shows the averaged results. If NO=16, we always have Rs>90, and it is safely considered that approaches in Subsect.3.2 allow us to reduce the probability of more than one error appearing per actual sequence. In the following, 49-NH-16 NN’s are therefore discussed.

4.2 Experimental Evaluations on MTTF’s

Let ns denote the maximum number of permissible faults. In other words, the (ns+1)-th fault makes it impossible to retrieve the lost function of an NN. The beginning point means the point in time when the NN begins normally working, after the initial learning converges. For the re-learning, the MTTF is as follows.

[Definition 1] The MTTF achieved by means of the re-learning is the length between the beginning point and the point in time when the (ns+1)-th fault occurs.

It is shown that a single NN is superior in MTTF to the triplication (i.e., combined three NN’s)[9], subject to applying the re-learning. The re-learning is originally proposed for such triplication. For the single NN, however, the re-learning probably achieves as high MTTF as the re-learning. The MTTF is then as follows.

[Definition 2] The MTTF equals np Δt, if the np-th invocation of the re-learning can retrieve the lost function of an NN, whereas the next invocation fails in retrieving it. The definitions similar to the above are also shown in [9].

For the proposed re-learning, the MTTF is defined as follows.
[Definition 3] It is assumed that the lost function of an NN is not retrieved, by means of the re-learning, immediately after some error is detected. The MTTF is the length between the beginning point and the point in time when the above error is detected.

### Table 1. Probability of atmost a single-bit error appearing per actual sequence

|               | \(p=5\) (%) | \(p=10\) (%) |
|---------------|--------------|--------------|
| \(N_h=20\)   | 96.6         | 90.4         |
| \(N_h=30\)   | 99.7         | 94.6         |
| \(N_o=16\)   | 47.4         | 42.3         |
| \(N_o=5\)    | 40.9         | 38.1         |

### Table 2. Evaluation in terms of MTTF and frequency of invoking re-learning

|               | \(\Gamma\)'s | \(\rho\)'s |
|---------------|--------------|------------|
| \(N_h=20\)   | \(N_h=30\)   | \(N_o=20\) | \(N_o=30\) |
| The re-learning\(^{\prime}\) | 38.5         | 0.134      | 0.018      |
| The re-learning\(^{p}\)   | 41.4         | 52.4       | 0.954      | 0.954      |

If the learning according to STBP is initially made just once and no re-learning is made, an MTTF is defined as length between the beginning point and the point in time when the first error appears in the output layer. Let MTTF\(_0\) denote such MTTF.

An MTTF and a frequency of invoking the re-learning per fault are experimentally measured. The \(s\)-\(a\)-\(F\) faults are injected into a 49-\(N_h\)-16 NN one after another from the beginning point. \(F\)'s and points in time when injecting the faults are determined randomly. Learning patterns are then fed to the NN randomly as input patterns, and each re-learning scheme is invoked if necessary. As metrics, \(\Gamma\) and \(\rho\) equal to MTTF/MTTF\(_0\) and the frequency of invoking each of the re-learning schemes divided by \(n_1\) are employed, respectively. Each of the resultant \(\Gamma\)'s and \(\rho\)'s in Table 2 is the average of 500 runs. Though the re-learning\(^{\prime}\) is very hard to actually begin without being aware of points in time when faults occur, it is safely invoked every time a fault is injected in this experiment. MTTF's achieved by means of the re-learning\(^{\prime}\) are acquired first. Then, \(\Delta t\) is adjusted so that the re-learning\(^{p}\) can achieve as high MTTF as the re-learning\(^{\prime}\). So \(\Gamma\)'s concerned with the re-learning\(^{p}\) approximately equal \(\Gamma\)'s concerned with the re-learning\(^{\prime}\) in Table 2. The re-learning\(^{\prime}\) is approximately equal to the re-learning\(^{p}\) in \(\Gamma\)'s, whereas \(\rho\)'s concerned with the former are far smaller than those concerned with the latter. This is because the re-learning\(^{p}\) requires excessively short \(\Delta t\) to achieve such high \(\Gamma\)'s. Since faults are injected in a same manner and approximately equal \(\Gamma\)'s are acquired for the schemes in Table 2, it is safely said that \(n_1\) takes an approximately equal value for any scheme. So the re-learning\(^{\prime}\) is not invoked so perpetually as the re-learning\(^{p}\). For 49-\(N_h\)-16 NN's (\(N_h=20, 30\)), adopting approaches in Subsect.3.2 requires about 1.7 times as long learning time as STBP[7]. The re-learning\(^{\prime}\) is therefore superior to the previous schemes in availability of the NN between the beginning point and the point in time when the NN is down.
5 Conclusions

This paper proposed the re-learning for NN’s where weight faults occur. A single-bit error is detected by examining the Hamming weight of the actual sequence while an NN is at work. Some fault-tolerant approaches are adopted to make the parity check possible at the output layer. Once an error is found, the re-learning is made to retrieve the lost function of the NN. Experimental results show that the re-learning is superior to the previously proposed re-learning in hours of normal operation.

Determining the number of neurons in the output layer implies determining the number of bits to be prepared to encode each target sequence. A method for determining appropriately the number of such neurons remains to be developed.

References

1. Kosaka, H., Shibata, T., Ishii, H., Ohmi, T.: An Excellent Weight-Updating-Linearity EEPROM Synapse Memory Cell for Self-learning Neuron-MOS Neural Networks. IEEE Trans. on Electron Devices, vol. 42 no. 1 (1995) 135-143
2. Murray, A. F., Smith, A. V. W.: Asynchronous VLSI Neural Networks Using Pulse Stream Arithmetic. IEEE J. Solid-State Circuits, vol. 23 (1998) 688-697
3. Emmerson, M. D., Damper, R. I.: Determining and Improving the Fault Tolerance of Multilayer Perceptions in a Pattern-Recognition Application. IEEE Trans. on Neural Networks, vol. 4 no. 5 (1993) 788-793
4. Murray, A. F., Edwards, P. J.: Enhanced MLP Performance and Fault Tolerance Resulting from Synaptic Weight Noise during Training. IEEE Trans. on Neural Networks, vol. 5 no. 5 (1994) 792-802
5. Ito, T., Takanami, I.: On Fault Injection Approaches for Fault Tolerance of Feedforward Neural Networks. Proc. of the 6th Asian Test Symposium, (1997) 88-93
6. Hammadi, N. C., Ito, H.: A Learning Algorithm for Fault Tolerant Feedforward Neural Networks. IEICE Trans. Inf. & Syst. vol. E80-D no. 1 (1997) 21-26
7. Kamiura, N., Taniguchi, Y., Hata, Y., Matsui, N.: A Learning Algorithm with Activation Function Manipulation for Fault Tolerant Neural Networks. IEICE Trans. Inf. & Syst. vol. E84-D no. 7 (2001) 899-905
8. Kamiura, N., Taniguchi, Y., Isokawa, T., Matsui, N.: An Improvement in Weight-Fault Tolerance of Feedforward Neural Networks. Proc. of The 10th Asian Test Symposium, (2001) 359-364
9. Tohma, Y., Abe, M.: Improvement of MTTF of Feedforward Neural Networks by Applying Re-Learning. (in Japanese) IEICE Trans. D-I, vol. J82-D-I no. 12 (1999) 1379-1386
Analyzing the Temporal Sequences for Text Categorization

Xiao Luo and A. Nur Zincir-Heywood

Faculty of Computer Science, Dalhousie University
6050 University Avenue, Halifax, NS, Canada B3H 1W5
{luo, zincir}@cs.dal.ca

Abstract. This paper describes a text categorization approach that is based on a combination of a newly designed text representation with a kNN classifier. The new text document representation explored here is based an unsupervised learning mechanism – a hierarchical structure of Self-Organizing Feature Maps. Through this architecture, a document can be encoded to a sequence of neurons and the corresponding distances to the neurons, while the temporal sequences of words as well as their frequencies are kept. Combining this representation with the power of kNN classifier achieved a good performance (Micro average F1-measure 0.855) on the experimental data set. It shows that this architecture can capture the characteristic temporal sequences of documents/categories which can be used for various text categorization and clustering tasks.

1 Introduction

Text categorization is one of the significant tasks of content-based document management. Research has been performed in this area since early ‘60s; but it became a subfield of the information system discipline [8] in the early ‘90s. The most popular technique for representing a text document is Vector Space Model (VSM). The basic VSM was introduced in 1975 by Salton et al [7]. In this model, a document is represented by a vector. The number of dimensions of the vector is the number of different words in the corpus. Each entry of the vector is indexed by a specific individual word, and the components of the vector are formed by a given weight of the term. However, many researches show that, in a linguistic sense, individual words could not be expressed as a textual unit because they have a larger degree of ambiguity than phrases [10]. Nevertheless, attempts to introduce more sophisticated text representation methods are not ceasing. These include selected n-grams representation [3], Natural Language Processing [1], [5], Bag-Of-Words [4], [6], [10]. However, with the increasing size of the document corpus, each document usually includes only a small fraction of it. Either n-gram or VSM faces statistical sparseness and high dimensionality problems. Moreover, neither of the representations above considers the significant sequences of words or phrases in the documents. Word sequences or position information is very important to a document when the document is fairly short and the words in each of the documents are very similar.
In this work, our objective is to explore a new way of representing a text document for text categorization by keeping information regarding the temporal sequences of words, as well as their frequencies. The new way of representation is based on hierarchical Self Organizing Feature Maps (SOMs) architecture, which was employed to encode those pertinent features (character probabilities) of a document. Then the encoded information can be used to measure the similarity between the characteristics of any given document. The machine learning categorization algorithm k-Nearest Neighbor (kNN) was employed for the stage of categorization. The results show that this encoding system can capture the characteristic temporal sequences of documents and categories. The sequence information can be utilized for document categorization. The results turned out good on the returned Micro F1-measure (0.855).

This encoding mechanism has several advantages. First, this representation naturally solves the high dimensionality and statistic sparse problem, which occur with the conventional representation for the high volume corpus. Second, it implicitly considers word correlations and position information. Finally, this encoding mechanism can encode both textual and non-textual data. It can also be utilized in analyzing other data where sequence information is significant.

The rest of the paper is organized as follows. Section 2 presents the hierarchical SOMs encoding system. The document representation and the categorization method are described in section 3 and 4. Section 5 gives the experiments performed and the results. Finally, conclusions are drawn and future work is discussed in section 6.

2 Hierarchical SOMs Encoding System

The core of our approach is to automate the identification of typical category characteristics by analyzing the temporal sequence information of the documents in the corpus. In this session, a three-level hierarchical SOM architecture for the process of encoding documents is described. Each of the three levels of the SOM hierarchy is employed to discover the patterns of characters, words, and word co-occurrences.

Indeed, pre-processing of data, which is employed before the encoding process, is shown in Fig. 1.

After preprocessing, SOMs may now be used to identify a suitable character encoding, then word encoding, and finally, word co-occurrence encoding. The hierarchical nature of the architecture is shown in Fig. 2.

1) Input for the First-Level SOMs: In order to train an SOM to recognize patterns in characters, the document data must be formatted in such a way as to distinguish characters and highlight the relationships between them. Characters can easily be represented by their ASCII representations. However, for simplicity, we enumerated them by the numbers 1 to 26, i.e. no differentiation between upper and lower case. The relationships between characters are represented by a character's position, or time index, in a word. For example, in the word “news”: “n” appears at time index 1, “e” appears at time index 2, “w” appears at time index 3, etc. It should be noted that it is
important to repeat these words as many times as they occur in the documents. The overall pre-processing process for the first-level SOM is therefore:

- Convert the word's characters to numerical representations between 1 and 26.
- Give the time index to the characters in a word. It is the actual time index plus 2, expect the first character in the word.

The indices of the characters are altered in this way so that when the list is input to an SOM, both data features (enumerated characters and indices) are spread out over a close range. The assumption at this level is that the SOM forms a code-book for the patterns in characters that occur in a specific document category.

2) Input for the Second-Level SOMs: When a character and its index are run through a trained first-level SOM, the closest neurons (in the Euclidian sense), or Best Matching Units (BMUs), are used to represent the input space. A two-step process is used to create a vector for each word, \( k \), which is input to the first-level SOM of each document:

- Form a vector of size equal to the number of neurons \( r \) in the first-level SOM, where each dimension corresponds to a neuron on the SOM, and is initialized to 0.
- For each character of word \( k \),
  - Observe which neurons \( n_1, n_2, \ldots, n_r \) are affected the most.
  - Increase entries in the vector corresponding to the 3 most affected BMUs by \( 1/j \), \( 1 \leq j \leq 3 \).

Hence, each vector represents a word through the sum of its characters. The result given by the second-level SOM is clusters of words on the second-level SOM.

3) Input for the Third-Level SOMs: In the context of this architecture, word co-occurrence is simply a group of consecutive words in a document. The consecutive words are from a single document with a sliding window of size three. The input space of the third-level SOMs is formed in a similar manner to that in the second-level, except that each word in the word co-occurrences is encoded to the indexes of the 3 most affected BMUs resulting from word vectors passed through the second-level SOMs. The result given by the hierarchical three level SOMs is: clusters of word co-occurrences on the third level SOM.
The sizes of the maps shown in Table 1 are chosen empirically according to the observed weight changes of neurons on the SOMs. Hence, we considered the balance between the computational cost and the weight change in choosing the size of a map.

| Level  | Size  |
|--------|-------|
| Level-1| 7 by 13 |
| Level-2| 8 by 8  |
| Level-3| 20 by 20|

3 Document Representation

After training the three levels of SOMs, we found that documents from the same category share some parts of the BMU sequences to each other. Fig. 3 shows that BMU sequences of two documents from category “Earn”. Moreover, those shared BMU sequences between the documents belong to the top frequent BMU sequences of their corresponding category. Experiments also showed that different categories have different top frequent BMU sequences. To this end, we proposed the document representation by using a sequence of BMUs on the third level SOM and distances to the corresponding BMUs as shown in Fig. 4.
4 K-Nearest Neighbor Classifier-Learning Algorithm

kNN stands for k-Nearest Neighbor (kNN) classification. It has been studied extensively for text categorization by Yang and Liu [9]. The kNN algorithm is quite simple: To classify a test document, the k-Nearest Neighbor classifier algorithm finds the k nearest neighbors among the training documents, and uses the category labels of the k nearest training documents to predict the category of the test document. In general, Euclidean distance, or cosine distance is used to measure the similarity between the documents. However, those distance calculations are for vector space representation. In this work, we designed a similarity measurement as (1), which fits to this sequential data representation.

$$Sim(D_i, D_j) = \sum_{k=1}^{n} \frac{100}{1 + \text{dist}(W_{ik}, W_{jk})} \times n$$ (1)

$D_i$: Test document to be categorized.
$D_j$: Document in the training set.
$n$: Total number of BMUs shared by BMU sequences of $D_i$ and $D_j$.
$\text{dist}(W_{ik}, W_{jk})$: The Euclidean distance between the $W$ (defined in Fig. 4) of the corresponding BMU in the shared BMU sequences of $D_i$ and $D_j$.

5 Experimental Setup and Categorization Result

In this work, we used the well-known multi-class, multi-labeled document set – Reuters-21578\(^1\), which is a large research-oriented corpus to evaluate the approaches.

---

\(^1\) Reuters data set, http://www.daviddlewis.com/resources/testcollections/reuters21578/
There are a total of 12612 news stories in this collection, where 9603 of them are in the training data set, and 3299 are in the test set. In general, when a document has more than one label, those category labels are very similar or strongly related to each other. Thus, to automatically explore the relationship of those similar categories becomes a challenge for text categorization.

In order to analyze the efficiency of the representation for the multi-class, multi-labeled document categorization, we analyzed complex relationships and overlap between the top 10 categories of this data set. Based on the information we got, we chose 6 of them to test the categorization performance. The size and relationship between these categories in training set are shown in Fig. 5, and is the same as their relationship in the test set. We could see that “Grain”, “Wheat” and “Corn” are strongly related to each other, as are “Money-fx” and “Interest”. “Earn”, the biggest category in the whole corpus, has no overlap with the other five categories.

In our experiments, we set the number of nearest neighbors – k = 3, 5, 10 and 15. It turns out by experiments that k = 5 gives the best performance in terms of the Micro F1-measure score.

Facing a multi-labeled (N labels) test document, we first calculate the similarity between the test document and the training documents in the selected categories using formula (1). After ranking those similarities, we select 5 of them and weight the categories they belong to. Finally, we classify the test document to the top N weighted categories which corresponding to the number of the labels of the test document.

The classical effectiveness measurements of multi-labeled text categorization: Recall(R), Precision (P) and F-measure (F) are used to measure the categorization performance. Table 2 summaries the results obtained through the experiments.

From the achieved performance above, this encoding architecture capture the characteristic sequences for documents and categories. Good performance is achieved by utilizing the sequences information for categorization. However, the results show that it works better for some categories, especially the category “Earn”. We conclude
Table 2. Categorization results of all six categories

| Category  | Size in test set | Recall | Precision | F1-measure |
|-----------|------------------|--------|-----------|------------|
| Earn      | 1087             | 0.958  | 0.956     | 0.957      |
| Money-fx  | 179              | 0.646  | 0.662     | 0.654      |
| Interest  | 131              | 0.539  | 0.738     | 0.623      |
| Grain     | 149              | 0.741  | 0.708     | 0.724      |
| Wheat     | 71               | 0.721  | 0.671     | 0.695      |
| Corn      | 56               | 0.623  | 0.702     | 0.660      |

Micro Average F1-measure: 0.855

the reasons behind this are: first, this data representation is based on the machine-learning algorithm to capture the characteristic word co-occurrence for categories, so the more frequent the word co-occurrence is, the more easily it can be caught and represented by the neurons of the third level SOM. Second, for some categories, there is less variety of word co-occurrences in the category, so that all of them can be captured and represented well by the SOM neurons. However, for those categories with more variety of word co-occurrences, because of the size of SOM, some word co-occurrences of them may not be represented well enough by the neurons on the SOM. The characteristic word co-occurrences from different categories may mix together on the same neuron. This impresses the performance for those categories.

6 Conclusions and Future Work

Through this work, we explored a new way of data representation specifically designed for text representation. The results (in section 5) show that this architecture works well for capturing the characteristics of documents/categories using temporal sequences of word co-occurrences. The performance of this new data representation has been tested for document categorization by using a kNN classifier on top of it. We end up with the Micro average F1-measure for the selected data set at 0.855.

The efficiency of this new representation presented in this paper is still far from being completely elaborated, so we definitely consider this as a work in progress. Future work will include testing the performance on different size of SOMs, performing experiments on utilizing the temporal sequences analysis for extensive document categorization as well as classification of data for other applications such as medical and/or business information systems in which the analysis of temporal sequences of information is very important. And other classifiers, which fit more to the sequences representation, will also be analyzed and utilized in the future.

References

1. R. Basili, A. Moschitti, and M. T. Pazienza. Language-sensitive text classification. In Proceedings of 6th International Conference “Recherche d’Information Assistee par Orindateur”, (2000) 331-343
2. E. Brill. A simple rule-based part of speech tagger. In Proceedings of 3rd Conference on Applied Natural Language Processing, (1992) 152-155
3. M.F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. Text Databases and Document Management: Theory and Practice, (2001) 78-102
4. S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In Proceedings of 7th ACM International conference on Information and knowledge management, (1998) 148-155
5. P.S. Jacobs. Joining statistics with NLP for text categorization. In Proceedings of the Third conference on Applied Natural Language Processing, (1992) 178-185
6. T. Joachims. Text Categorization with support vector machines: learning with many relevant features. In Proceedings of ECML’98, 10th European Conference on Machine Learning, (1998) 137-142
7. G. Salton, A. Wang, and C.S. Yang. A vector space model for information retrieval. Journal of the American Society for information Science, 18, (1975) 613-620
8. F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1), (2002) 1-47
9. Y. Yang, X, Liu. A re-examination of text categorization methods. In Proceedings of SIGIR’99, (1999) 42-49
10. S. M. Weiss, C. Apte, F. J. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. IEEE Intelligent Systems, 14(4), (1999) 63-69
Prediction of Women’s Apparel Sales Using Soft Computing Methods

Les M. Sztandera, Celia Frank, and Balaji Vemulapali
Philadelphia University, Philadelphia, PA 19144, USA
SztanderaL@PhilaU.edu

Abstract. In this research, forecasting models were built based on both univariate and multivariate analysis. Models built on multivariate fuzzy logic analysis were better in comparison to those built on other models. The performance of the models was tested by comparing one of the goodness-of-fit statistics, $R^2$, and also by comparing actual sales with the forecasted sales of different types of garments. Five months sales data (August-December 2001) was used as back cast data in our models and a forecast was made for one month of the year 2002. The performance of the models was tested by comparing one of the goodness-of-fit statistics, $R^2$, and also by comparing actual sales with the forecasted sales. An $R^2$ of 0.93 was obtained for multivariate analysis (0.75 for univariate analysis), which is significantly higher than those of 0.90 and 0.75 found for Single Seasonal Exponential Smoothing and Winters’ three parameter model, respectively. Yet another model, based on artificial neural network approach, gave an $R^2$ averaging 0.82 for multivariate analysis and 0.92 for univariate analysis.

1 Introduction

Sales Forecasting is an integral part of apparel supply chain management and very important in order to sustain profitability. Apparel managers require a sophisticated forecasting tool, which can take both exogenous factors like size, price, color, and climatic data, price changes, marketing strategies and endogenous factors like time into consideration. Although models built on conventional statistical forecasting tools are very popular they model sales only on historic data and tend to be linear in nature (Kincade et. al., 1998). Soft computing tools like fuzzy logic and Artificial Neural Networks (ANN) can efficiently model sales taking into account both exogenous and endogenous factors and allow arbitrary non-linear approximation functions derived (learned) directly from the data (Kuo and Xue, 1999).

In order to reduce their stocks and to limit stock out, textile companies require specific and accurate sale forecasting systems. One of the approaches (Thomassey et. al., 2004) analyses two complementary forecasting models, appropriate to textile market requirements. The first model (AHFCCX) allows to automatically obtain mean-term forecasting by using fuzzy techniques to quantify influence of explanatory variables. The second one (SAMANFIS), based
on a neuro-fuzzy method, performs short-term forecasting by readjusting mean-term model forecasts from load real sales. In yet another approach (Ansuj et. al, 1996) the researchers compared the use of time series ARIMA model with interventions, and neural network back propagation model in analyzing the behavior of sales in a medium size enterprise. The forecasts obtained using the back propagation model was reported to be more accurate than those of ARIMA model with interventions.

In our approach a multivariate fuzzy model has been built based on important product variables of color, time and size. This model is being currently further extended to include other variables like climate, economic conditions etc., which would be used in building a comprehensive forecasting software package.

2 Methodology and Results

Since our present research is based on multivariate analysis, sales data containing multiple independent variables are being used in a multivariable fuzzy logic and ANN models. Two product variables color, and combined time and size, which significantly affect apparel sales, were chosen to model sales. The converted data were grouped based on different class-size combinations, trained, and then sales were forecasted for each grouping using fuzzy logic and ANN modeling.

Fuzzy Logic Approach. Fuzzy logic allows the representation of human decision and evaluation in algorithmic form. It is a mathematical representation of human logic. The use of fuzzy sets defined by membership function constitutes fuzzy logic (Von Altrock, 1995). The basic terms are summarized below.

Fuzzy Set: is a set with graded membership over the interval [0, 1].

Membership function: is the degree to which the variable is considered to belong to the fuzzy set.

A sales fuzzy logic controller is made of:

Fuzzification: Linguistic variables are defined for all input variables (color and size).

Fuzzy Inference: rules are compiled from the database and based on the rules, the value of the output linguistic variable is determined. Fuzzy inference is made of two components:

Aggregation: Evaluation of the IF part of the rules.
Composition: Evaluation of the THEN part of the rules.

Defuzzification: linguistic value(s) of output variable (sales) obtained in the previous stage are converted into a real output value. This can be accomplished by computing typical values and the crisp result is found out by balancing out the results.

Fuzzy logic model was applied to grouped data and sales values were calculated for each size-class combination. Total sales value for the whole period was calculated by summing up the sales values of all the grouped items. The daily sales were calculated from grouped sales using two different methods: fractional contribution method and Winters’ three parameter model. The forecasted daily
sales were then compared with actual sales by using goodness-of-fit statistics, $R^2$.

**Fractional Contribution Method.** It was observed that the fraction contribution of each weekday towards total week sales was constant (Frank et. al., 2002). Figure 1 depicts the average fractional contribution of a weekday towards total sales of a week, which can be used to forecast the daily sales from the forecasted weekly sales. The daily sales were calculated as a fraction of total sales.

![Figure 1. Fraction of Weekly Sales Distributed Among 7 Days](image)

The $R^2$ of the model was 0.93 and the correlation coefficient $R$ between actual and forecasted daily sales for October 2002 was 0.96. Figure 2 shows the actual versus forecasted sales values for October-2002 month.

![Figure 2. Actual vs. forecasted sales for October 2002 using fuzzy model](image)

**Winters’ Three Parameter Exponential Smoothing Model.** Winters’ smoothing model assumes that:

$$Y_{t+m} = (S_t + b_t)I_{t-L+m}$$  \hspace{1cm} (1)

where: $S_t$ = smoothed nonseasonal level of the series at end of $t$, $b_t$ = smoothed trend in period $t$, $m$ = horizon length of the forecasts of $Y_{t+m}$, $I_{t-L+m}$ = smoothed seasonal index for period $t + m$. That is, $Y_{t+m}$ the actual value of a series equals a smoothed level value $S_t$ plus an estimate of trend $b_t$ times a
seasonal index $I_{t-L+m}$. These three components of demand are each exponentially smoothed values available at the end of period $t$ (DeLurigo, 1998). The smoothed values were estimated as follows:

$$S_t = \alpha \left( \frac{Y_t}{I_{t-L}} \right) + (1 - \alpha)(S_{t-1} + b_{t-1})$$  \hspace{1cm} (2)$$

$$b_t = \beta(S_t - S_{t-1}) + (1 - \beta)b_{t-1}$$  \hspace{1cm} (3)$$

$$I_t = \gamma \left( \frac{Y_t}{S^t} \right) + (1 - \gamma)I_{t-L+m}$$  \hspace{1cm} (4)$$

$$Y_{t+m} = (S_t + b_t m)I_{t-1+m}$$  \hspace{1cm} (5)$$

where: $Y_t = \text{value of actual demand at end of period } t$, $\alpha = \text{smoothing constant used for } S_t$, $S_t = \text{smoothed value at end of } t$ after adjusting for seasonality, $\beta = \text{smoothing constant used to calculate the trend (} b_t \text{)}, b_t = \text{smoothed value of trend through period } t, I_{t-L} = \text{smoothed seasonal index } L \text{ periods ago, } L = \text{length of the seasonal cycle (e.g., 5 months), } \gamma = \text{smoothing constant, gamma for calculating the seasonal index in period } t, I_t = \text{smoothed seasonal index at end of period } t, m = \text{horizon length of the forecasts of } Y_{t+m}$. Equation 2 is required to calculate the overall level of the series. $S_t$ in equation 3 is the trend-adjusted, deseasonalized level at the end of period $t$. $S_t$ is used in equation 5 to generate forecasts, $Y_{t+m}$. Equation 3 estimates the trend by smoothing the difference between the smoothed values $S_t$ and $S_{t-1}$. This estimates the period-to-period change (trend) in the level of $Y_t$. Equation 4 illustrates the calculation of the smoothed seasonal index, $I_t$. This seasonal factor is calculated for the next cycle of forecasting and used to forecast values for one or more seasonal cycles ahead. Alpha, beta, and gamma values were chosen using minimum mean squared error (MSE) as the criterion. Applying a forecast model built on five months sales data, a daily forecast of sales ratio was done for October of 2002. Figure 3 shows the actual versus forecasted sales values for October-2002 month. The parameters used were: $\alpha=0.6$, $\beta=0.01$, $\gamma=1$, and $R^2=0.97$, $R=0.98$.

![Figure 3. Actual vs. forecasted for fuzzy approach with Winters three par. model](image)

**Neural Network Model.** In our research, a feed forward neural network, with back propagation, was implemented with 10 neurons in the input layer, 30 neurons in the hidden layer and 1 neuron in the output layer. Grouped sales data over a period of 10 months was used, out of which the first 32 rows were used as
training set, next 34 rows were used in test set and the last 234 rows were used in production set.

**Fractional Contribution Method.** The fractional contribution method described under fuzzy logic section was implemented for NN model. $R^2$ of the model was 0.82, and the correlation coefficient $R$ between actual and forecasted daily sales for October 2002 was 0.93. Figure 4 shows the actual versus forecasted sales values for October-2002 month.

![Fig. 4. Actual vs. forecasted sales by using ANN](image)

**Winters’ Three Parameter Model.** The winters’ three parameter model method described under fuzzy logic section was implemented for NN model. The following parameters were used: $\alpha = 0.6$, $\beta = 0.01$, $\gamma = 1$, and $R^2 = 0.44$, $R = 0.67$ were obtained. Figure 5 shows the actual versus forecasted sales values for October-2002 month.

![Fig. 5. Actual vs. forecasted sales using ANN](image)

**Univariate Forecasting Models.** Forecasting models were built on univariate analysis using both conventional statistical models as well as unconventional soft-computing methods. Among all the models, the ANN model performed the best. However all the models could not forecast with precision because they were built using a single variable time. A plot of actual versus forecasted sales for various models done using univariate analysis are shown in Figures 6, 7 and 8.
Fig. 6. Actual vs. forecasted sales for SES model ($R^2=0.90$)

Fig. 7. Actual vs. forecasted sales for Winters’ three parameter model ($R^2=0.75$)

Fig. 8. Actual vs. forecasted sales for ANN model ($R^2=0.92$)

Fig. 9. Goodness of fit statistic for models based on multivariate analysis
3 Conclusions

Multivariable fuzzy logic model can be an effective sales forecasting tool as demonstrated by our results. A correlation of 0.93 was obtained, better than that obtained by using the NN model, which showed a correlation of 0.82 (for the fractional contribution method). The values for the three parameter model were: 0.97 and 0.44, respectively. The poor correlation in the case of the NN model can be attributed to the noise in the sales data. The fuzzy model performed best because of its ability to identify nonlinear relationships in the input data. However, the correlation was better for short-term forecasts and not as good for longer time periods. However the multivariate fuzzy logic model performed better in comparison to those based on univariate analysis, which goes on to prove that multivariate analysis is better compared to that of univariate analysis. A much more comprehensive model can be built by taking into account other factors like climate, % price change, marketing strategies etc.

References

Ansuj A. P., Camargo M. E., Radharamanan R. (1996), Sales Forecasting using Time Series and Neural Networks, Computers and Industrial Engineering 31(1/2), 421-425.

Frank C., Garg A., Raheja A., Sztandera L. (2002), Forecasting Women’s Apparel Sales Using Mathematical Modeling, International Journal of Clothing Science and Technology 15(2), 107-125.

Kincade D. H., Cassill N., Williamson N. (1993), The Quick Response Management System: Structure and Components for the Apparel Industry, Journal of Textile Institute 84, 147-155.

Kuo, R. J., Xue K. C. (1999), Fuzzy Neural Networks with Application to Sales Forecasting, Fuzzy Sets and Systems 108(2), 123-155.

Thomassey S., Happiette M., Castelain J. M. (2004), A Short and Mean-term Automatic Forecasting System - Application to Textile Logistics, European Journal of Operational Research, in press.

Von Altrock C., (1995), Fuzzy Logic and Neuro Fuzzy Applications Explained, Prentice-Hall, Upper Saddle River.
A Try for Handling Uncertainties in Spatial Data Mining

Shuliang Wang¹,², Guoqing Chen¹, Deyi Li³, Deren Li⁴, and Hanning Yuan³

¹School of Economics and Management, Tsinghua University, Beijing 100084, China
²International School of Software, Wuhan University, Wuhan 430072, China
³China Institute of Electronic System Engineering, Fuxing Road 20, Beijing 100039, China
⁴School of Remote Sensing Engineering, Wuhan University, Wuhan 430072, China

hnyuanslwang@yahoo.com

Abstract. Uncertainties pervade spatial data mining. This paper proposes a method of spatial data mining handling randomness and fuzziness simultaneously. First, the uncertainties in spatial data mining are presented via characteristics, spatial data, knowledge discovery and knowledge representation. Second, the aspects of the uncertainties in spatial data mining are briefed. They often appear simultaneously, but most of the existing methods cannot deal with spatial data mining with more than one uncertainty. Third, cloud model is presented to mine spatial data with both randomness and fuzziness. It may also act as an uncertainty transition between a qualitative concept and its quantitative data, which is the basis of spatial data mining in the contexts of uncertainties. Finally, a case study on landslide-monitoring data mining is given. The results show that the proposed method can well deal with randomness and fuzziness during the process of spatial data mining.

1 Introduction

There are uncertainties in spatial data mining. People are faced with large amounts of spatial data but are short of knowledge, which promotes spatial data mining [1]. The uncertainties are the major component of spatial data quality, and a number of methods has been tried to deal with the elements, measurement, modeling, propagation, and cartographic portrayal [4]. The uncertainties are inherent in most of the data capturing and data analyzing because of the limitations of current instruments, technologies, capitals, and human skills. Because the spatial data are the objectives of spatial data mining, the uncertainties are brought to spatial data mining along with spatial data at the beginning [5]. Then, new uncertainties will further come into being during the process of spatial data mining. It is an uncertain process for spatial data mining to discover the little-amount knowledge from the large-amount data because of variously mining angles, scales, and granularities [3]. And the indices of the discovered knowledge, e.g. interesting degree, supporting degree and confidential degree, are all uncertain. The uncertainties may directly or indirectly affect the quality of a spatial decision-making based on spatial data mining.

However, the uncertainties have not been addressed to the same degree to spatial data mining itself [6]. Although there have been some methods and techniques on spatial data mining, or on spatial data uncertainties [5], each of them is developed in its own direction. First, most of the existing models may describe some specific situation. It is difficult for them to deal with the case where more than one uncertainty
appears at the same time, e.g. both fuzziness and randomness. In fact, the cases with many uncertainties often happen in spatial data mining. Second, some models may be far beyond the comprehension of the common users. Without enough background knowledge, these users may have difficulty in making sense of the exact nature of uncertainty that an expert specifies. Third, it is an essential issue for spatial data mining to transform between a qualitative concept and its quantitative data. Commonly, the transition models are of rigid specification and too much certainty, which comes into conflict with the human recognition process. Fourth, almost none of the existing models are unable to well deal with the uncertainties in spatial data mining, and it is strange to find out the integration of spatial data mining and spatial data uncertainties. In order to continue enjoying its success, spatial data mining should think of the uncertainties carefully, and the theories to handle the uncertainties may have to be further studied.

2 Uncertainties Inherent in Spatial Data Mining

Spatial uncertainties indicate the unknown degree of the observed entities. In spatial data mining, they may arise from the objective complexity of the real world, the subjective limitation of human recognition, the approximate weakness of computerized machine, and the computerized shortcomings of techniques and methods, the amalgamation of heterogeneous data, the discovery, representation and interpretation of knowledge, and so on. During the process of spatial data mining, the original uncertainties in spatial data may be further propagated from the beginning to the end, and they are also affected by the scale, granularity and sampling in spatial data mining. And these uncertainties may have to be identified instead of presenting them as being correct [5].

First, there are many sources and causes of uncertainties, e.g. instruments, environments, observers, projection algorithms, slicing and dicing, coordinate system, image resolutions, spectral properties, temporal changes, etc. Spatial data stored in the databases are to describe and represent how the spatial entities are in the infinitely complex world via binary digits to approach them. The spatial database is only an abstracted representation with uncertainties. For it works with the spatial database as a surrogate for the real entities, spatial data mining is unable to avoid the uncertainties.

Second, spatial data mining is an uncertain process. In a spatial computerized system that observes and analyzes the same spatial entities on variant levels of granularity, and/or on different worlds of different granularities, it is common to be faced with having to use data that are less detailed than one would like, and then some data will be further eliminated or got rid of when the spatial data are edited, stored, and analyzed. The unknown knowledge is refined with a high abstraction level, small scales, and small granularities, whereas the existing data are coarse with a low abstraction level, big scales, and big granularities. Sampling creates a representation from limited data, leaving uncertainty as to what actually exists between the sample points. As to the same dataset, different knowledge may be mined when different people apply the same technologies, or the same people apply different technologies.
Third, there exist uncertainties in knowledge representation. The discovered knowledge is unknown in advance, potentially useful, and ultimately understandable. Knowledge uncertainty arises when roll-up or drill-down is carried out in spatial data mining, and there is also a gap to be bridged between the rigidity of computerized spatial data and the uncertainty of the spatial qualitative concept, i.e. spatial transition between the qualitative concept and the quantitative data.

Fourth, the performance and nature of uncertainty are various, i.e. randomness, fuzziness, chaos, positional uncertainty, attribute uncertainty, incompleteness. For example, randomness is included in a case with a clear definition but not always happens every time, and fuzziness is the indetermination between a proposed and incomplete value but cannot be defined exactly.

3 Cloud Model on Randomness and Fuzziness

A cloud model \(^2\) is a mathematical model of the uncertainty transition between a linguistic term of a qualitative concept and its numerical representation data. It is named after the natural cloud in the sky for both are visible in a whole shape but fuzzy in detail. A piece of cloud is not a membership curve but is composed of many cloud-drops, any one of which is a stochastic mapping in the discourse universe from a qualitative fuzzy concept. As well, the degree of any cloud-drop is specified to represent the qualitative concept when the one-to-many transition is carried out. The cloud model integrates the fuzziness and randomness via three digital characteristics \{Ex, En, He\} (Fig.1).

![Cloud Model Diagram](image)

**Fig. 1.** \{Ex, En, He\} of the linguistic term” displacement is 9 millimeters around”

In the discourse universe, Ex (Expected value) is the position corresponding to the center of the cloud gravity, the elements of which are fully compatible with the spatial linguistic concept; En (Entropy) is a measure of the concept coverage, i.e. a measure of the spatial fuzziness, which indicates how many elements could be accepted to the spatial linguistic concept; and He (Hyper-Entropy) is a measure of the dispersion on the cloud-drops, which can also be considered as the entropy of En.
Cloud generators may be forward or backward in the context of the integrity \{Ex, En, He\}. Given \{Ex, En, He\}, the forward cloud generator can produce as many cloud-drops as you would like, which may visualize the discovered knowledge. The input of the forward cloud generator is \{Ex, En, He\}, and the number of cloud-drops to be generated, N, while the output is the quantitative positions of N cloud-drops in the data space and the certain degree that each cloud-drop can represent the linguistic term. On the other hand, the backward cloud generator may mine \{Ex, En, He\} of cloud-drops specified by many precise data points, which discovers the knowledge from the given spatial database. The input of the backward cloud generator is the quantitative positions of N cloud-drops, \(x_i\) (i=1,\ldots,N), and the certainty degree that each cloud-drop can represent a linguistic term, \(y_i\) (i=1,\ldots,N), while the output is \{Ex, En, He\} of the linguistic term represented by the N cloud-drops.

During the process of knowledge discovery with the cloud model, the quantitative data first produce several essential cloud models. Then the roll-up is carried out one by one, and the linguistic atoms also become linguistic terms, and further concepts. The higher the roll-up, the more generalized the qualitative concept. The concept that can attract the interest, match the demand, and support the decision-making will be the knowledge. The top hierarchy of spatial data mining is the most generalized knowledge, while the bottom hierarchy of spatial data mining is the objective data in the spatial database. It is the virtual cloud model that implements the roll-up and drill-down in spatial data mining, i.e. floating cloud, synthesized cloud, resolved cloud, and geometric cloud.

4 A Case Study

The spatial database is 1G bytes on the displacements of Baota landslide, on which 2,000 people are living. The properties of \(dx\), \(dy\) and \(dh\), are the measurements of displacements in \(X\), \(Y\) and \(H\) direction of the landslide-monitoring points. In Baota landslide data mining, there exist uncertainties, e.g. randomness and fuzziness, and different people may discover various rules with different techniques. In the following, all spatial knowledge is discovered from the databases with \(dx\).

![Fig. 2. Pan-concept hierarchy tree of different displacements](image-url)
From the observed values, the backward cloud generator can mine \( \text{Ex} \), \( \text{En} \) and \( \text{He} \) of the linguistic term indicating the average level of those landslide displacements. Based on landslide-monitoring characteristics, let the linguistic concepts of “smaller (0~9mm), small (9~18mm), big (18~27mm), bigger (27~36mm), very big (36~50mm), extremely big (50mm)” with \( \text{Ex} \), “lower (0~9), low (9~18), high (18~27), higher (27~36), very high (36~50), extremely big (50)” with \( \text{En} \), “more stable (0~9), stable (9~18), instable (18~27), more instable (27~36), very instable (36~50), extremely instable (50 and over)” with \( \text{He} \) respectively depicting the displacements, scattering levels and stabilities of the displacements. Then, the linguistic terms of different displacements on \( dx \), \( dy \) and \( dh \) may be depicted by the conceptual hierarchy tree in the conceptual space (Fig. 2). Fig. 3 presents the cloud models of Fig. 2 in the discourse universe.

![Concept hierarchy](image)

**Fig. 3.** The cloud models of pan-concept hierarchy tree of different displacements

It can be seen from Fig. 2 and Fig. 3 that the nodes “very small” and “small” both have the son node “9 millimeters around”, so the concept hierarchy tree is a pan-tree structure. In the context of the cloud model, the qualitative concept from the quantitative data may be depicted via the cloud generators. Based on the gained \( \{\text{Ex}, \text{En}, \text{He}\} \), the forward cloud generator can reproduce as many deterministic cloud-drops as you would like, i.e. producing synthetic values of landslide displacements. These cloud-drops are reproduced with randomness, and they can be further taken as virtually monitoring Baota landslide under the umbrella of given conditions. The virtual monitoring data may further fill in the incompleteness when it is unable to establish monitoring points on some typical surfaces of Baota landslide. With the forward cloud generator and backward cloud generator, the level of monitoring-points’ displacements is extended to the whole landslide. This may
approach the moving rules of Baota landslide well. Thus the rules on Baota landslide in X direction can be discovered from the databases in the conceptual space (Table 1). Because large amounts of consecutive data are replaced by discrete linguistic terms in Table 1, the efficiency of spatial data mining can be improved. Meanwhile, the resulting knowledge will be stable due to the randomness and fuzziness of concept indicated by the cloud model. Fig. 4 visualizes the displacing rule of each point with 30,000 pieces of cloud-drops, where the symbol of “+” is the original position of monitoring point, the different rules are represented via the different pieces of cloud, and the level of color in each piece of cloud denotes the discovered rules of a monitoring-point.

| Points | Rules |
|--------|-------|
| BT11   | The displacements are big south, high scattered and instable. |
| BT12   | The displacements are big south, high scattered and very instable. |
| BT13   | The displacements are small south, lower scattered and more stable. |
| BT14   | The displacements are smaller south, lower scattered and more stable. |
| BT21   | The displacements are extremely big south, extremely high scattered and extremely instable. |
| BT22   | The displacements are bigger south, high scattered and instable. |
| BT23   | The displacements are big south, high scattered and extremely instable. |
| BT24   | The displacements are big south, high scattered and more instable. |
| BT31   | The displacements are very big south, higher scattered and very instable. |
| BT32   | The displacements are big south, low scattered and more instable. |
| BT33   | The displacements are big south, high scattered and very instable. |
| BT34   | The displacements are big south, high scattered and more instable. |

Fig. 4 indicates that all monitoring points move to the direction of Yangtze River, i.e. south, or the negative axe of X. Moreover, the displacements are different from each other. BT21 are extremely big south, extremely high scattered and extremely instable, and followed by BT31. At least, BT14 is smaller south, lower scattered and more stable. In a word, the displacements of the back part of Baota landslide are bigger than those of the front part in respect of Yangtze River, and the biggest exceptions are BT21.

When the Committee of Yangtze River investigated in the region of Baota landslide, they found out that the landslide had moved to Yangtze River. By the landslide BT21, a small size landslide had taken place. Now there are still two pieces of big rift. Especially, the wall rift of the farmer G. Q. Zhang’s house is nearly 15 millimeters. These results match the discovered spatial knowledge very much, and indicate that the method of randomness and fuzziness -based spatial data mining in the context of cloud model are creditable.
5 Conclusions

There were inherent uncertainties in spatial data mining. This paper proposed a method to handle randomness and fuzziness simultaneously in spatial data mining, by giving the cloud model to realize the transition between a qualitative concept and its quantitative data. It includes the algorithms of forward and backward cloud generators in the contexts of three digital characteristics, \{Ex, En, He\}. The case study of Baota landslide monitoring showed that the method was practical and confident, and the discovered knowledge with a hierarchy can match different demands from different users.

Acknowledgements

This study is supported by the funds from National Natural Science Foundation of China (70231010), Wuhan University (216-276081), and National High Technology R&D Program (863) (2003AA132080).

References

1. ESTER M. et al., 2000, Spatial data mining: databases primitives, algorithms and efficient DBMS support. Data Mining and Knowledge Discovery, 4, 193-216
2. LI D.Y., 1997, Knowledge representation in KDD based on linguistic atoms. Journal of Computer Science and Technology, 12(6): 481-496
3. MILLER, H.J., HAN, J., 2001, Geographic Data Mining and Knowledge Discovery (London: Taylor & Francis)

4. VIKTOR H.L., PLOOY N.F. D., 2002, Assessing and improving the quality of knowledge discovery data. In: Data Warehousing and Web Engineering, edited by Becker S. (London: IRM Press), pp. 198-205

5. WANG S.L., 2002, Data field and cloud model based spatial data mining and knowledge discovery. Ph.D. Thesis (Wuhan: Wuhan University)

6. ZEITOUNI K., 2002, A survey of spatial data mining methods databases and statistics point of views. In: Data Warehousing and Web Engineering, edited by Becker S. (London: IRM Press), pp. 229-242
Combining Evidence from Classifiers in Text Categorization

Yaxin Bi, David Bell, and Jiwen Guan

School of Computer Science,
Queen's University of Belfast, Belfast, BT7 1NN, UK
{y.bi, da.bell, j.guan}@qub.ac.uk

Abstract. In this paper, we describe a way for modelling a generalization process involved in the combination of multiple classification systems as an evidential reasoning process. We first propose a novel structure for representing multiple pieces of evidence derived from multiple classifiers. This structure is called a focal element triplet. We then present a method for combining multiple pieces of evidence by using Dempster's rule of combination. The advantage of the novel structure is that it not only facilitates the distinguishing of trivial focal elements from important ones, but it also reduces the effective computation-time from exponential as in the conventional process of combining multiple pieces of evidence to linear. In consequence, this allows Dempster's rule of combination to be implemented in a widened range of applications.

1 Introduction

In text categorization, a number of researchers have shown that combining multiple classifiers consisting of different classification methods can improve the classification accuracy [1]. In the literature [2], Sebastiani provides a state-of-the-art review on text categorization, including this aspect. It identifies three combination functions or rules used for combining multiple classifiers in text categorization, including majority voting (MV) [3], weighted linear combination (WLC) [1, 4], and dynamic classifier selection (DCS) [3]. In this paper, we present an investigation into an evidential approach for combining multiple classifiers. This work is inspired by an idea from artificial intelligence research, viz. a decision made on the basis of the multiple pieces of evidence should be more effective than one based on a single piece of evidence.

Machine learning offers many different techniques ranging from concept learning to reinforcement learning. The best-understood form of learning is supervised learning. Supervised learning methods for classification roughly span two categories: statistical similarity-based methods and induction of production rules or decision trees. A method from either category is usually applicable to the problem of text categorization, but the two categories of procedure can differ radically in their underlying models and the final format of their solution [5].

Generally speaking, text categorization systems built on the basis of supervised similarity-based learning methods consist of two sub-methods: learning algorithms, and decision making mechanisms. The former aims at finding how best to use
historical data – training data – to learn general regularities which can be presented in the form of production rules or classification models (classifiers) depending on the specific learning methods and tasks. The latter is concerned with generalizing these regularities to any document instances through an inference or a decision making mechanism in order to determine which classes they should belong to. Therefore, the predictive accuracy of a text categorization system not only relies on the quality of learned the models learned, but it also depends on an effective inference mechanism. In this study, we focus on the latter aspect, and model such an inference mechanism as an evidential reasoning process – deriving evidence from classifiers and combining them by using Dempster’s rule of combination in order to make a final decision.

2 Background

In this section, we start by giving an overview of the Dempster-Shafer theory of evidence, and then introduce a general form that similarity-based learning methods may have, which suits for the task of text categorization well.

2.1 Overview of the Dempster-Shafer (D-S) Theory of Evidence

Consider a number of exhaustive and mutually exclusive propositions \( h_i, i = 1, \ldots, m \), which form a universal set \( \Theta \), called the frame of discernment. For any subset \( H \subseteq \Theta \), where \( h_i \) is called a focal element \((0 < r \leq k)\), and when \( H \) is a one element subset, i.e. \( H = \{ h_i \} \), it is called a singleton. All the subsets of \( \Theta \) constitute a powerset \( 2^\Theta \), that is, for any subset \( H \subseteq \Theta \), \( H \in 2^\Theta \). The D-S theory uses a numeric value in a range \([0, 1]\) to represent the strength of supporting a proposition \( H \subseteq \Theta \) based on a given piece of evidence, denoted by \( m(H) \), called a mass function. The D-S theory uses a sum of the strengths of all the subsets of \( H \) to indicate the degree of belief to the proposition \( H \), denoted by \( \text{bel}(H) \), called a belief function. When \( H \) is a singleton, \( m(H) = \text{bel}(H) \). The formal definition of the orthogonal sum operation is given below [6]:

**Definition 1.** Let \( m_1 \) and \( m_2 \) be two mass functions on the frame of discernment \( \Theta \), and for any subset \( H \subseteq \Theta \), the orthogonal sum \( \oplus \) of two mass functions on \( H \) is defined as:

\[
(m_1 \oplus m_2)(H) = \frac{\sum_{X \cap Y = H} m_1(X) \ast m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) \ast m_2(Y)}
\] (1)

Another name for the orthogonal sum is Dempster’s rule of combination. It allows two mass functions to be combined into a third mass function, and thus it pools pieces of evidence which support propositions of interest [6].
2.2 A General Output Form of Classification Methods

Generally speaking, a similarity-based learning algorithm for text categorization is aiming at learning a classifier or mapping which enables the assignment of documents into the predefined categories. In this setting, each text document which consists of a vector of features is denoted by \( d \), each category is denoted by \( c \), and it is assumed that there is some underlying function \( f \) such that \( c = f(d) \) for each pair of training documents \( \langle d, c \rangle \). The goal of the learning algorithm is to find a good approximation \( \phi \) to \( f \) that can be invoked to assign categories to new document instances. This function \( \phi \) is called a classifier or a mapping function.

Regardless of what internal structure a learning algorithm has and what theory and methodology it is based on, generally the algorithm works by searching through a space of possible functions, called hypotheses, to find one function \( \phi \). This function is the best approximation to unknown function \( f \)[5]. Due to \( \phi \) being an approximation to \( f \), function \( \phi \) cannot guarantee that an assignment of categories to a document is absolutely true or absolutely false. Instead it supplies a set of numeric values, denoted by \( S = \{ s_1, s_2, \ldots, s_{|\mathbb{C}|} \} \), which represents a measure of the similarity between the document and the set of categories in the form of similarity scores or probabilities, where the greater the score of the category, the greater the possibility of the document being under the corresponding category.

Formally, let \( \phi \) be a classifier, \( \mathbb{C} = \{ c_1, c_2, \ldots, c_{|\mathbb{C}|} \} \) be a set of categories, and \( \mathbb{D} = \{ d_1, d_2, \ldots, d_{|\mathbb{D}|} \} \) be a set of test documents, for any document \( d \in \mathbb{D} \), then we have \( \phi(d) = \{ s_1, s_2, \ldots, s_{|\mathbb{C}|} \} \). We regarded this expression as a piece of evidence derived from the classifier. For convenience of discussion, we define a function \( \mathfrak{m}, \mathfrak{m}(c_i) = s_i \) for all \( c_i \in \mathbb{C} \). Substituting \( \mathfrak{m}(c_i) \) for \( s_i \), we have \( \phi(d) = \{ \mathfrak{m}(c_i), \mathfrak{m}(c_2), \ldots, \mathfrak{m}(c_{|\mathbb{C}|}) \} \).

Analogously, if we refer to \( \mathbb{D} \) as an evidence space and regard \( \mathbb{C} \) as a hypothesis space, then the relationship between the two spaces is inherently uncertain. This uncertain relationship can be depicted by a mapping function \( \phi \), it alternatively can quantitatively be represented as mass functions and belief functions.

3 Define an Application-Specific Mass Function

Having introduced a general form of classifiers and a representation of the output information yielded by the classifiers, we now turn to the problem of estimating the degrees of belief for the evidence obtained from classifiers and the specific definitions of mass and belief functions for this domain. We then look at how to fuse multiple pieces of evidence in order to make a final decision.

**Definition 2.** Let \( \mathbb{C} \) be a frame of discernment, where each category \( c_i \in \mathbb{C} \) is a proposition that the document \( d \) is of category \( c_i \), and let \( \phi(d) \) be a piece of evidence that indicates a possibility that the document comes from each category \( c_i \in \mathbb{C} \). Then a mass function is defined a mapping, \( m: 2^\mathbb{C} \rightarrow [0,1] \), i.e. a basic probability assignment (bpa) to \( c_i \in \mathbb{C} \) for \( 1 \leq i \leq |\mathbb{C}| \) as follows:
\[ m(\{c_i\}) = \frac{\sigma(c_i)}{\sum_{j=0}^{\mid C \mid} \sigma(c_j)} \text{ where } 1 \leq i \leq \mid C \mid \quad (2) \]

This expresses the degrees of beliefs in respective propositions corresponding to each category to which a given document could belong.

With formula (4), the expression of the output information \( \varphi(d) \) is rewritten as \( \varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{\mid C \mid}\})\} \). Therefore, two or more outputs derived from different classifiers as pieces of evidence can be then combined using the orthogonal sum formula (1). In order to improve the efficiency of computing orthogonal sum operations and the accuracy of a final decision on the basis of the combined results, we have developed a new structure, called a focal element triplet, which partitions \( \varphi(d) \) into three subsets. A number of empirical evaluations have been carried out to examine its effectiveness, and the result shows it is particularly useful for decision making under uncertainty when there is insufficient or incomplete information. More theoretical work based on its validity can be found in [7].

**Definition 3.** Let \( C \) be a frame of discernment and \( \varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{\mid C \mid}\})\} \), where \( \mid \varphi(d) \mid \geq 2 \), a focal element triplet is defined as an expression of the form \( Y = \langle A_1, A_2, A_3 \rangle \), where \( A_1, A_2 \subseteq C \) are singletons, and \( A_3 \) is the whole set \( C \). These elements are given by the formulae below:

\[
\begin{align*}
A_1 &= \{c_i\}, c_i = \max\{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{\mid C \mid}\})\} \\
A_2 &= \{c_j\}, c_j = \max\{\{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{\mid C \mid}\})\} - m(\{c_i\})\} \\
A_3 &= C 
\end{align*}
\]

The associated mass function is given as follows:

\[
\begin{align*}
m(A_1) &= m(\{c_i\}) \\
m(A_2) &= m(\{c_j\}) \\
m(A_3) &= 1 - m(\{c_i\}) - m(\{c_j\}) \quad (4)
\end{align*}
\]

We call \( m(A_3) \) the uncommitted belief or ignorance associated with this mass function. It represents the belief that we do not wish to assign to either \( A_1 \) or \( A_2 \), so that their belief is committed to the frame of discernment \( C \).

Note that the alternative element \( A_2 \) is defined to be a singleton. However it can alternatively be defined as the complementary subset of \( A_1 \), i.e. \( A_2 = C - A_1, A_2 \subseteq 2^C \), referred to as a dichotomous structure, which is used in [9]. In this work we prefer to choose \( A_2 \) as a singleton. This choice is inspired by the observations of the ranking process. As we know, to assign one and only one category to a test document, a common approach is to rank the category assignments \( \varpi(c_1), \varpi(c_2), \ldots, \varpi(c_{\mid C \mid}) \) in descending order. The top choice is seen as the true category of the document. However, this is the ideal case and it really depends on the performance of classification methods. It is not possible for the true category always to end on the top of the ranking list. Instead, it may be in another position of the list, such as second top. Because of this, here we make the assumption that the classes to be assigned to a
Given instance only be among the top choice, the top second choice, or the whole of the frame in descending order. It is then possible that the top second choice will be ranked as the top choice when we combine multiple classifiers. This assumption forms the rationale behind dividing $\varphi(d)$ into a triplet.

4 The Combination Method

Given a classification algorithm and a set of training data, the algorithm can generate one or more classifiers based on the training method chosen. For example, using ten-fold cross-validation, ten classifiers will be generated.

In general, $n$-fold cross-validation divides the set of training data into $n$ folds or subsets. Based on these $n$ subsets, $n$ different classifiers are constructed, each using $n$-1 subsets for training and the remaining data for validation. This procedure yields a statistically sound estimate for gauging the performance of the classifiers. Once all the $n$ cross-validation iterations have been performed, the performance of the classification method is the average of the performances of all the $n$ classifiers on the corresponding validation sets.

Given a group of classification methods for the same classification task, using $n$-fold cross-validation on a set of training data, each method will generate $n$ classifiers. Formally, let $\{\varphi_0^k, \varphi_1^k, \ldots, \varphi_n^k\}_{k=1}^K$ be a group of classifiers generated by $K$ learning methods. Then for a test document $d$, each classifier can produce a triplet $Y$, denoted by $\varphi_i^k(d) = Y_i^k$, where $1 \leq i \leq n$. With $K$ pieces of evidence $Y_i^k$, we will have a collection of $K$ triplet mass functions $m_1, m_2, \ldots, m_K$. In order to combine these triplet mass functions, we can perform pairwise orthogonal sums as follows:

$$m = m^1 \oplus m^2 \oplus \ldots \oplus m^K = [...[[m^1 \oplus m^2] \oplus m^3] \oplus \ldots \oplus m^K]$$  \hspace{1cm} (5)

Suppose we are given two triplets $\langle A_1, A_2, C \rangle$ and $\langle B_1, B_2, C \rangle$ where $A_1 \in C, B_1 \in C$, and two associated triplet mass functions $m_1, m_2$. Prior to combining them, we first examine the relationships between two pairs of focal elements $A_1, A_2$ and $B_1, B_2$ below:

1) if $A_1 = B_1$ and $A_2 = B_2$, then $A_1 \cap B_1 = \phi$ and $A_2 \cap B_1 = \phi$, so the combination of two triplet functions involves three different focal elements

2) if $A_1 = B_1$ and $A_2 \neq B_2$, then $A_1 \cap B_2 = \phi$, $A_2 \cap B_1 = \phi$ and $A_2 \cap B_2 = \phi$ or if $A_2 = B_2$ and $A_1 \neq B_1$, then $A_1 \cap B_2 = \phi$, $A_2 \cap B_2 = \phi$ and $A_1 \cap B_1 = \phi$, so the combination of two triplet functions involves four different focal elements.

3) if $A_1 \neq B_1, A_2 \neq B_2, A_1 \neq B_1,$ and $A_2 \neq B_2$, then $A_1 \cap B_2 = \phi, A_2 \cap B_1 = \phi, A_1 \cap B_1 = \phi$ and $A_2 \cap B_2 = \phi$, so the combination involves five different focal elements.

The three different cases above require different formulae to combine two triplet mass functions. Equations (6)-(10) give formulae for computing the case of $A_1 = B_1$ and $A_2 \neq B_2$.

$$m_1 \oplus m_2(A_1) = \frac{1}{N} (m_1(A_1)m_2(B_1) + m_1(A_1)m_2(C) + m_1(C)m_2(B_1))$$  \hspace{1cm} (6)

$$m_1 \oplus m_2(A_2) = \frac{1}{N} (m_1(A_2)m_2(C))$$  \hspace{1cm} (7)
\[
m_1 \oplus m_2(B_2) = \frac{1}{N} (m_1(C)m_2(B_2))
\]

\[
m_1 \oplus m_2(C) = \frac{1}{N} (m_1(C)m_2(C))
\]

where

\[
N = 1 - \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y) = 1 - m_1(A_1)m_2(B_2) - m_1(A_2)m_2(B_2)
\]

Let us consider two pieces of evidence obtained from two classifiers kNN (k-nearest neighbours) and SVM (Support Vector Machine), respectively, represented in XML below [8]:

**Output 1 (SVM):**

```xml
<DOCUMENT name="37928" category ="c_2">
  <CATEGORY bpa = "0.724">{c_1}</CATEGORY>
  <CATEGORY bpa = "0.184">{c_2}</CATEGORY>
  <CATEGORY bpa = "0.092">{c_1, c_2, c_3, c_4, c_5, c_6}</CATEGORY>
</DOCUMENT>
```

**Output 2 (kNN):**

```xml
<DOCUMENT name="37928" category ="c_2">
  <CATEGORY bpa = "0.688">{c_2}</CATEGORY>
  <CATEGORY bpa = "0.208">{c_4}</CATEGORY>
  <CATEGORY bpa = "0.104">{c_1, c_2, c_3, c_4, c_5, c_6}</CATEGORY>
</DOCUMENT>
```

**Fig. 1.** Outputs produced by kNN and SVM (c_1:comp.windows.x; c_2: comp.graphics; c_3:comp.sys.ibm.pc.hardware; c_4:comp.sys.mac.hardware; c_5: comp.os.ms-windows.misc; c_6: alt.atheism)

In this example, \( C = \{c_1, c_2, c_3, c_4, c_5, c_6\} \) is a frame of discernment, and we use triplets \( \langle A_1, A_2, C \rangle \) and \( \langle B_1, B_2, C \rangle \) to represent the outputs 1 and 2, i.e. \( \langle A_1, A_2, C \rangle = \langle \{c_1\}, \{c_2, c_3, c_4, c_5, c_6\} \rangle \) and \( \langle B_1, B_2, C \rangle = \langle \{c_1\}, \{c_2, c_3, c_4, c_5, c_6\} \rangle \), respectively. The corresponding mass functions on these propositions for document 37928 are shown in Figure 1. For example, the mass function given by SVM is \( m(\{c_1\}) = 0.724 \), \( m(\{c_2\}) = 0.184 \), and the ignorance \( m(\{c_1, c_2, c_3, c_4, c_5, c_6\}) = 0.092 \). By using Equations (6)-(10) to combine these results produced by the two classifiers, we can obtain the set of aggregated results below:

\[
m(A_1) = \frac{m_1(A_1)m_2(C)}{1 - (m_1(A_1)m_2(B_1) + m_1(A_1)m_2(B_2) + m_1(A_2)m_2(B_2))} = 0.24
\]

\[
m(A_2) = \frac{m_1(A_2)m_2(B_1) + m_1(A_2)m_2(C) + m_1(C)m_2(B_1)}{1 - (m_1(A_1)m_2(B_1) + m_1(A_1)m_2(B_2) + m_1(A_2)m_2(B_2))} = 0.67
\]
Since $A_1$, $A_2$, $B_2$ are singletons, the belief function is the same as the new mass function $m$. Therefore we have a set of strengths of belief with 3 possible categories as a combined result: \{\text{bel}(A_1), \text{bel}(A_2), \text{bel}(B_2)\}. By choosing the category with the maximum degree of belief as a final decision, we have $D(37928) = A_2 = c_2$. Thus the final decision made by the combined classifier is category $c_2$ – the decision made by the kNN classifier.

By repeatedly computing pairwise orthogonal sums in Equation (5), we combine all of the triplet mass functions.

5 Conclusion

In this work, we present a novel method and technique for representing outputs from different classifiers – a focal element triplet – and a method for combining multiple classifiers based on this new structure. The empirical comparison with the dichotomous structure proposed in [9] has been carried out, and the result shows our structure is better than the dichotomous structure in terms of both efficiency and accuracy. More theoretical justifications on the triplet structure and general formulae for evidential reasoning have also been provided in [7]. The structure, and the associated methods and techniques developed in this research is particularly useful for data analysis and decision making under uncertainty in the case of which knowledge and information are insufficient and incomplete.

References

1. Larkey, L.S. and Croft, W.B. (1996) Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 289-297.
2. Sebastiani, F., (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34 (1), 2002.
3. Li, Y. H. and Jain, A, K. (1998). Classification of Text Documents. The Computer Journal, Vol 41(8), pp537-546.
4. Yang, Y., Thomas Ault, Thomas Pierce. (2000). Combining multiple learning strategies for effective cross validation. The Seventeenth International Conference on Machine Learning (ICML’00), pp1167-1182.
5. Mitchell, T., (1997). Mitchell. Machine Learning. McGraw-Hill.
6. Shafer, G., (1976). A Mathematical Theory of Evidence, Princeton University Press, Princeton, New Jersey.
7. Bi, Y., (2004). Combining Multiple Classifiers for Text Categorization using Dempster’s rule of combination. Internal report.
8. Bi, Y., Bell, D. (2003). Specification of Dempster-Shafer’s Uncertainty Reasoning Engine. The ICONS deliverable 21.
9. Barnett, J. A. (1981). Computational methods for a mathematical theory of evidence. Proceedings of Seventh Joint Conference of Artificial Intelligence (IJCAI’81), pp 868-875.
Predicting the Relationship Between the Size of Training Sample and the Predictive Power of Classifiers

Natthaphan Boonyanunta and Panlop Zeephongsekul

Department of Mathematics and Statistics, RMIT University, Melbourne, Australia
natthaphanb@yahoo.com, panlopz@rmit.edu.au

Abstract. The main objective of this paper is to investigate the relationship between the size of training sample and the predictive power of well-known classification techniques. We first display this relationship using the results of some empirical studies and then propose a general mathematical model which can explain this relationship. Next, we validate this model on some real data sets and found that the model provides a good fit to the data. This model also allow a more objective determination of optimum training sample size in contrast to current training sample size selection approaches which tend to be ad hoc or subjective.

1 Introduction

In the area of machine learning and pattern recognition, much effort has been put into improving the predictive power of classification technique ([1],[2],[3],[4],[5]). To the best of our knowledge, none of these studies has attempted to objectively explain the relationship between the size of training sample and the predictive power of various classification techniques. In practice, it is widely accepted that the predictive power improves as more training samples are used. However, the improvement follows the law of diminishing return where the rate of improvement declines as more training samples are used. Based on this observed trend, a number of practitioners suggested an incremental approach whereby the size of training sample is increased up until the point where the predictive power of the model show no further improvement ([6],[7],[8]). Alternatively, some practitioners recommended using a very large training sample size to ensure that an acceptable level of predictive power is achieved. These recommendations are usually ad hoc or subjective, relying on experiences of practitioners, for instance, Berry & Linoff [9] suggested using a training sample of at least 10,000 for any common data mining problem. Given that current approaches for selecting the appropriate training sample size are subjective and that there is a trade-off between using more training data and increase in cost and time required in the model development process [7], these considerations have motivated us towards our main objective of this study which is to investigate the relationship between the size of training sample and the predictive power of a certain classification technique and then develop a general mathematical model which can best explain this relationship. Practitioners can use this model to determine the appropriate training sample size when applying a classification technique to a certain problem. After this introduction, we report on some empirical analyses of some real data sets which highlight the
relationship between training sample size and predictive power of a number of standard classification techniques. A mathematical model is then derived from this and validated in Section 3. Finally, in the last section, we conclude with some useful recommendations.

2 Empirical Analyses and Mathematical Model

In this section, we develop classifiers using various sizes of training sample on some real data sets, ie. Data Sets A, B and C. Data Set A contains application and performance information of 6,200 home loan borrowers. Half of the data set is the information on the borrowers who repay their loan in advance and the reminders are those who do not repay their loan in advance. For this data set, our objective is to predict the likelihood of the borrowers repaying their loan in advance by using the information from their loan application. Data Set B contains 7,600 observations. Half of the data set is the information of good credit risk borrowers and the reminders are the information of the poor credit risk borrowers. Our objective was to predict future performance of existing borrowers using their financial transaction information. This classification problem is widely known in banking industry as behavioural scoring [4]. For data set C, a sample data of 22,000 observations were taken from the ‘Adult’ data set from UCI Machine Learning Repository [13]. For this data set, the task is to predict whether a person earn over $50,000 per year by using the information from Census Bureau Database.

![Fig. 1. Predictive power of classification techniques versus training sample sizes using Data Set A](image)

In order to investigate the relationship between the training sample size and the predictive power of each classification technique, we partitioned each data set into 2 subsets: training and test sets. At each training sample size, ten re-samplings were made in order to reduce the impact of the sampling variation between training and test samples. The predictive power of a technique on this training sample size was then computed as the average percentage of the number of correctly classified borrowers over ten different test
data sets. For neural networks & classification tree analysis, cross-validation was undertaken to reduce the chance of overfitting by equally dividing the test data set into two subsets, the first subset was used as a validation data set and the other data set was used to test the performance of the model. Figures 1, 2 and 3 show the predictive power achieved (ie. average % of observations that are correctly classified over 10 test sets) by using different sizes of training sample on data set A, B and C. We note that the classification techniques used in this study are linear discriminant analysis (LDA), logistic regression analysis (LGT), backpropagation neural networks with 10 and 15 hidden neurons (NN10 and NN15), K nearest neighbour classifiers (with K = 1, 5, 10) and classification tree (TREE).

![Fig. 2. Predictive power of classification techniques versus training sample sizes using Data Set B](image1)

![Fig. 3. Predictive power of classification techniques versus training sample sizes using Data Set C](image2)

(Note: KNN refers to nearest neighbour classifier with k = 1; KNNi refers to nearest neighbour classifier with k = i; NNi refers to neural networks with i hidden neurons in the hidden layer)
Based on the empirical results shown in Figures 1, 2 and 3, it is observed that there is a distinct common pattern in the relationship between the size of training sample and the predictive power of a certain classification technique which is consistent with the relationship found by a number of practitioners (eg. [6],[7],[8]).

This relationship can be divided into three stages (refer to Figure 4). Initially, as the size of training sample increases, the rate of improvement in predictive power is fairly rapid, ie. from 0 to A, and then begins to decline, ie. from point A to B. In the last stage (from point B onward), no significant improvement in predictive power can be seen as the predictive power reaches a plateau which we will call its efficiency threshold. Based on this trend, we propose the following mathematical model relating predictive power to training sample size. But first let us define a few terms:

\[ P(p) = \text{predictive power at the training sample size } p; \]

\[ p = \text{training sample size}; \]

\[ T = \text{efficiency threshold}; \]

\[ k = \text{efficiency rate, ie. rate of improvement in predictive power per unit increase in efficiency.} \]

Based on our empirical analyses and previous discussion, the following relationship between \( P(p) \) and \( p \) is postulated:

\[
\frac{dP(p)}{dp} = k(T - P(p)) \quad (1)
\]

The justification of (1) follows directly from the previous discussion and illustration in Figure 4, ie. predictive power as a function of \( p \) improves very rapidly at first with fairly steep gradient, then the rate of improvement gradually decline as more training samples are used. Almost no improvement in predictive power occurs
after a certain training sample size, its optimum level, has been used for training (refer to the curve from point B onward in Figure 4). Solving the differential equation is straightforward resulting in

\[ P(p) = T(1 - e^{-kp}) + P(0)e^{-kp} \]  

(2)

The value \( P(0) \), i.e., predictive power when no training data is used to produce the classification rule, is related to the learnability of a classification technique ([11],[12]). For strong learners such as the techniques we have used in this study, i.e., where each technique is associated a polynomial time algorithm that achieves low error with high probability, it is expected that \( P(0) > 0.5 \), i.e., their performances are guaranteed to be significantly better than that obtained from a coin toss.

One of the advantages in unravelling the relationship between training sample size and predictive power of a classification technique is that it will enable us to identify the optimum training sample size for a classification technique for a certain classification problem. We refer here to the training sample size where predictive power of a classification technique is not significantly improved with further increase in training sample size beyond that optimum (in Figure 4, the optimum training sample size would be at around point B). Using a larger training sample size than B to fit a classification model would impose unnecessary wastage in both time and cost. On the other hand, using training sample size smaller than B will not produce a fitted model with the optimum level of efficacy.

3 Model Validation and Trend Prediction

In this section, we fit the proposed model given by (2) to our data sets using a non-linear optimization routine. We only utilize training sample sizes of 100, 500 and 1000 in the fitting process and use the resulting fitted model to forecast the predictive power for other training sample sizes. Table 1 below displays the observed predictive power (Technique-Actual) and the forecasted predictive power (Technique-Predict) on Data Set C. The mean absolute errors for each technique are provided in the last column of Table 1.

The low values of Mean Absolute Error for all techniques lend support to the hypothesis that the relationship between training sample size and predictive power indeed follows a precise mathematical law given by (2). Furthermore, it is noteworthy that the trend can be forecasted accurately using only small training samples in the fitting process (i.e., 100, 500 and 1,000 samples). We remark that the same analyses performed on the other two real data sets give similar results.

Based on the result in Table 1, we graphically display the actual and forecasted predictive power for some classification techniques in Figure 5. This figure confirms that there is a certain point in the relationship between training sample sizes and predictive powers where the rate of improvement beyond this point is very minimal
Table 1. Actual Predictive power Vs. Forecasted Predictive Power by Model (2) for each classification technique on Data Set C

| Techniques   | 140  | 420  | 700  | 1400 | 2100 | 2800 | 4200 | 5600 | 7000 | 8400 | 9800 |
|--------------|------|------|------|------|------|------|------|------|------|------|------|
| Knn-Actual   | 65.71% | 55.93% | 66.60% | 69.09% | 69.51% | 69.94% | 71.12% | 71.94% | 72.11% | 71.95% | 72.40% |
| Knn-Predict  | 64.21% | 55.6%  | 66.76% | 68.89% | 70.25% | 71.12% | 72.04% | 72.41% | 72.57% | 72.63% | 72.66% |
| Knn5-Actual  | 66.85% | 57.88% | 68.83% | 71.78% | 71.96% | 72.49% | 73.30% | 74.16% | 74.00% | 74.75% | 74.77% |
| Knn5-Predict | 65.47% | 57.51% | 69.09% | 71.65% | 73.00% | 73.71% | 74.29% | 74.45% | 74.50% | 74.51% | 74.51% |
| Knn15-Actual | 65.13% | 59.30% | 69.97% | 72.02% | 73.06% | 73.36% | 74.20% | 74.64% | 74.93% | 75.30% | 75.23% |
| Knn15-Predict| 64.41% | 57.82% | 69.98% | 72.51% | 73.32% | 73.58% | 73.69% | 73.70% | 73.71% | 73.71% | 0.82  |
| NN10-Actual  | 71.96% | 54.15% | 75.26% | 77.48% | 77.98% | 78.01% | 78.84% | 79.39% | 79.25% | 79.12% | 79.12% |
| NN10-Predict | 71.42% | 53.99% | 75.50% | 77.07% | 77.49% | 77.60% | 77.63% | 77.63% | 77.63% | 77.63% | 0.89  |
| NN15-Actual  | 71.24% | 53.92% | 75.27% | 77.36% | 78.64% | 78.54% | 78.79% | 79.27% | 79.24% | 79.50% | 79.55% |
| NN15-Predict | 71.36% | 53.99% | 75.59% | 77.39% | 77.92% | 78.07% | 78.14% | 78.14% | 78.14% | 78.14% | 78.14% |
| LGT-Actual   | 69.43% | 54.44% | 77.17% | 78.76% | 79.15% | 79.26% | 79.44% | 79.56% | 79.60% | 79.72% | 79.72% |
| LGT-Predict  | 67.54% | 75.52% | 77.64% | 78.38% | 78.40% | 78.40% | 78.40% | 78.40% | 78.40% | 78.40% | 78.40% |
| Tree-Actual  | 73.14% | 75.28% | 78.71% | 79.76% | 79.75% | 80.23% | 80.79% | 81.14% | 80.39% | 80.98% | 79.18% |
| Tree-Predict | 73.21% | 78.61% | 79.57% | 79.78% | 79.78% | 79.78% | 79.78% | 79.78% | 79.78% | 79.78% | 79.78% |
| LDA-Actual   | 71.67% | 75.81% | 76.28% | 77.46% | 77.88% | 78.01% | 78.13% | 78.20% | 78.27% | 78.32% | 78.64% |
| LDA-Predict  | 70.78% | 75.08% | 76.93% | 78.16% | 78.31% | 78.32% | 78.33% | 78.33% | 78.33% | 78.33% | 0.40  |

Fig. 5. Graphical plots of actual and predicted trend of TREE, KNN, KNN5 and NN15 on Data Set C

and can be considered as an insignificant improvement in predictive power. Since there is a degree of subjectivity in determining what represents an insignificant level of improvement, data storage, extraction costs and time required in developing the model should be taken into account when considering this fact. This can vary significantly between different practitioners. For example, some practitioners would tolerate using an additional 3000–4000 training samples in order to gain around 1% improvement in predictive power, while in some situations, obtaining additional 500 training samples may incur cost that is much higher than the benefit gaining from 1% improvement in the predictive power.
4 Conclusion

In this paper, we have undertaken an empirical study to investigate the relationship between the size of training sample and the predictive power of a number of classification techniques. Our aim was to develop a more objective approach in selecting the appropriate training sample size for certain classification problems. Beginning with the premise that size of training sample is one of the most important factors affecting the predictive power of a classification technique, we proceed to investigate how these two factors are related. Graphical plots of the results clearly indicate that as the training sample size increases, the predictive power improves until it reaches a plateau where further increase in the training sample size has no effect on the predictive power. It is important to note that the relationship between the predictive power of classification techniques and the size of training sample found in this study is also consistent with the trend found by other practitioners ([6],[7],[8]).

A nonlinear functional relationship, based on observing the rate of improvement in predictive power with respect to the change in training sample size, was proposed and then validated on some real data sets. The results show that this nonlinear model, which can be fitted by using relative small training sample sizes, can be used to predict the overall relationship trend between the predictive power and the training sample size for a certain classification technique. By observing this relationship trend obtaining from the fitted model (2) together with understanding the cost and benefit in obtaining additional training samples, practitioners can then determine the appropriate size of training sample for a certain classification problem (refer to point B in Figure 4).

References

1. Freund, Y.: An adaptive version of the boosting by majority algorithm. *Machine Learning* **43**(3) (2001) 293-318
2. Schapire, R.E.: Drifting games. *Machine Learning* **43**(3) (2001) 265-291
3. Ueda, N.: Optimal linear combination of neural networks for improving classification performance. *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(2) (2000) 207-215
4. Webb, G.I.: MultiBoosting: a technique for combining boosting and wagging. *Machine Learning* **40**(2) (2000) 159-196
5. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(3) (1998) 226-239
6. Weiss, S.M., Indurkhya, N.: *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers: California (1998)
7. Witten, I.H., Frank E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers: California (2000)
8. Groth, R.: *Data Mining: Building Competitive Advantage*. Prentice Hall: New Jersey (2000)
9. Berry, M.J.A., Linoff G.: *Data Mining Techniques For Marketing, Sales and Customer Support*. John Wiley & Sons: New York (1997)
10. Lewis, E.M.: *An Introduction to Credit Scoring*. Athena Press: California (1992)
11. Valiant, L.G.: A theory of the learnable. *Communications of the ACM* **27** (1984) 1134-1142
12. Schapire, R.E.: The strength of weak learnability. *Machine Learning* **5** (1990) 197-227
13. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases (1998)
Topographic Map Formation Employing kMER with Units Deletion Rule

Eiji Uchino, Noriaki Suetake, and Chuhei Ishigaki
Yamaguchi University, Yamaguchi 753-8512, Japan

Abstract. A kernel-based topographic map formation, kMER, was proposed by Van Hulle, and some effective learning rules have been proposed so far with many applications. However, no guide is indicated concerning the determination of the number of units in kMER. This paper describes a units deletion rule, which enables to construct automatically an appropriate-sized map to acquire the global topographic features underlying the input data. The effectiveness of the present rule has been confirmed by some preliminary computer simulations.

1 Introduction

In the study of self-organizing map (SOM) [1][2], kernel-based topographic map [3] was proposed in order to increase the description ability of the probability distribution of the input data, which has more biological knowledge as its background. This kernel-based topographic map has the lattice units with local kernel functions as shown in Fig. 1, e.g., Gaussian type, rather than introducing winner-take-all (WTA) scheme.

As a leaning rule for this map, kMER (kernel-based maximum entropy learning rule) was proposed by Van Hulle M. M.[4]. The kMER efficiently forms a map where all the units activate with an equal probability. This is successfully applied to clustering, estimation of the distribution of the input data, and so on. However, a guide is not provided how many units should be prepared beforehand for the given data, which makes it difficult to grasp globally the topographic features of the data when many units are used on the map.

This paper proposes a new learning rule for a kernel-based topographic map in order to cope with this problem. In the new rule, deletion of units is introduced in the conventional kMER, which enables to understand easily the global topographic features underlying the input data. The effectiveness of the proposed rule has been confirmed by computer simulations.

2 Kernel-Based Maximum Entropy Learning Rule: kMER

In this section, kMER[4] is briefly introduced. Now consider a lattice A, with a regular and fixed topology, with dimension $d_A$ in the $d$-dimensional space $V \subseteq \mathbb{R}^d$. Each unit $i (i = 1, \ldots, N)$ corresponds a kernel $K(v - w_i, \sigma_i)$, $v \in V$, with a radially-symmetric receptive field (RF), which has center $w_i$ and radius
Fig. 1. Kernel-based topographic map. Example of a $2 \times 2$ map with four units. Each unit has a Gaussian type kernel in this example. A circle is a receptive field of a kernel, called a range of a kernel, with its center on each unit.

Fig. 2. A receptive field (RF) kernel $K(v - w, \sigma)$ and a RF region $S_i$.

Fig. 3. Update of the receptive fields (RF). The direction of the arrow stands for the updating direction of the RF center $w_i$, given the input $v$. The dashed circles stand for the updated RF regions $S_i$ and $S_j$.

$\sigma_i$. A Gaussian function is often used for this kernel. The RF region $S_i$ is defined as shown in Fig. 2 by the the unit’s activation threshold $\tau_i$.

When the input $v$ falls on $S_i$, then the activation threshold will be increased, otherwise the activation threshold will be decreased. These events are formalized by introducing the following code membership function:

$$\xi_i(v) = \begin{cases} 1, & \text{if } v \in S_i, \\ 0, & \text{if } v \notin S_i. \end{cases}$$  \hspace{1cm} (1)

Since the RF regions usually overlap as shown in Fig. 3, a fuzzy code membership function is defined as follows:

$$\Xi_i(v) = \frac{\xi_i(v)}{\sum_{j \in A} \xi_j(v)}, \quad i, j \in A,$$  \hspace{1cm} (2)
to get $0 \leq \Xi_i(v) \leq 1$ and $\sum_i \Xi_i = 1$. Depending on the activation state of units, the parameters $w_i$ and radii $\sigma_i$ are adapted by using the following two learning rules.

In “batch” mode, for a set of $M$ input data samples, the RF centers $w_i$ are updated in proportion to $\Xi_i$ in the direction of $v$ as shown in Fig. 3:

$$\Delta w_i = \eta \sum_{v^\mu \in M} \sum_{j \in A} A(i, j, \sigma_A(t)) \Xi_i(v^\mu) \text{Sgn}(v^\mu - w_i), \forall i \in A,$$

where $\text{Sgn}(\cdot)$ is a sign function applied component wise for each neighborhood range, $\sigma_A(t)$ (in lattice space coordinates), of a neighborhood function $A(\cdot)$. $t$ is present time step, and $\eta$ is a learning rate.

The kernel radii $\sigma_i$ are updated as follows so that the activation probability for unit $i$ converges to $P(\xi_i(v) \neq 0) = \rho/N, \forall i$, with a constant scale factor $\rho$:

$$\Delta \sigma_i = \eta \sum_{v^\mu \in M} \{\rho_r/N(1 - \xi_i(v^\mu)) - \xi_i(v^\mu)\}, \forall i \in A \tag{4}$$

with $\rho_r = \rho N/(N - \rho)$. For the mathematical details including a proof of convergence in kMER, refer to [3,4].

### 3 kMER with Units Deletion Rule

In this section, kMER with units deletion rule is discussed. In the units deletion rule, overlapping degree $m_i(j)$ between units $i$ and $j$ are calculated at first. The overlapping degree is defined as follows:

$$m_i(j) = \begin{cases} 0, & l(i, j) > 2\sigma_i, \\ 1, & l(i, j) < 0, \\ l(i, j)/2\sigma_i, & \text{otherwise} \end{cases} \tag{5}$$

with

$$l(i, j) = \sigma_i + \sigma_j - ||w_i - w_j||. \tag{6}$$

As shown in Fig. 4, $m_i(j)$ approximately shows the value which is proportional to the overlapping area between $S_i$ and $S_j$. After calculation of $m_i(j)$, it
is determined whether to delete the unit $i$ according to $m_i(j)$. That is, if $m_i(j)$ is greater than the threshold $\theta$, then the unit $i$ is deleted from the map.

In kMER, the probability, to which the unit is activated, converges to $P(\xi_i(v) \neq 0) = \rho/N$. In our case, we use the joint activating probability of units $i$ and $j$ as an index to measure the overlapping between $S_i$ and $S_j$. The threshold, $\theta$ is given by:

$$\theta = \left( \frac{\rho}{N} \right)^2.$$  

Further, in the present rule, topology-preserving procedure is taken after a deletion of unit $i$ as shown in Fig. 5. In this procedure, unit $j$ inherits from unit $i$ the neighborhood relationship on a competitive layer as follows:

$$C_j \leftarrow k, \{k | k \in C_i, k \neq j\},$$

where $C_j$ is a set of units in the neighborhood of unit $j$. And the position of unit $j$, $r_j \in R$, is updated as:

$$r_j \leftarrow \frac{r_j + r_i}{2}.$$  

The present units deletion rule is added to the conventional kMER, and the kernel-based topographic map is obtained, which is automatically tuned to have an appropriate size of map to acquire the global topographic features of the input data.

4 Simulation Results

The effectiveness of the present units deletion rule is verified by some preliminary computer simulations.

First, kMER with units deletion rule is applied to the input data shown in Fig. 6(a). The input space consists of two-dimensional 600 vectors ($M = 600$), and has a probability distribution on $[0, 1]^2$ as shown in Fig. 6(b).

Figs. 7 and 8 show the evolutional change in time of the RF (receptive field) regions of the kernel-based topographic map obtained by the respective learning rule in “batch” mode. In both kMER and kMER with units deletion rule, the following neighborhood function $A(i, j, \sigma_A(t))$ was employed:

$$A(i, j, \sigma_A(t)) = \exp \left( -\frac{||r_i - r_j||^2}{2\sigma_A(t)} \right)$$.
Fig. 6. Input data. (a) Distribution of the input data. (b) Probability distribution. “pdf” stands for the probability density function.

Fig. 7. Evolutional change in learning of the receptive field (RF) regions on a $10 \times 10$ lattice space by using the “batch” version of the kMER. (a) 1,000 steps. (b) 3,000 steps. (c) 6,000 steps. (d) 9,000 steps. (e) 21,000 steps. (f) 30,000 steps. The dots in the figure show the input sample points, and the circles show the receptive fields of the kernels with

$$\sigma_A(t) = \sigma_{A0} \exp\left(-2\sigma_{A0} \frac{t}{t_{\text{max}}}ight).$$

(11)
Fig. 8. Evolutional change in learning of the receptive field (RF) regions on a $10 \times 10$ lattice space by using the “batch” version of the kMER with the proposed units deletion rule. (a) 1,000 steps. (b) 3,000 steps. (c) 6,000 steps. (d) 9,000 steps. (e) 21,000 steps. (f) 30,000 steps. The dots in the figure show the input sample points, and the circles show the receptive fields of the kernels.

Fig. 9. Input data. (a) Distribution of the input data. (b) Probability distribution. “pdf” stands for the probability density function.
Fig. 10. The resulting receptive field (RF) regions on a $10 \times 10$ lattice space obtained after 30,000 steps of learning. (a) The results by the kMER. (b) The results by the kMER with the proposed units deletion rule. The dots in the figure show the input sample points, and the circles show the receptive fields of the kernels.

The common parameters for the two learning rules are $N = 100(10 \times 10)$ as an initial value, $\eta = 0.0005$, $\sigma_{A0} = 5$, $t_{\text{max}} = 30,000$ and $\rho = \max(1, 30N/M)[3]$. The present units deletion rule is used every 3,000 learning steps of the conventional kMER method. From Figs. 7 and 8, it is observed that the kMER with present units deletion rule makes a map with an appropriate size to catch the global topographic features underlying the input data.

The kMER with units deletion rule is also applied to the input data shown in Fig. 9(a). The input space consists of two-dimensional 1,000 vectors ($M = 1,000$) and has a probability distribution on $[0, 1]^2$ as shown in Fig. 9(b). Other parameters are the same as the simulation for the input data in Fig. 6.

Fig. 10 shows the resulting RF regions obtained after 30,000 steps of the respective learning. In this figure, it is seen that the kMER with units deletion rule grasps roughly the global topographic features of the input data.

5 Conclusions

In this paper, we have proposed the units deletion rule for the conventional kMER. The proposed rule enables to acquire the global topographic features of the input data with only a small number of units.

The validity of the proposed rule has been confirmed by computer simulations. The simulation results are promising for a possible practical application to such as clustering problems, etc.

Future study is to analyze in detail the relationship between the threshold parameter $\theta$ of Eq. (7) and the performance of the proposed rule.

References

[1] Kohonen T. “Self-organized Formation of Topologically Correct Feature Maps,” Biol. Cybern., Vol.43, pp.59-63, 1982.
[2] Kohonen T. *Self-organizing maps*, Heidelberg, Springer, 1995.

[3] Van Hulle M. M., *Faithful Representations and Topographic Maps: From Distortion- to Information-based Self-organization*, John Wiley & Sons, Inc., 2000.

[4] Van Hulle M. M., “Kernel-based Equiprobabilistic Topographic Map Formation,” *Neural Computation*, Vol.10, pp.1847-1871, 1998.
Study on Weld Quality Control of Resistance Spot Welding Using a Neuro-Fuzzy Algorithm

Yansong Zhang, Guanlong Chen, and Zhongqin Lin

School of Mechanical Engineering,
Shanghai Jiaotong University, Shanghai, 200030, China
{zhangyansong, glchen, zqlin}@sjtu.edu.cn

Abstract. Resistance spot welding (RSW) is widely utilized as a joining technique for automobile industry. However, good weld quality evaluation method has not yet been developed in plant environment. It is necessary to achieve real-time inspection of RSW. This paper proposed a neuro-fuzzy algorithm to predict weld quality online. An experimental system was developed to measure electrode displacement curve. Accordingly based on electrode displacement curve nugget diameter will be inferred. Inference results showed that proposed neuro-fuzzy algorithm is suitable as a weld quality monitoring for resistance spot welding.

1 Introduction

Resistance spot welding (RSW) is widely utilized as a joining technique for automobile structure due to flexibility, robustness and high-speed of process combining with very high quality joints at very low cost. In the case of the automobile, the welding process is performed on thousands of spots in order to complete the body of a single car. The quality of the RSW welds determines the durability and corrosion resistance of the automobile body. Therefore, the quality of the welds is very important to the strength and quality of the automobile. However, an effective quality evaluation method has not yet been developed. [1] Current practice of evaluating weld quality in industry is still destructive testing. A small number of vehicle structures have to be chiseled apart every day to evaluate the weld quality in automotive assembly plants, the confidence level on weld quality is considerable low. To reduce the risk of part failures, manufacturers often require that more welds be performed than would be needed if each one were reliably made.

It is difficult to determine the quality of RSW with the naked eye, so non-destructive testing (NDT), which uses ultrasonic waves or X-rays, is useful to examine the quality and defects of the weld. However, applying NDT to in-process manufacturing causes many problems in terms of cost and facilities, especially in real-time quality inspection. In an assembly process, it is desirable to estimate the weld quality right after each weld has been made such that remedial action can be taken when defective welds have been detected. Studies have shown that the strength of a weld is correlated to the size of the weld nugget. Therefore, the nugget size (especially nugget diameter) has been widely used as a quality measurement in the industry.
Electrode displacement, which gives good indication of thermal expansion, melting, and expulsion, has proven to be a particularly useful signal to monitor. It is believed that the amount of thermal expansion, melting, and expulsion can be corrected to the slope and magnitude of the displacement curve. A number of control systems have been developed based on maximum electrode displacement or its changing rates. [2,3,4]

In order to handle effectively the non-linear dynamic effect, which occurs within a very short time during RSW, and to explain the functional relationship between the process variables and welds quality, attempts are being made to apply fuzzy algorithm, or artificial neural networks, to RSW. However, these algorithms are not industrially applicable because it requires too many input variables and fuzzy rules. [5,6,7]

This paper describes the real-time development of spot weld nugget diameter using an electrode displacement signal based on neuro-fuzzy algorithm. Electrode displacement and electrode velocity are regards as input parameters of neuro-fuzzy inference system. The adaptive neuro-fuzzy inference system is applied to the measurement of nugget diameter. This inference system is then verified by a comparison between the inferred and real nugget diameter.

2 Electrode Displacement Characteristics

Because of the current flow, heat will be generated and temperatures in the system will start to increase. Electrode displacement will rise due to thermal expansion caused by heat. It can be concluded that electrode displacement indicates the relative movement of the electrodes during the welding process. In this paper, the trace of relative electrode movement was measured by an OMRON laser displacement sensor. The displacement signal was fed into a low-pass filter for decreasing noise. Then the signal conditioning unit scales the signals to suitable voltage levels via an A/D converter in the computer. The data acquisition software used is National Instrument’s Labview. A simplified experimental data acquisition system is shown in fig. 1. The measured signals of a whole welding cycle for a typical weld is shown in fig. 2(a).

From the measured electrode displacement curves we could see that the electrodes will approach due to the electrode force, then the electrodes separate at a constant velocity because of thermal expansion of the sheet. Maximum expansion occurs when electric current is cut off. Finally the electrode displacement starts to fall owing to cooling. This curve has been verified to reflect the physical phenomenon occurring during the weld formation [3]. When expulsion occurs, electrode displacement signals will show abrupt changes in fig. 2(b). Experimental results show that electrode displacement and electrode velocity values of the curve are well corrected with nugget size. Based on the results we can draw a conclusion that a desirable nugget size can be obtained by monitoring electrode displacement curve.

The above experiment results showed that electrode displacement curve could reflect the nugget formation during RSW. And Electrode displacement and electrode velocity not only can reflect growth of a spot weld nugget but also are two measurable output parameters based on electrode displacement curve. Thus, electrode displacement and electrode velocity values were selected as fuzzy input variables for a neuro-fuzzy inference system under the non-expulsion condition [8]. When
expulsion occurs, weld schedules were adjusted to meet welding quality demand according to electrode displacement signal.

**Fig. 1.** A data acquisition system of electrode displacement curve

**Fig. 2.** (a) electrode displacement curve (no expulsion) (b) electrode displacement curve (with expulsion)

### 3 Neuro-Fuzzy Inference System

The neuro-fuzzy modeling has been used as a powerful tool which can facilitate the effective development of models. The combined use of the learning ability of neural networks and the representation ability of fuzzy systems can partially overcome vague
and imprecise data related to a fuzzy system. The approach is especially useful for large complex and nonlinear systems, which cannot be represented reasonably as simple and unique. Thus, the approach is ideally suited to investigate the complex spot welding control problems.

Neuro-fuzzy models describe systems by means of fuzzy if-then rules represented in a network structure; to which learning algorithms known from the area of ANN can be applied. They provide new directions in the application of on-line measurement to spot welding systems.

This paper proposed the neuro-fuzzy inference system with two input variables (electrode displacement and electrode velocity) and one output variable (nugget diameter). The neuro-fuzzy scheme is shown in Fig. 3. Firstly, the two inputs are codified into linguistic values by the set of Gaussian membership functions. The next step will calculate its respective activation degree to each rule. Lastly, the inference mechanism weights each conclusion value. The error signal between the inferred output value and the respective desired value is used by the gradient-descent method to adjust each rule conclusion.

The fuzzy inference system consists of three main blocks: membership functions selection, fuzzy rules, and conclusion value output. The following subsections represent the neural structure which is proposed here to map the fuzzy inference to ANN. This neuro-fuzzy scheme consists of three layers.

### 3.1 Membership Functions Selection

The first layer is composed of neurons with Gaussian activity functions which are determined by the centers $c_j$ and the variances $\sigma_j^2$. Membership functions denoted
by $\mu_{A_j}(x_i)$ as we expressed in Equation (1). This layer performs the fuzzification of crisp network input values in that neuron.

$$\mu_{A_j}(x_i) = a_{ij} e^{-\left(\frac{x_i - c_{ij}}{\sigma_{ij}}\right)^2}$$  \hspace{1cm} (1)

### 3.2 Fuzzy Rules

The second layer represents the rule layer in which the logical operators are implemented and the antecedent’s possibilities are aggregated. The most common neuro-fuzzy network is used to develop or adjust a fuzzy model in Mamdani form. A Mamdani fuzzy model consists of a set of fuzzy if-then rules in the following form:

$$R: \text{IF } (x_1 \text{ is } A_1 \text{ and } x_2 \text{ is } A_2 \text{ and } \ldots x_m \text{ is } A_m) \text{ THEN } (y \text{ is } B);$$

Each if-then rule defines a fuzzy implication between antecedent and consequent. The reasoning process combines all rule contributions using the defuzzification formula in a weighted form.

### 3.3 Conclusion Value Output

The third layer performs the defuzzifications to achieve a crisp value of the variable. The output of the inference process so far is a fuzzy set, specifying a possibility distribution of control action. In the on-line control, a nonfuzzy (crisp) control action is usually required. This paper used defuzzification operator is center of area (COA). It generates the center of gravity of the possibility distribution of the inferred fuzzy output.

### 3.4 The Learning Mechanism

At the computational level, a fuzzy system can be seen as a layered network structure, similar to artificial neural networks of the RBF-type. In order to optimize parameters in a fuzzy system, gradient-descent training algorithms known from the area of neural networks can be applied [9].

The gradient-descent algorithm changes the conclusion values to minimize an objective function $E$ usually expressed by equation (2). By changing the learning rate parameter and number of learning iterations executed by the algorithm each conclusion value was adjusted.

$$E = \frac{1}{2} \left[ Y(x(k)) - y^\prime(k) \right]^2$$  \hspace{1cm} (2)

where the value $y^\prime(k)$ is the desired output value and $Y(x(k))$ is the inferred output value.
4 Result

Fig. 4 shows a comparison result between the inferred nugget diameter and the real nugget diameter when no expulsion happens. The analysis result showed that the linear correlation coefficient was 0.9653. The standard nugget diameter of common low-carbon steel is 5.5mm with a 1.5mm thickness. Fig. 5 shows the relationship
between the inference error and real nugget diameter. Among the total number of specimens, 88% were successfully inferred within a range of 1.5% error.

5 Conclusion

The neuro-fuzzy modeling and the learning mechanism to inference nugget diameter in resistance spot welding were investigated. Inference results showed that proposed neuro-fuzzy algorithm is suitable as a weld quality monitoring for resistance spot welding. We believe that emerging technologies as neuro-fuzzy systems have to be used together with genetic algorithms to produce more intelligent weld quality control systems.

References

1. Stiebel, A., Ulmer, C., Kodrack, D., and Holmes, B.B., Monitoring and control of spot weld operations, 860579, SAE, Feb., 1986.
2. Tsai, C.L., Dai, W.L., Dickinson, D.W., and Papiritan, J.C., Analysis and development of a real-time control methodology in resistance spot welding, Welding Research Supplement, 1991, 12: 339-351.
3. Cho, H.S., and Chun, D.W., A microprocessor-based electrode movement controller for spot weld quality assurance. IEEE Transactions on Industrial Electronics, 1985, 32(3):234-238.
4. Chang, H.S., Cho, Y.J., Choi, S.G., and Cho, H.S., A proportional-integral controller for resistance spot welding using nugget expansion, ASME Journal of Dynamic Systems, Measurement, and Control, 1989, 111: 332-336.
5. Khoo, L.P. and Young, H.Y. A prototype fuzzy resistance spot welding system. International Journal of Production Research. 1995, 33(7), 2023-2036.
6. Dilthey, U. and Dickersbach, J. Application of neural networks for quality evaluation for resistance spot welds. ISIJ International, 1999, 39(10), 1061-1066.
7. S.R.Lee and Y.J.Choo., A quality assurance technique for resistance spot welding using a neuro-fuzzy algorithm, Journal of manufacturing systems, 2001: 320-328.
8. Robert W. Messler, Jr., Min Jou, An intelligent control system for resistance spot welding using a neural network and fuzzy logic, Conference Record IAS Annual Meeting, 1995:1757~1763.
9. Jang, J.-S.R. (). ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man & Cybernetics, 1993, 23(3), 665–685.
Exploring Benefits of Neuro Fuzzy Controller with
Vehicle Health Monitoring

Preeti Bajaj¹ and Avinash Keskar²

¹Assistant Professor, Electronics Department, G. H. Raisoni College of Engineering,
CRPF Gate 3, Digdoh Hills, Hingna Road, Nagpur, India- 440016,
Phone (R: 91-712-2562782, M: 91-9822237613)

²Professor, Department of Electronics and Computer Sciences, Visvesvarya National
Institute of Technology, Nagpur-440011,
Phone (R: 91-712-2225060, M: 91-9823047697)
{preetib123, avinashkeskar}@yahoo.com

Abstract. This paper presents the architecture and learning procedure-underlying Adaptive- Neural Fuzzy Inference System, a Fuzzy Inference System implemented in the framework of adaptive neural networks. It elaborates the work to employ the ANFIS architecture to model linear function for creating a closed loop system. It tests the comparative performance with simple and hybrid controllers with respect to ‘Pre-trip plan assistance system with vehicle health monitoring’. The system calculates safest traveling distance based on six inputs. The approach is an extension to design of pre trip plan assistance with open loop fuzzy and hybrid fuzzy using genetics [1-3]. This paper mainly proposes performance of ANFIS for calculating safest distance that can be traveled by a vehicle under pre trip plan assistance system with vehicle health monitoring. ANFIS is used in the current work to create a closed loop system and to test comparative performance with simple and hybrid controllers. Nearly 5-8 % improvement in the results is obtained if we incorporate hybrid (Mamdani Sugeno) model and neural fuzzy closed loop controller gives further 4% improvement. This system will be helpful for the travelers where long journey is to be commenced and where cost of traveling matters, and frequent variation in weather dominates success of the journey.

1 Introduction

Traveler information plays a critical role in supporting safety, security, mobility, and in improving the reliability of travel. The study of Vehicle Health monitoring plays very important role in deciding the possibility of completing journey successfully. Neural approach has recently been demonstrated to be a vital approach to learning fuzzy logic membership functions for transportation. Already work has been done on Pre Trip plan assistance with simple Fuzzy and Hybrid Fuzzy systems[1-3]. In general for Fuzzy Logic Controllers Mamdani or Sugeno model are employed depending upon its suitability to the problem. Since both have some restrictions, an attempt has been made in this paper to make comparison of best of both the world by merging...
special membership function shapes of the both the models in open loop that with closed loop by incorporating adaptive neural fuzzy inference system.

In the present work, a Neural-Fuzzy system is proposed; to promote the learning performance of logic rules that can extract information from multiple fuzzy knowledge source and put into a single knowledge base. The work elaborates a fuzzy system, in which the membership function shapes, types are evolved using a Neural Network. The neurons are trained by a fuzzy system. The Adaptive Neural Fuzzy Inference System (ANFIS) calculates safest distance that can be traveled by a vehicle under pre trip plan assistance system with vehicle health monitoring by creating a closed loop system and to test comparative performance with simple and hybrid controllers. Its property of changing the parameters of the membership functions serves as the feedback link. This system will be helpful for the travelers where long journey is to be commenced and also can be used where cost of traveling matters, and frequent variation in weather dominates success of the journey. The benefits of the methodology are illustrated in the process of calculating safest Distance Traveled by a vehicle depending upon status of various inputs.

2 Vehicle Health and Vehicle Maintenance

The success of completing a desired journey depends on various factors, which can be categorized as journey classification, journey type, journey condition, vehicle condition, vehicle maintenance, vehicle characteristics and the proficiency of the driver. The vehicle health is the current condition of the vehicle based on sensor output about the Air pressure in tyres, Coolant level, Engine oil level, Break oil level, Battery condition, Fuel condition, Tyre age, Last servicing report, part replacement, age of the car etc. The detail history of servicing and life of different parts of vehicle is maintained as per the characteristics of the vehicle. Since all these factors are measured by sensors and vary in specifying the minimum requirement with the make, type of journey, load etc, Fuzzy controller is chosen to judge the Vehicle Health and to calculate the safest distance to be traveled by a vehicle with current sensors output. The adaptive ness of the Controller is tested with six dominating inputs as coolant, engine oil; break oil and fuel, visibility and air pressure.

3 Pre Trip Plan Assistance Using ANFIS

The design and the performance of the ANFIS controller for Pre Trip Plan Assistance system with vehicle Health Monitoring is evaluated over 6 dominated inputs. These inputs and their ranges are given in Table 1. All these factors play an important role in deciding the optimum distance that can be traveled. These factors generally affect driver’s pre-trip and reroute choice decision. Fig 1 shows open loop Sugeno Fuzzy system with six inputs and one output as safest distance to be traveled by a vehicle.

The sample of input and output membership function for Hybrid fuzzy (sugino-mamdani) is as shown in Figure 2 & 3 respectively.
Table 1. Input & Output membership mapping in Adaptive fuzzy

| In/Out labels     | Membership function and levels                  | In range |
|-------------------|-------------------------------------------------|----------|
| Coolant           | Trapezoidal, Dangerous, safe, almost full        | 0, 500ml |
| Engine oil        | Trapezoidal, Dangerous, Safe, Almost Full        | 0, 500ml |
| Break oil         | Triangular, Dangerous, safe, almost full         | 0, 500ml |
| Fuel              | Trapezoidal, Less, Moderate, Full                | 0, 40    |
| Air pressure      | Trapezoidal, Less, Moderate, Full                | 0, 30    |
| Visibility        | Trapezoidal, Less, Moderate, Full                | 0, 200m  |
| Safest traveling distance -Km | Trapezoidal, Short, moderate, large | 0, 300km |

Fig. 1. Open Loop Sugeno System For Pre Trip Plan Assistance System With Vehicle Health Monitoring

Training has been given to the FIS using the training data. These inputs are fuzzified and applied to the fuzzifier. The Inference engine fires appropriate rules and gives defuzzified output for safest distance to be traveled. All these factors play an important role in deciding the optimum distance that can be traveled. These factors will also affect driver’s pre-trip and reroute choice decision.

Sample of Input Rules for Open Loop Sugino Fuzzy Inference System: If air pressure is not perfect or coolant is very low or break oil is not full or fuel is 10 or engine oil is 10 or visibility is low then the distance covered is 0(1).
5 Execution of Adaptive Neural Fuzzy System

A Fuzzy Inference System is designed with six inputs and then a model is trained to get the desired output using ANFIS. The training data is generated using a MATLAB program (M-file), which accepts the data from a variable ‘g’ from the workspace. By training the set of inputs [4-5], error can be reduced & better accuracy can be achieved. Membership functions form the part and parcel of the Fuzzy Inference Systems. The closed loop nature of the system is because of the property of the alternating of membership function of both inputs and output by ANFIS.

**Fig. 2. Hybrid Membership Function for fuel**

**Fig. 3. Membership Function Safest traveling distance**
ANFIS models the FIS to best suit the training data available. This modeling is done by alteration of the Membership function parameter of inputs and output. Altering of the parameters is done by ANFIS using back propagation. The simplest implementation of back propagation learning updates the network weights and biases in the direction in which the performance function decreases most rapidly - the negative of the gradient. The gradient descent algorithm can be implemented in batch mode, where the weights and biases of the network are updated only after the entire training set has been applied to the network. The gradients calculated at each training example are added together to determine the change in the weights and biases. The program is executed with the 10 different samples as given in the table 2.

**Table 2. Sample of Input and output for ANFIS**

| Air pressure | Coolant (ml) | Brake Oil (ml) | Fuel (lit) | Engine Oil (ml) | Visibility (m) | Open loop output | Closed loop output (m) |
|--------------|--------------|----------------|------------|-----------------|----------------|------------------|------------------------|
| 24           | 250          | 250            | 10         | 250             | 150            | 75               | 82                     |
| 27           | 200          | 200            | 20         | 200             | 20             | 89               | 99                     |
| 23           | 260          | 280            | 35         | 250             | 40             | 130              | 145                    |
| 21           | 320          | 320            | 26         | 350             | 70             | 169              | 180                    |
| 25           | 300          | 375            | 15         | 300             | 50             | 207              | 227                    |
| 26           | 350          | 375            | 25         | 400             | 25             | 340              | 375                    |
| 22           | 400          | 420            | 30         | 450             | 30             | 360              | 387                    |
| 25           | 330          | 400            | 37         | 160             | 90             | 363              | 388                    |
| 24           | 370          | 430            | 38         | 240             | 150            | 376              | 389                    |
| 26           | 450          | 450            | 39         | 320             | 180            | 387              | 396                    |
Fig. 4. Sample of open loop and closed loop Membership function

Fig. 4 shows the features of the ANFIS, which modifies the input membership functions of the open loop fuzzy systems and provides a close link so as to become system as closed loop feedback system that gives more stable output. The structure [6] for Pre trip plan assistance system with six inputs and one output is as shown in fig 5.

Fig. 5. Neural Network Equivalent structure
6 Results and Conclusion

By using a hybrid learning procedure, the proposed ANFIS can identify the structure of input-output membership mapping based on human knowledge (in the form of Fuzzy if-then rules). In the simulation, an attempt is made to employ the ANFIS architecture to model linear function with respect to ‘Pre-trip plan assistance system’. The proposed approach allows improving the system performance approximately by 7-10% as shown in Fig 6.

This paper has demonstrated an approach to Vehicle health monitoring and hence to determine the safest distance to be traveled by a vehicle with current health. ANFIS was applied to discover the fuzzy controllers capable of determining the safest distance allowed by the vehicle. ANFIS evolved a good and reasonable membership function set of rules for an FLC that demonstrated satisfactory responsiveness to various initial conditions while utilizing minimal human interface. Such systems strike a balance between learned responsiveness and explicit human knowledge makes the system very robust, extensible and suitable for solving a variety of problems. These results now will be used for Pre trip plan assistance system with soft computing tools.

![Comparison of Sugino Hybrid & ANFIS Model](image)

**Fig. 6.** Comparative Performance of Hybrid and Simple Fuzzy Controller Showing Safest Distance To be Traveled

References

1. Ms Bajaj, Dr. A.G. Keskar, “Smart Pre trip Plan assistance system using Vehicle health monitoring”, ITSC-02, fifth international conference on Intelligent transportation systems, IEEE, 3-6th sept 2002, Singapore.
2. Ms Bajaj and Dr. Keskar, “Smart Pre Trip Plan Assistance System With hybrid Fuzzy Algorithm”, Hong Kong Transportation Association
3. Ms Bajaj, Dr. Keskar, “Soft computing based adaptive approach for vehicle health monitoring”, KES 2003, seventh International conference on Knowledge Based Intelligent Information and Engineering Systems, university of Oxford, UK
4. Fuzzy Toolbox, Matlab user’s Guide
5. Artificial Neural Networks By B.Yegnanarayana
6. Introduction to Artificial Neural System By Jacek M. Zurada
Improvement of Low Frequency Oscillation Damping in Power Systems Via an Adaptive Critic Based NeuroFuzzy Controller

Farzan Rashidi and Behzad Moshiri
Control and Intelligent Processing Center of Excellence, Department of Electrical and Computer Engineering, University of Tehran, Iran
f.rashidi@ece.ut.ac.ir, moshiri@ut.ac.ir

Abstract. This paper proposes an Adaptive Critic Based Neuro-Fuzzy Controller (ACBNFC) to Thyristor Controlled Series Capacitor (TCSC), which might have a significant impact on power system dynamics. The function of the ACBNFC is to control a firing angle of the TCSC. The proposed method is used for damping the low frequency oscillations caused by disturbances such as a sudden change of small or large loads or an outage in the generators or transmission lines. To evaluate the usefulness of the proposed method, the computer simulation for single machine infinite system is performed and compared with fuzzy PD controller. Obtained results show the proposed control strategy is very robust, flexible and could be used to get the desired performance levels. The response time is also very fast. Simulation results that have been compared with fuzzy PD controller show that our method has the better control performance than fuzzy PD controller.

1 Introduction

Series capacitive compensation in AC transmission systems can yield several benefits, such as increased power transfer capability and enhanced transient stability. Thyristor controlled series capacitors (TCSC) are beginning to find applications as adjustable series capacitive compensators, as they provide a continuously variable capacitance by controlling the firing angle delay of a thyristor controlled reactor (TCR) connected in parallel with a fixed capacitor. Besides controlling the power flow, TCSCs have a potential to provide other benefits, such as transient stability improvement, damping power swing oscillations, mitigating subsynchronous resonance (SSR) and fault current reduction. Hence, effective firing control strategies are required to exploit all advantages that a TCSC installation might offer. Several different control and intelligent strategies have been developed in recent years to achieve the stated goals fully or partially [1-4]. Though they show a good controller performance in a specific operating point because they are designed using the linearized system, it is difficult to obtain a good controller performance in a different operating condition. In particular, because the dynamic characteristic of power system with the reactive power compensator has a strong nonlinearity, the controller designed based on linear control can not show an optimal control performance. The purpose of this paper is to suggest another control strategy, based on adaptive neuro-fuzzy controller [5], for damping the low frequency oscillations caused by disturbances such as a sudden...
change of small or large loads or an outage in the generators or transmission lines. Simulation results show that, the proposed method is very robust and the response time can achieve satisfactory performance. To evaluate the usefulness of ACBNFC, we perform the computer simulation for a single machine infinite system. We compare the response of this method with fuzzy PD controller. Obtain results show that the performance of ACBNFC is better than fuzzy PD controller. In the subsequent sections, we discuss the mathematical model of power system, our proposed controller, and its application in the closed loop control system, simulation and some concluding remarks.

2 Mathematical Model of Generator and TCSC

The differential equations of a single-Machine Infinite System are expressed in equations (1)-(3), which is for designing TCSC controller to maximize the usage rate of power transmission facilities and power increase of delivery capacity. If a rapid response typed exciter is used, we can model the generator sufficiently using only automatic voltage regulator (AVR) removing the exciter as shown in (4). The turbine/regulator characteristic of the synchronous machine is not considered because of its long time constant, relatively slight variation.

\[
\frac{dE_q'}{dt} = -\frac{1}{T_{do}}[E_q' + (X_d - X_d')I_d - E_{fd}] \\
\frac{d\delta}{dt} = \omega - \omega_{ref} \\
\frac{d\omega}{dt} = \frac{\omega_{ref}}{2H}[T_m - E_q I_q - (X_q - X_d')I_d I_q] \\
\frac{dE_{fd}}{dt} = \frac{K_a}{T_a}(V_{ref} - V_t + V_s) - \frac{1}{T_a}E_{fd}
\]

The conventional series compensator using a breaker is restricted to the usage frequency owing to abrasion and can not compensate dynamically because its compensator speed is slow and include an abnormal oscillation such as subsynchronous resonance (SSR). But TCSC can control promptly and precisely by using a high speed switching thyristor and be operated all time with not restricted to usage frequency and contributes to the improvement of the transient stability. TCSC function as a fixed capacitor in the steady state, so that it can control the power flow and improve the steady state stability by increasing voltage stability margin. It can increase the dynamic stability of power system by controlling a capacity value in disturbances and protect devices from over voltage and/or over current by bypassing the capacity with adequate protection device in fault and reinstall the capacity promptly in fault restoration. The line reactance in a conventional PSS analysis model is a fixed, constant value, but the line reactance in the model including TCSC can no longer be considered as a fixed value because of its variation. So, in this paper, we use
the modulation control changing the reactance of TCSC continuously by a firing angle control. The fundamental wave of TCSC reactance is (5).

\[
X_{TCSC} = -\frac{1}{\omega C} + \frac{A}{\pi \omega C} [2\sigma + \sin 2\sigma] - \frac{4A}{\pi \omega C(k^2 - 1)} \cos^2 \sigma [k \tan k\sigma - \tan \sigma] \tag{5}
\]

Where \( A = \frac{\omega_0^2}{\omega_0^2 - \omega^2} \), \( k = \frac{\omega_0}{\omega} \), \( \sigma = \frac{\beta}{2} \), \( \omega_0^2 = \frac{1}{LC} \), \( \pi/2 < \beta < \pi \). \( \beta \): firing angle

3 Adaptive Critic Based Neuro-Fuzzy Controller

According to psychological theories, some of the main factors of human beings’ learning are emotional elements such as satisfaction and stress. We always search for a way to lower our stress with respect to our environment [6]. This is a key idea in using of ACBNFC in control systems. ACBNFC is a kind of unsupervised learning methods for autonomous agents to acquire action rules to adapt clue of reward and punishment. In this approach the teacher of conventional supervised learning is replaced by an intelligent critic that assesses the performance of controller and evaluates the current states of system and generates proper reinforcement signal. In ACBNFC there exists an element in the control system called critic whose task is to assess the present situation which has resulted from the applied control action in terms of satisfactory achievement of the control goals and to provide the so called reinforcement signal (the stress). The controller should modify its characteristics so that the critic’s stress is decreased. Figure 1 shows the Configuration of the proposed ACBNFC for Firing Angle Control of TCSC. As shown in this figure, it contains three main items as: neuro-fuzzy controller, plant and fuzzy critic agent. In the subsequent sections, we briefly discuss the above elements.

![Fig. 1. Configuration of ACBNFC for TCSC](image_url)

Two major approaches of trainable neurofuzzy models can be distinguished. The network based Takagi-Sugeno fuzzy inference system and the locally linear neurofuzzy model. It is easy to see that the locally linear model is equivalent to Takagi-Sugeno fuzzy model under certain conditions, and can be interpreted as an extension of normalized RBF network as well. The Takagi-Sugeno fuzzy inference system is based on fuzzy rules of the following type:
**Rule:** If $u_i = A_i$ And ... And $u_p = A_p$

then $\hat{y} = f_i(u_1, u_2, ..., u_p)$  \hspace{1cm} (6)

Where $i=1...M$ and $M$ is the number of fuzzy rules. $u_1, ..., u_p$ are the inputs of network, each $A_{ij}$ denotes the fuzzy set for input $u_j$ in rule $i$ and $f_i(.)$ is a crisp function which is defined as a linear combination of inputs in most applications

$$\hat{y} = \omega_{i0} + \omega_{i1}u_1 + \omega_{i2}u_2 + ... + \omega_{ip}u_p$$ \hspace{1cm} (7)

Thus the output of this model can be calculated by

$$\hat{y} = \frac{\sum_{i=1}^{M} f_i(u)\mu_i(u)}{\sum_{i=1}^{M} \mu_i(u)}, \mu_i(u) = \prod_{j=1}^{p} \mu_{ij}(u_j)$$ \hspace{1cm} (8)

A simple form of $f_i(u)$ can be as following form:

$$f_i(u) = a_i u_1 + b_i u_2 + c_i$$ \hspace{1cm} (9)

The out put of controller is in the following form:

$$y = \frac{\sum_{i=1}^{m} \mu_i(a_i u_1 + b_i u_2 + c_i)}{\sum_{i=1}^{m} \mu_i}$$ \hspace{1cm} (10)

Where $m$ is number of controller fuzzy rules, $\mu_i$ is the firing strength of $i$th rules, $u_1$ is the first and $u_2$ is the second one for two input type controller. Here we choose $u_1$ and $u_2$ as error and its derivative respectively.

The most important block in figure 1 is the critic. The performance of the critic can be compared with the performance of emotional hue in humans. In absence of an exact evaluation of the present state in term of the objective value function, emotional cues like stress, satisfaction and etc. can be guide our control action into changing in the right direction so as to produce desired response. Similarly, the critic evaluates the state of system and generates a signal called emotion signal ($es$). This emotional signal is used to train and fine tune the neurofuzzy controller. Basically the critic acts as intelligent guide for the controller. The learning mechanism will be adapted the controller in order to satisfy critic and reduce the stresses of it. Here the critic is defined in fuzzy form. Fuzzy systems are very useful for critic modeling because the critic just gives us an approximate evaluation of current states of system. For this plan, the inputs of critic are error of plant output from desired response and its derivative. The emotion (output) of critic is a signal between $[-1, 1]$ and shows the performance of the system. If this emotion becomes zero, it means that the critic is satisfied by the performance of controller. If the emotion becomes larger, it shows the more stress and more dissatisfaction.
The main objective of learning mechanism is to satisfy total emotion and reduces total stress. This aim can be extracted through the following energy function:

$$J = \frac{1}{2} \sum_{i=1}^{N} K_i e_{s_i}^2$$  \hspace{1cm} (11)

Where $e_{s_i}$ is the $i$th emotional signal. $K_i$ is the corresponding output weights and $N$ is the total number of outputs. Learning is adjusting the weights of model by means of a nonlinear optimization method, e.g. the steepest descent or conjugate gradient. With steepest descent method the weights will be adjusted by the following variations:

$$\Delta w_i = -\eta \frac{\partial J}{\partial w_i}$$  \hspace{1cm} (12)

Where $\eta$ is the learning rate of the corresponding neurofuzzy model. After some calculation [5] the learning algorithm and adaptation parameters $a_i$, $b_i$ and $c_i$ in (10) is obtained as follow:

$$a_i(n) = a_i(n-1) + \eta e_{s_i}(n) e(n) \frac{u_i(n)}{\sum_{i=1}^{m} u_i(n)}$$

$$b_i(n) = b_i(n-1) + \eta e_{s_i}(n) e(n) \frac{u_i(n)}{\sum_{i=1}^{m} u_i(n)}$$  \hspace{1cm} (13)

$$c_i(n) = c_i(n-1) + \eta e_{s_i}(n) \frac{u_i(n)}{\sum_{i=1}^{m} u_i(n)}$$

## 4 Simulation Results

To evaluate the usefulness of the proposed method for damping the low frequency oscillations in power systems, the computer simulation for a single-machine infinite system is performed and compared with fuzzy PD controller. The analysis conditions, which are used for comparing control performance of fuzzy PD controller with ACBNFC, are summarized in Table 1. This table is classified according to the power system operating conditions used in designing ACBNFC and evaluating the robustness of the ACBNFC. As shown in Table 1, case-1 is used in designing the ACBNFC and we used case-2 to case-6 in evaluating the robustness of the PSS. Figure 2 shows the generator angle and firing angle when the three-phase fault occurs under the Case-1 of Table 1. As shown in Figure 2, the ACBNFC shows the better control performance than fuzzy PD controller in terms of setting time and damping effect. To evaluate the robustness of the proposed method, figure 3 shows the generator response characteristic in case that fuzzy PD controller and the proposed ACBNFC are applied under the Case-2 of Table 1. As seen in Figure 3, ACBNFC shows the better control performance than fuzzy PD controller in terms of setting time and damping effect.
Table 1. Simulation cases used in evaluation of controller performance
A: Three phase fault, B: Mechanical torque was changed as 0.2pu

| Simulation Cases | Operating condition | Disturbance | Fault time [msec] |
|------------------|---------------------|-------------|------------------|
| Case-1           | Heavy load $P_e = 1.5 [pu]$ | A           | 45               |
| Case-2           | $Q_e = 0.02 [pu]$    | B           | –                |
| Case-3           | Nominal load $P_e = 1.0 [pu]$ | A           | 45               |
| Case-4           | $Q_e = 0.02 [pu]$    | B           | –                |
| Case-5           | Light load $P_e = 0.5 [pu]$ | A           | 45               |
| Case-6           | $Q_e = 0.02$         | B           | –                |

Fig. 2. Generator responses when three-phase fault was occurred (Heavy Load)
Fig. 3. Generator responses when mechanical torque is changed by 0.2pu (Heavy Load)

To evaluate the robustness of the ACBNFC, Figs. 4 and 5 show the generator response characteristic in case that fuzzy PD controller and the ACBNFC are applied under the Case-3 and 4 of table2. As shown from these figures, the ACBNFC has the better control performance than fuzzy PD controller.

Fig. 4. generator responses when three-phase fault was occurred (Nominal Load)
Fig. 5. Generator responses when mechanical torque is changed by 0.2pu (Nominal Load)
To evaluate the robustness of the ACBNFC, Figs. 6-7 show the generator response characteristic in case that fuzzy PD controller and ACBNFC are applied under the Case-5 and 6 of table1. As seen from these figures the ACBNFC has the better control performance than fuzzy PD controller in terms of settling time and damping effect.

Fig. 6. Generator responses when three-phase fault was occurred (Light Load)  
Fig. 7. Generator responses when mechanical torque is changed by 0.2pu (light load)

5 Conclusion

The purpose of this paper, as seen, was to suggest another control approach, based on a modified version of Adaptive Critic Based Neuro-Fuzzy Controller (ACBNFC), for low frequency oscillation damping of power system. Simulation results showed that, the proposed method is very robust and the response time can achieve satisfactory performance. To evaluate the usefulness of ACBNFC, some computer simulations for a single machine infinite system were performed and compared with fuzzy PD controller. Obtained results showed that the performance of proposed method was very better than fuzzy PD controller in terms of settling time and damping effect. Then, to evaluate the robustness of the ACBNFC, we simulated dynamic characteristic of generator for a changeable mechanical torque and three-phase fault in nominal and light load. As seen from results the proposed method was very robust.

References

1. Gama. C., M. Noroozian, “Control Strategy for Damping of Power Swings Using TCSC”, Cigré Symposium, Kuala Lumpur, 1999
2. Lei. X., D. Jiang, and D. Retzmann, “Stability Improvement in Power Systems with Non-Linear TCSC Control Strategies”, ETEP, vol 10, No. 6, pp. 339–345, 1998
3. W.G. Kim, G.H. Hwang, H. T. Kang, S.O. Lee, “Design of Fuzzy Logic Controller for Firing Angle of TCSC Using Real-Type Tabu Search”, IEEE conference, pp.575-579, 2001
4. Y. Wang, Y. L. Tan, and G. Guo, "Robust Nonlinear Coordinated Excitation and TCSC Control for Power Systems," IEE Proc. Generation, Transmission and Distribution (UK), Vol. 149, No. 3, pp 367 - 372, 2002.
5. Rashidi. F., Rashidi M., Hashemi Hosseini A., “Emotional Temporal Difference Learning Based Intelligent Controller”, IEEE Coference, CCA, pp.341-346, 2003.
6. Rashidi. F., Lucas C., Khaki Sedigh A. “Design of Context Based Emotional learning for multiobjective systems”, ICEE03, Shiraz, Iran, 2003.
Use of Artificial Neural Networks in the Prediction of Kidney Transplant Outcomes

Fariba Shadabi¹, Robert Cox², Dharmendra Sharma², and Nikolai Petrovsky¹

¹ Medical Informatics Center, Division of Health, Design and Science
² School of Information Sciences and Engineering, University of Canberra, ACT, 2601, Australia

Abstract. Traditionally researchers have used statistical methods to predict medical outcomes. However, statistical techniques do not provide sufficient information for solving problems of high complexity. Recently more attention has turned to a variety of artificial intelligence modeling techniques such as Artificial Neural Networks (ANNs), Case Based Reasoning (CBR) and Rule Induction (RI). In this study we sought to use ANN to predict renal transplantation outcomes. Our results showed that although this was possible, the positive predictive power of the trained ANN was low, indicating a need for improvement if this approach is to be useful clinically. We also highlight potential problems that may arise when using incomplete clinical datasets for ANN training including the danger of pre-processing data in such a way that misleading high predictive value is obtained.

1 Introduction

AI modeling techniques such as Artificial Neural Networks (ANNs), Case Based Reasoning (CBR) and Rule Induction (RI) can be utilized in problems where large databases may contain useful knowledge that are far too large to be analysed by hand [1]. They can also be applied to environments where programs must adapt to changing conditions.

Artificial Neural Networks are known to provide good solutions to many problems [2, 3, 4]. However, there are a number of aspects of neural network implementation that make it a difficult task. Some important design decisions must be made in order to generate a reasonably accurate solution. Such decisions include choice of training method, number of layers and number of nodes. Detailed information about the foundations of ANNs can be found in [5].

In this study we considered the use of ANNs for predicting outcomes of medical procedures. We specifically used a trial data set made available to us from a kidney transplant database. The main objective of this project was to study the usefulness of ANNs in the prediction of successful or unsuccessful kidney transplants. This study was built upon work by Tam Soh Khum [6] and Petrovsky et al. [7]. This previous study reported good results (using trained ANN to predict kidney rejections at six months post transplantation).
2 Renal Transplant Challenges

The treatment for patients with end-stage renal diseases is either dialysis or kidney transplantation. There are thousands of men, women and children on the kidney transplantation waiting list. Given the critical shortage of kidneys for transplantation, a variety of techniques have been proposed to estimate the chance that a kidney graft will be functioning for a certain period of time after surgery. This study is concerned with creating an ANN-based approach for prediction of outcomes of kidney transplants.

3 Experimental Methodology

It should be noted that there is a fair amount of judgment in the use of ANN classifiers, which generally can be classed into two distinct categories:

a. The pre-processing of the data set into a form suitable for input into the ANN as training data. This includes decisions about appropriate parameters to be included in the data set and the input representation as well as the size of training, testing and validation set.

b. The selection of the ANN training parameters (number of nodes, training epochs, the training constants and output representation).

We conducted five separate experiments:

Experiment 1. We tried to re-produce the previous results by using the pre-processing strategy reported in their studies and the ‘original dataset’.

Experiment 2. We contacted the original authors and obtained their data in a post – preprocessed form, we then fed this data into our experimental set up.

Experiment 3. We constructed our own pre-processing method, using the original data. In doing this we changed the horizon of rejection from 6 months to 2 years. In effect we were trying to predict success or failure of the kidney in the 2 years after the transplant. We also removed the experimental bias found in Experiment 2.

Experiment 4. In an attempt to determine sensitivity to individual factors we systematically removed each input variable in turn and re-trained the ANN.

Experiment 5. We processed the same data used in experiment 4 into an RBF neural network (as opposed to previous experiments where we used a MLP neural network).

The data used in the project was obtained from a kidney transplant and dialysis database. The data was given as text files, with each file containing the data of a table. There are all together 35 tables in the database. The data dictionary of the database was used in interpreting the data contained in the tables [8].
Table 1. Variables used in data set for neural network predictions

| No. | Variable/code | Description                                      | Type and size |
|-----|---------------|--------------------------------------------------|---------------|
| 1   | AGE           | Age at transplant (Recipient)                    | NUMBER (2)    |
| 2   | MISA          | Number mismatches A                              | NUMBER (3)    |
| 3   | MISB          | Number mismatches B                              | NUMBER (3)    |
| 4   | MISD          | Number mismatches DR                             | NUMBER (3)    |
| 5   | MISDQ         | Number mismatches DQ                             | NUMBER (3)    |
| 6   | REFHOSP       | Referring hospital                               | CHARACTER (4) |
| 7   | REFSAT        | Referring state                                  | NUMBER (1)    |
| 8   | DONHOSP       | Donor hospital                                   | CHARACTER (4) |
| 9   | DONSTAT       | Donor state                                      | NUMBER (1)    |
| 10  | TRANHOSP      | Transplant hospital                              | CHARACTER (4) |
| 11  | TRANSTAT      | Transplant state                                 | NUMBER (1)    |
| 12  | CMV           | Recipient CMV antibody status                    | NUMBER (1)    |
| 13  | EBV           | Recipient EBV antibody status                    | NUMBER (1)    |
| 14  | DONSOUCRC     | Donor source                                     | NUMBER (2)    |
| 15  | DONAGE        | Donor age                                        | NUMBER (2)    |
| 16  | DONSEX        | Donor sex                                        | CHARACTER (1) |
| 17  | ISCHEMIA      | Total ischemia (to nearest hour)                 | NUMBER (2)    |
| 18  | MULTIPLE      | Has recipient had another organ transplanted?     | CHARACTER (1) |
| 19  | BLTRANA       | Ever transfused? (Before the first graft only)    | NUMBER (1)    |
| 20  | BLTRANB       | Number of units transfused                       | NUMBER (2)    |
| 21  | INSITU        | Insitu Y?                                       | CHARACTER (1) |
| 22  | KIDPRESI      | Initial kidney preservation                      | NUMBER (2)    |
| 23  | KIDPRESM      | Machine kidney preservation                      | NUMBER (1)    |
| 24  | TXSTATUS      | Did graft succeed or fail?                       | NUMBER (1)    |

The parameters used in the training set were selected from two tables. Some variables from these tables were removed because they were actually an indication of the outcomes of the transplant, and they are measured after the transplant has been made. The variables that were retained are shown in the table 1 excluding patient number and graft number. We pre-processed the data by performing normalization. We also followed the classification scheme that has been proposed by [6] in order to define the set of possible outcomes and to provide a general structure for the analysis. Notably we use a 6 month rejection horizon in experiments one and two and a two year rejection horizon in experiments 3 to 5.
For experiments 1 to 4 a Multilayer Perceptron (MLP) with a single hidden layer was trained to differentiate between successful and unsuccessful transplants. The training algorithm uses a trial and error approach to determining these three parameters. The 3-training parameters were: number of nodes in the hidden layer, number of training epochs and the training constant. The algorithm uses the following method:

1. For the number of nodes:
   • Given 2 nodes, it will add 1 node at a time until testing set accuracy stops increasing then it sets the final number of hidden nodes.
2. For the number of training epochs and training constant:
   • Using the test data: it sets the training constant to 0.1 and add 0.02 until there is no improvement in the training. It remembers this value, number of epochs and the accuracy.
   • Using the test data: it sets the training constant to 0.09 and subtract 0.01 until there is no improvement in the training after 3 tries. It remembers this value, number of epochs and the accuracy.
3. It takes the best training constant, number of epochs (peak value) and number of nodes from the above steps and retrains the network using the peak values.

Another point to note here is that in experiment 4 we ran 30 experiments (changing the initial random allocation of weights in the ANN) and averaged the results. In experiment 5 we tested the data on a Radial Basis Function (RBF) network of conventional design.

4 Research Results

4.1 Replicate the Work Done Previously

In this experiment we tried to replicate previous results reported [6, 7] in which an Artificial Neural Network (MLP network) was trained to predict whether a kidney transplant is going to be a success or failure. Using the classification scheme reported in the previous studies [6, 7], we achieved only an accuracy of prediction of 58.94% for successful transplant prediction and 53.75% for unsuccessful transplant prediction. This was considerably below the previous experimental results.

4.2 Use the Previous Pre-processed Input

Having received the pre-processed data from the original researchers we fed this data into our own neural network and we achieved 66.95% for successful transplant prediction and 98.65% for unsuccessful transplant prediction. This was a better result than the 84% and 71% accuracy rate previously reported (the improvement may well come from our automated trial and error trainer which optimizes the overall result and tends to outperform an ANN in which a human has set the training parameters).

We compared the data points used in experiment 1 and experiment 2 and found that after the selection and pre-processing stage, experiment 2 had a ‘biased’ training data set. It was noted that a slightly different selection strategy was used during the pre-
processing stage of successful and unsuccessful records. Consequently, records with missing information were removed from the successful dataset, but left in the unsuccessful data set. The neural network simply had to learn that if a column had a missing data point (which was represented by the number –1) it was a fail point. Figure 1 shows the size of missing data points in each target category in their training dataset.

![Figure 1. The percentage of missing data points in fail and success target category for the 23 variables.](image)

The missing data distribution shown in figure 1 has an impact on the network performance because of the unbalanced missing points in each category. It is easy to see that missing data is skewed to favor one or the other target category. This is particularly true for column 5 but is visible in columns 4, 5, 9, 12, 13, 17, 19, 20, 22 and 23. Therefore the skewing of missing data is a very plausible reason for the high neural network accuracy rate reported in the previous study. We suspect this was done mainly because significant numbers of records belong to the successful transplant category and data cleaning can be done without dramatically reducing the size of data. Therefore in order to match the size of successful dataset, records containing unsuccessful transplant with low quality points (ones with missing data) were not removed from the unsuccessful data set.

### 4.3 Pre-processing with Our Own Methodology

In this experiment we used our own pre-processing methodology. We decided to pre-process the data in a slightly different manner to avoid creating variables with skewed distribution of missing data. We also changed the pre-processing rejection period to two years. This gave us a final data set of 672 fail points. We randomly sampled 672 success points all of which had no skewed missing data. We used all the 23 attributes as inputs of the neural networks. Importantly we also divided the data into three equally sized sets, the training set, the test set and the validation set (the previous study had a 70/30 split into training data and testing data). The validation set gave a 61.64% overall accuracy rate which is disappointing.
4.4 Search for a Subset

We tried to improve the performance of our network and find the best subset of features by using different combinations of variables. The best the validation set reported was 62% accuracy. As a part of our investigation, we removed each column of data one at a time to see if any single column of data was having a significant influence on prediction accuracy. As it can be seen in table 2, the highest accuracy rate is obtained when we removed the variable number 14. However the results are still around 61%. We conclude that no single variable has significant effect on the graft outcome. These results are still not satisfactory for the purpose of clinical prediction.

| Removed Variable | Final Validation % | Final Validation % "Average After 30 Run" |
|------------------|---------------------|-------------------------------------------|
| 1                | 60.96               | 61.35                                     |
| 2                | 61.87               | 61.60                                     |
| 3                | 62.10               | 62.20                                     |
| 4                | 61.87               | 61.93                                     |
| 5                | 61.64               | 61.58                                     |
| 6                | 60.50               | 62.40                                     |
| 7                | 61.42               | 61.82                                     |
| 8                | 62.33               | 60.92                                     |
| 9                | 61.64               | 61.89                                     |
| 10               | 62.10               | 61.68                                     |
| 11               | 64.38               | 62.06                                     |
| 12               | 64.84               | 62.13                                     |
| 13               | 64.38               | 62.45                                     |
| 14               | 62.79               | 62.86                                     |
| 15               | 60.50               | 60.80                                     |
| 16               | 59.36               | 61.86                                     |
| 17               | 57.99               | 58.65                                     |
| 18               | 59.82               | 61.54                                     |
| 19               | 59.59               | 59.03                                     |
| 20               | 60.73               | 61.81                                     |

4.5 Performance of RBF Network

The next approach taken was to implement a RBF network. RBFs are one possible choice for the ANN transfer function. We used a RBF of conventional design; it uses the K-Means algorithm to allocate the location of the radial nodes. Each node has a separately calculated width based on the distance to the nearest 4 nodes. The output layer is trained using a linear optimization algorithm. Like the MLP, the data is split into test, training and validation sets and training was done by an automated trail and
error mechanism. In this experiment, the validation set with all 23 columns gave 61.42% accuracy. We conclude that the RBF and MLP performance is generally similar on this data set.

5 Conclusion and Further Work

Despite using a range of pre-processing and ANN solutions for prediction of outcomes of kidney transplants, we found that the resultant accuracy of approximately 62% was probably too low to be of any clinical use.

Further work will involve using a more complete dataset and drawing together our recent research results with previous relevant work, with a view towards developing a methodology for generating better predictions in medical domains.

References

1. Wolberg W. H., Street W. N., and Mangasarian O. L.: Machine learning techniques to diagnose breast cancer from needle aspirates. Cancer Letters, Vol. 77 (1994) 163-171
2. Pantel P.: Breast Cancer Diagnosis and Prognosis. University of Manitoba (1998)
3. Zainuddin Z.: Advanced Neural Network Learning Applied to Breast Cancer Detection. Proceedings of the Eight Australian and New Zealand Intelligent Information Systems Conferences-ANZIIS (2003) 367-472 ISBN 1-74107-039-2
4. Street, W. N.: A neural network model for prognostic prediction. Proceedings of the Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann (1998)
5. Simpson, P. K.: Foundation of Neural Networks. Artificial Neural Networks-Paradigms, Applications, and hardware implementation, IEEE Press (1992) ISBN 0-87942-289-0
6. Tam, S. K.: What Determines The Outcome of Kidney Transplants. Master thesis, National University of Singapore (2001)
7. Petrovsky, N., Tam, S. K., Brusic, V., Russ, G., Socha, L., Bajic, V. B.: Use of artificial neural networks in improving renal transplantation outcomes. Graft (2002) 6-13
8. Data Dictionary, ANZDATA Registry Database (2000)
An SoC-Based Context-Aware System Architecture

Keon Myung Lee†, Bong Ki Sohn†, Jong Tae Kim†, Seung Wook Lee†,
Ji Hyong Lee‡, Jae Wook Jeon‡, and Jundong Cho†

† School of Electric and Computer Engineering, Chungbuk National University, Korea
kmlee@cbnu.ac.kr
‡ School of Information and Communication Engineering,
SungKyunKwan University, Korea

Abstract. Context-aware computing has been attracting the attention as an approach to alleviating the inconvenience in human-computer interactions. This paper proposes a context-aware system architecture to be implemented on an SoC (System-on-a-Chip). The proposed architecture supports sensor abstraction, notification mechanism for context changes, modular development, easy service composition using if-then rules, and flexible context-aware service implementation. It consists of the communication unit, the processing unit, the blackboard, and the rule-based system unit, where the first three components reside in the microprocessor part of the SoC and the rule-based system unit is implemented in hardware. For the proposed architecture, an SoC system has been designed and tested in an SoC development platform called SystemC. This SoC-based context-aware system architecture has been developed to apply to mobile intelligent robots which would assist old people at home in a context-aware manner.

1 Introduction

Users are getting tired of feeding into a computer every single detail on how to accomplish the task when using it. Context-aware computing has been attracting the attention as an approach to alleviating the inconvenience in human-computer interaction. Context is any information that can be used to characterize the situation of an entity, where an entity is either a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves. The typical instances of contexts include location, identity, time and activity. These pieces of context information may enable computers to answer the questions of who, what, when, and where. A system is said to be context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user’s task.

* This work has been supported by the Center for Intelligent Robotics, which carries out one of the 21st century’s Frontier R&D Projects sponsored by the Korea Ministry of Science & Technology.
Context may be used in various ways on developing context-aware applications. The following is one of categorization of features for context-aware applications introduced by Dey\cite{Dey_2001}: *contextual presentation* of information and services which is the ability to present information and services to a user automatically based on available context, automatic *contextual execution* of a service which is the ability to execute or modify a service automatically based on the current context, and *contextual augmentation* of information for later retrieval which is to associate digital data with the user’s context. Thanks to increased availability of commercial, off-the-self sensing technologies, prevalence of powerful, networked computers and mobile computing devices, there will be increasing demands on context-aware applications.

Context-aware applications strongly depend on sensors to get context information. Various applications have been developed in a way to be tightly coupled with sensors due to the inherent nature that context information is acquired from non-traditional devices like cameras, floor pressure sensors, active badges, and the like. It is not a good software engineering practice since such style of implementation makes it difficult to reuse or replace already-implemented components. Sometimes, context interpretation is needed to transform one or more types of context into another type of context. There are some situations at which the notification about relevant changes to context value is required for context-aware services. It would be great to have a platform with which we can easily construct context-aware applications with small amount of effort. There will appear many context-aware applications embedded into mobile devices or mobile robots. From these observations, we developed a context-aware system architecture to be implemented in an SoC(System-on-a-Chip) for mobile devices or mobile robots.

2 Related Works

There are several architectures that have been developed specifically to support context-aware applications. These architectures were designed to be applicable to a range of context-aware applications, but most of them, in fact, deal with only a portion of the context-aware system features.

The Stick-e Notes system\cite{Stick-eNotes_2001} is a general framework for supporting a certain class of context-aware applications, of which goal is to allow non-programmers to easily author context-aware services using *if-then* rules. In this system, semantics for writing rules for service description is quite limited since it is intended for non-programmers.

CoolTown\cite{CoolTown_2002} is an infrastructure that supports context-aware applications by mapping each real world object, including people, places and device, to a Web page called *Web presence*. Depending on the obtained context information, the corresponding Web presence is updated and displayed, if decided it is needed. It was not intended to support other context-aware features like automatic execution of services, contextual augmentation.
The TEA project[5] proposed an architecture to provide awareness of context to personal mobile devices. It uses a blackboard model that allows sensor representations to write their acquired context to it and context interpreters to read context and write interpreted or abstracted context, and applications to read relevant context from it.

Dey’s Context Toolkit[1] is a framework to support building and execution of context-aware applications. The framework consists of the following components: context widgets for acquiring context from sensors, context interpreters to interpret context, context aggregators to collect related context together, context services to provide reusable context-aware behaviors or services to applications, and discoverer for discovering resource. It allows developers to use these components and modify them, if needed, which are implemented in library functions.

Our context-aware system architecture aims at the followings: sensor abstraction to separate sensor data processing from context-aware service, notification of context changes, system evolution to allow flexible addition and replacement of sensors and contexts, easy modular implementation of applications, implementation on a peanut processor (i.e., an SoC) for mobile devices or robots.

3 The Proposed Context-Aware System Architecture

The proposed context-aware system architecture consists of communication unit, processing unit, rule-based system, and blackboard as shown in Figure 1. The communication unit takes charge of taking external and internal messages and delivering them to corresponding processing modules, and receiving results produced by the system and sending them to the corresponding external counterparts. The processing unit contains the context acquisition modules, query processing modules, action modules, and task processing modules. The rule-based system unit plays the role of representing *if-then* rules and finding the rule to be executed next. The blackboard is a shared data repository through which those units share data and information. A portion of the blackboard serves as the working memory for the rule-based system.

![Diagram of the proposed Context-aware System Architecture](image-url)
To separate the details of dealing with the sensors from the applications and thus to allow the applications to deal with only the context they are interested in, the proposed architecture takes the strategy to use independent processing modules for sensor data processing and for context-aware services. For easy construction of context-aware services, the architecture allows the developer to write if-then rules which state what actions to be taken under what contexts. These if-then rules are maintained in the rule-based system and the matching operations for if-parts of rules are performed by the rule-based system and the then-parts of rules can be made to be executed by action modules in the processing unit.

**Communication Unit.** The communication module works in event-driven programming style, where there are an input message queue and an output message queue. For each queue, there is a message delivering loop which fetches a message from the queue one at a time, finds the processing module corresponding to the message, and then activate the processing module with the message. In the proposed architecture, there are four types of input messages: sensor messages, query messages, task messages, and internal module call messages. Sensor messages encapsulate sensor data received from external sensors. They are transferred to their corresponding context acquisition modules which extract context information from sensor data. Query messages are generated by an external system (e.g., a context-aware application) to inquire some information about another system. Once query messages are received, they are handed over the relevant query processing modules. Task messages are requests from external systems to ask for the recipient system to do some tasks on behalf of them. Those messages are tossed to corresponding task modules. Internal module call messages are the invocation request for a processing module by another module. To avoid tight coupling between processing modules, processing modules are invoked in an indirect manner using internal module call messages. The proposed architecture controls its own peripherals using its processing modules and uses the communication unit to send commands and/or messages to external sensors and systems. The outgoing messages are enqueued into the output message queue, and the output message queue manager employs proper handlers to deal with each message one by one. The communication unit maintains the yellow page about which module can handle what kind of messages. Each time new messages are added, the information of its corresponding processing modules is registered into the yellow page.

**Processing Unit.** The processing unit contains four types of processing modules: context acquisition modules, query processing modules, action modules, and task processing modules. The context acquisition modules play the role of extracting context information from sensor data along with other existing information available through the blackboard. The proposed architecture allows to notify context changes to processing modules and/or if-then rules. To provide the notification mechanism, each context acquisition module maintains two registries, one for keeping the list of processing modules to be notified and the other for keeping the list of flags for the rules to be notified. To notify the context change to if-then rules, we take the strategy to append to the if-part of rules an extra
flag which is used to inform rules of context change. Once the related context is changed, its related flags are set by the context acquisition modules. If a rule is fired and executed, then all its notification flags are reset by the action module. A context acquisition module contains the following basic routines: sensor data processing routine, update routine to generate internal module call messages for the processing modules and to modify rule flags in the blackboard. To tell the context change, a context acquisition module retains the states about the previous sensor data, the corresponding context and the update time of the context. The query processing modules retrieve requested data from the blackboard, generate an output message containing the data, and put it into the output message queue. The action modules take care of the then-parts of rules. The rule-based system enables the then-part of a rule to invoke one or more action module(s). In this way, the proposed architecture allows the developer to implement complicated and delicate works using if-then rules. On the other hand, action modules can be used to perform context interpretation for primitive context. The task processing modules perform the tasks asked from the external systems and send back to them a message containing the processing results, if any.

Rule-Based System Unit. The rule-based system unit contains if-then rules which describe high-level context processing and context-aware services. Some high-level context extraction can be expressed using if-then rules, which can be encoded into the rule-based system. The following is such a rule exerted from our old people assistance demo system we are developing: IF the master’s present location is not of one of rest spots and has been staying at the location during more than the specified time, THEN master is in the state of ‘long-resting-at-improper-location’. On the other hand, the rule-based system is used to compose the context-aware services as follows: context-aware presentation, context-aware automatic execution, and context-aware augmentation. For context-aware presentation, an if-then rule is constructed of which if-part describes the considered context and then-part describes the name of the action module to be invoked. Based on the current context, an action module collects the relevant information from the blackboard and presents it to the user. By loading such action modules into the processing unit, registering the corresponding if-then rule into the rule-based system, and registering the correspondence of the internal module call message to action module into the communication module, the developer can initiate a new context-aware presentation service. In context-aware services, the then-part invokes an action module which automatically executes some tasks. In context-aware augmentation services, the corresponding action modules append related information to the considered context object of the blackboard and create new objects for the augmented data in the blackboard, if necessary. The internal module call messages generated by the rule-based system unit contain the information about the name of an action module to be invoked and related context information referred in the if-part.

Blackboard. The blackboard is the repository of context state variables, rule notification flags, temporary variables to store intermediate results, and databases
of some related data, which can be accessed by the communication unit, the processing unit, and the rule-based system unit. A portion of the blackboard serves as the working memory for the rule-based system unit. In the SoC-based implementation, the working memory elements are hardwired to the SoC hardware components. All blackboard elements including databases have unique identifiers and are regarded as objects by processing modules. For databases, there are corresponding processing modules in the processing unit which insert, delete, and update records, and, in addition, process queries.

4 An SoC-Based Implementation

An SoC is designed to implement the proposed context-aware system architecture, where the communication unit, the processing unit, and the blackboard reside in the microprocessor part of the SoC and the rule-based system unit is implemented in hardware as shown in Figure 2. The rule-based system hardware executes matching operations for the if-parts of all rules in parallel. Rule Base CAM(content-addressible memory) block encodes the if-parts of all rules, in which a CAM array is allocated for a rule (in fact, for the if-part of a rule) and matching operations based on equivalence tests(e.g., \(x = True\)) are performed at a time for all rules. The enabled rules are informed to the Conflict Resolution module in which a rule is selected as the one to be executed next. The current design supports two conflict resolution strategies: Priority-based resolution and LRU(least recently used)-based resolution. For the selected enabled rule, the Conflict Resolution module either generates an action module call message for its then-part or performs update operations for the working memory. The update operations by the Conflict Resolution module are all simple value setting to change the value of state variables. For the fast parallel matching, the conditions checked in the Rule Base CAM are all equivalence tests. Along with equivalence tests, conventional rules may use other comparison tests including such operators as \(>, \geq, \neq, <, \leq\). The architecture employs a specially designed crossbar

![Fig. 2. The SoC Architecture for the Rule-based System Unit](image-url)
network (Crossbar Network module) for preprocessing condition tests other than equivalence tests. The module has several preprocessing modules (PPM) capable of performing any comparison tests. The Switch Input module controls the Crossbar Network to make PPMs perform the comparison operations occurring in rules in a load-balanced way and then store the results into SoC Working Memory. Once all non-equivalence condition tests are preprocessed, Rule Base CAM module is activated. The designed rule-based SoC architecture has been described and simulated and verified using the SystemC package[8]. SystemC is one of famous design and simulation tools for SoC development, which is a modeling language based on C++ language intended to enable system level design and IP (Intellectual Property) exchange. The microprocessor part of the SoC implements the communication unit, the processing unit, and the blackboard. Figure 3 shows a SystemC simulation result wave form.

5 Conclusions

An SoC-based context-aware system architecture was designed to implement context-aware applications embedded in a mobile intelligent robot or a mobile device. The rule-based system hardware was successfully tested on SystemC simulation platform for a rule base with about fifty rules engineered to assist old people at home. The contributions of this study are as follows: The proposed architecture provides sensor abstraction by which the developers can easily add or replace sensors and contexts, and implement new context-aware services. It provides notification mechanism for context changes which simplifies system implementation by making it possible to avoid continuous polling of context values. It supports modular system development by employing the message-driven processing module invocation method. It makes the developer easily author context-
aware services by associating context with actions through *if-then* rules. It is
developed to fit into a peanut processor like SoC, and thus may be employed
in mobile devices or mobile robots. As the further works, there remains to de-
velop standard interfaces and library functions of those software modules for the
proposed architecture.

References

[1] A. K. Dey. Providing Architectural Support for Building Context-Aware Applica-
tions. Ph.D. dissertation. Georgia Institute of Technology. (2000).
[2] P.J. Brown. The stick-e document: a framework for creating context-aware appli-
cations. *Electronic Publishing* 9(1). (1996) 1-14.
[3] D. Caswell, D. P. Debaty. Creating Web representations for places. *Proc. of the 2nd
Int. Symp. on Handheld and Ubiquitous Computing.* (Bristol, UK. 2000). 114-126.
[4] J. Pascoe. Adding generic contextual capabilities to wearable computers. *Proc. of
the 2nd IEEE Int. Symp. on Wearable Computers.* (Pittsburgh, PA. 1998). 92-99.
[5] A. Schmidt, K. A. Aidoo, et al. Advanced Interaction in Context. *Proc. of HUC’99.
(1999). 89-101.
[6] A. Dey, J. Mankoff, G. Abowd, S. Carter. Distributed mediation of ambiguous
context in aware environments. *Proc. of the 15th annual ACM symp. on User in-
terface software and technology.* (Paris 2002).
[7] S. Swan, *An Introduction to System Level Modeling in SystemC 2.0.* May 2001.
http://www.systemc.org
An Intelligent Control of Chaos in Lorenz System with a Dynamic Wavelet Network

Yusuf Oysal
Anadolu University, Computer Engineering Department, Eskisehir, Turkey
yoysal@anadolu.edu.tr

Abstract. This paper proposes a dynamic wavelet network based intelligent adaptive controller design to regulate the chaotic states of the Lorenz equations. The “Dynamic Wavelet Network (DWN)” has lag dynamics, non-orthogonal mother wavelets as activation function and interconnection weights. Adaptation is done by adjusting parameters of the DWN to minimize the cost functional of the Lorenz system operating state errors. The cost gradients with respect to the network parameters are calculated by adjoint sensitivity analysis. It is illustrated in simulations that this control approach is more successful than the previous controllers for eliminating the tracking errors due to the set point changes.

1 Introduction

In many practical industrial applications, the most undesirable situation is the chaotic behavior that makes them more complex and irregular. Controlling or ordering of chaos is receiving increasing attention in order to remove the chaos and to give the system stable and predictable behavior. The most popular case is the Lorenz chaos control, because the Lorenz system captures many of the features of chaotic dynamics and it has been extensively studied as a benchmark system. For example, Vincent and Yu [1] used a bang-bang controller to control Lorenz chaos, showing that it is possible to drive a Lorenz chaotic system to one of the unstable equilibrium points. In addition to that, Yu [2] proposed a variable structure control strategy to stabilize the well-known Lorenz chaos. Moreover, Hwang et al. [3] presented a nonlinear feedback controller to regulate the chaotic states of the Lorenz equation and obtained successful result for faster settling time than perturbation linearization based classical analysis of Hartley and Mossayebi [4].

Beside these studies, there are several new controllers such as intelligent controllers and adaptive controllers applied for chaos control in Lorenz system. For example, Yeap and Ahmed [5], designed a suboptimal feedback controller implemented by a multiplayer feed-forward neural network to control the unpredictable behavior of chaotic systems such as Lorenz and Rossler systems. Zeng and Singh [6] presented an adaptive control law to regulate the chaos in Lorenz system in the presence of system parameter uncertainty to a specified point in the state space. Liao [7] proposed an adaptive control law such that two Lorenz systems are to be synchronized by guaranteeing the robustness against a bounded disturbance.
This paper proposes an intelligent adaptive controller based on neural network and wavelet technologies for chaos control in a Lorenz system to combine the effectiveness of wavelets in representation of nonstationary (transient) signals and the best advantages of each. A “Dynamical Wavelet Network (DWN)” ([8],[9]) that contains dynamical elements such as delayers or integrators in their processing units is used in the adaptive controller design to control the Lorenz system with a better performance and without any singularity.

2 Lorenz System

Lorenz proposed a simple model for the unpredictable behavior of the weather in 1963. He obtained the model equations of the motion of a two dimensional cell of fluid cooled from above and warmed from below as:

\[
\begin{align*}
\dot{x} &= P(y - x) \\
\dot{y} &= Rx - xz - y \\
\dot{z} &= xy - bz
\end{align*}
\]

where \(P > 0\) is the Prandtl number, \(R\) is the Rayleigh number, \(b > 0\) is a spatial constant, \(x\) is the fluid velocity in the loop, \(y\) and \(z\) are the vertical and horizontal temperature differences, respectively.

Generally, the Rayleigh number \(R\) is allowed to vary, while \(P\) and \(b\) held constant. The system has three stable equilibrium points. One of them is at the origin for \(0 < R < 1\), and the others are at \((\pm\sqrt{b(R - 1)}, \pm\sqrt{b(R - 1), R - 1})\) for \(1 < R < R^*\). \(R^*\) is the critical value for these two equilibrium points to become unstable. For example, if \(P = 10\), and \(b = 8/3\), then \(R^* = 24.74\). Stability analysis of these points can be found in [10].

With the feedback control signal \(u\) to the input \(R\), equation (2) can be written as:

\[
\dot{y} = R_{eff} x - xz - y
\]

where \(R_{eff} = R + u\) is the effective Rayleigh number. Our aim in this study is to design an adaptive dynamic wavelet network based controller design to calculate \(u\), so that we will be able to construct a stable equilibrium region of points that are trackable.

3 The Architecture of a Dynamic Wavelet Network

DWN model is a network that has unconstrained connectivity and has dynamical elements in the wavelet processing units, “wavebons” (wavelet neurons), with wavelet activation functions. The details of the DWN can be found in [8].

The computational model of a DWN can be written as in [8] and [11] with the following equations:
The output of a unit \( y_i \) is a wavelet activation function \( \phi_i(x_i) \) of a state variable \( x_i \) associated with the unit. \( r_{ij} \) is the input connection weights from \( j^{th} \) input to \( i^{th} \) wavelon, \( w_{ij} \) is the interconnection weights from \( j^{th} \) wavelon to \( i^{th} \) wavelon and \( q_{ij} \) is the output connection weights from \( j^{th} \) wavelon to \( i^{th} \) output. \( T_i \) is the dynamic constant of \( i^{th} \) wavelon and \( d_i \) is the bias term added to the unit input (or polarization) of \( i^{th} \) wavelon.

The initial conditions on the state variables \( x_i(0) \) must be specified. An example of a DWN open diagram with single-input/single-output (SISO) and two-wavelon is shown in fig 1.

![Fig. 1. The state diagram of a single input / single output DWN with two-wavelons](image-url)

Wavelets are usually explained as basis functions, which are compact (closed and bounded), orthogonal (or orthonormal) and have time-frequency localization properties. Basis functions are called “activation functions” in ANN literature and can be with global or local feature in time. In this study, only the local basis functions have been used. For this reason, the generalized nonorthonormal Mexican Hat basis function is selected as the \( i^{th} \) activation function which can be written as:
where $\mu_i$ and $\sigma_i$ are translation (center) and dilation (standard deviation) parameters, of the $i^{th}$ wavelet respectively. These types of wavelet functions have efficient time-frequency localization properties as shown from the frequency spectrum in [12].

## 4 Intelligent Chaos Control

In this section, the design details for the intelligent chaos control to regulate the chaotic states of the Lorenz equation is given. An adaptive controller was designed by a SISO DWN with two wavelons (Fig. 1) connected as in Fig. 2. The input of the DWN is the error ($e$) between the vertical temperature difference ($y$) and the reference state ($y_{ref}$). The output of the DWN is the control signal for chaotic Lorenz system, which makes the reference states to be tracked. Adaptation is based on the DWN training that is adjusting parameters of the DWN. This may be online or offline according to the process to be controlled as in Lorenz system control design.

The total state vectors can be written as ($\vec{X} = [x \ y \ z \ x_1 \ x_2]^T$), where $x_1$ and $x_2$ are the state variables of the DWN model with the following differential equations:

$$
\dot{x}_1 = \frac{1}{T_1} (-x_1 + w_{11} \phi_1(x_1) + w_{12} \phi_2(x_2) + d_1 + \eta_1 e) \quad x_1(0) = 0 \tag{9}
$$

$$
\dot{x}_2 = \frac{1}{T_2} (-x_2 + w_{21} \phi_1(x_1) + w_{22} \phi_2(x_2) + d_2 + \eta_2 e) \quad x_2(0) = 0 \tag{10}
$$

Hence the Lorenz system with controller is expressed as $\dot{\vec{X}} = f(\vec{X}, u)$ where the control signal can be written as:

$$
u(t) = q_{11} \phi_1(x_1) + q_{12} \phi_2(x_2) \tag{11}$$
Adaptation is done by minimizing the cost functional of state errors with respect to the DWN parameters \( p = [w \ r \ d \ q \ T \ \mu \ \sigma] \) as seen in fig. 2. As a performance index or cost structure a simple quadratic form is selected as follows:

\[
J = \int_{0}^{t_f} F(\bar{x}(t),t)dt = \frac{1}{2} \int_{0}^{t_f} (y_{ref}(t) - y(t))^2 dt
\]  

so that, the state errors can be minimized with this cost minimization.

For these type optimization problems, gradient-based algorithms have been used. In this study, cost gradients with respect to the network parameters were calculated by adjoint sensitivity analysis. “Adjoint” method is used for sensitivity computation ([8],[9],[11],[13]) based on a new dynamical system defined with adjoint state variables \( \lambda \):

\[
-\dot{\lambda} = \left( \frac{\partial \bar{f}}{\partial \bar{x}} \right)^T \lambda + \left( \frac{\partial F}{\partial x} \right), \quad \lambda(t_f) = 0
\]

The size of adjoint vector is \( n \) and is independent of network parameters. There are \( n \) quadratures for computing the sensitivities. The integration of the differential equations must be performed backwards in time, from \( t_f \) to 0. In this study, 5th order Runga-Kutta-Butcher integration rule is used [14]. Let \( p \) be a vector containing all network parameters. Then, the cost gradients with respect to parameters are given by the following quadratures:

\[
\frac{\partial J}{\partial p} = \int_{0}^{t_f} \left( \frac{\partial \bar{f}}{\partial p} \right)^T \lambda dt
\]

Calculation of cost gradients is straightforward from equations (12)-(14) with the total controlled Lorenz system equations. But we also need to calculate the gradients of the activation functions of each wavelon with respect to its own states from the equation:

\[
\frac{\partial \phi_j(x_j)}{\partial x_j} = \left[ 3\phi_j + 2\left( \frac{x_j - \mu_j}{\sigma_j} \right) \right] \left( \frac{x_j - \mu_j}{\sigma_j^2} \right) j = 1,2
\]

Broyden-Fletcher-Golfarb-Shanno (BFGS) gradient method is used for updating of network weights [15]. This method provides the history of parameter and gradient changes yielding approximate second order information.

5 Simulation Results

In this section, a series of the numerical examples to verify the performance of the new intelligent adaptive controller for the Lorenz equations will be given. The parameters are selected as \( P = 10 \), \( b = 8/3 \), and \( R = 28 \).
In the numerical examples, we first let the reference input $y_{ref} = 8.485$ as the fixed point of the original Lorenz system. The control is activated at time $t = 5$. For this case, DWN parameters are calculated offline with the BFGS gradient update laws. Calculated DWN interconnection parameters and time constants are:

$$
w = \begin{bmatrix} 26.95 & 5.325 \\ 18.055 & 9.253 \end{bmatrix}, \quad r = \begin{bmatrix} -8.978 \\ -13.187 \end{bmatrix}, \quad q = \begin{bmatrix} 8.325 & 51.67 \end{bmatrix}, \quad d = \begin{bmatrix} 6.347 & -15.083 \end{bmatrix}, \quad T = \begin{bmatrix} 1.13 & 0 \\ 0 & 0.754 \end{bmatrix}
$$

And respective translation parameters are $\mu_1 = 4.3496$, $\mu_2 = -2.14$ and dilation parameters are $\sigma_1 = 2.04$, $\sigma_2 = 1.23$.

As shown solid line in fig. 3, the present controller successfully controls the flow velocity to $y_{ref}$, thus, it successfully eliminates the chaotic behavior. But, we cannot say that this controller design gives the best results. For comparison we investigate the nonlinear feedback controller [3] with $k_p = 1$, and the P (proportional) controller [4] with $k_p = 1$. In fig. 3, the system responses with each controller are expressed as the broken lines. It is seen that our controller design and the P controller have underdamp with larger overshoot and longer settling time than the nonlinear feedback controller.

As another numerical example, a tracking problem is considered. It is desired to control the state form the fixed point $y_{ref} = 8.485$ to $y_{ref} = 12$. The control is actuated at time $t = 5$ first, then the tracking state changed from 8.485 to 12 at time $t = 10$. Adaptation of the DWN is now online after time $t = 5$. To see the intelligence of our control system, only the change of the parameter value of $q_{11}$ is permitted. The variation of this parameter with respect to each time steps is shown in fig. 4. It is observed in fig. 4 that, the system by using the PI controller [4] has undergone overshooting and also has a longer settling time.

The numerical results show that the system with adaptive DWN based controller achieves settling time than the nonlinear feedback controller in [3], and PI controller
Fig. 4. Chaos control into a fixed point with three different controllers. The control is activated at time $t = 5$ and the reference signal is changed from 8.485 to 12 at time $t = 10$ in [4]. Therefore, this simulation result proves the robustness and effectiveness of the present intelligent chaos control scheme due to a good performance for tracking the states along the equilibrium manifold.

6 Conclusion

In this paper, a dynamic wavelet network is applied to chaos control in a Lorenz system. This application is an alternative and successful controller for eliminating the tracking errors due to the set point changes. Proposed DWN based intelligent controller can be easily adapted to different chaotic processes online. There is one other important motivation for investigating dynamic wavelet networks for chaos control is that they suggest a simple hardware implementation. It is the simple and plausible VLSI implementation for a continuous-time dynamic wavelet network [16].

References

1. Vincent, T.L., and Yu J.: Control of a Chaotic System. Dynamics and Control, Vol. 1(1) (1991) 35-52
2. Yu, X.: Controlling Lorenz Chaos. International Journal of Systems Science, Vol. 27(4) (1996) 355-359
3. Hwang, C.C., Fung, R.F., Hsieh, J.Y., and Li, W.J.: A Nonlinear Feedback Control of the Lorenz Equation. International Journal of Engineering Science, Vol. 37(14) (1999) 1893-1900
4. Hartly, T.T., and Mossayebi, F.: Classical Control of a Chaotic System. International Journal of Bifurcation and Chaos, Vol. 2 (1992) 881
5. Yeap, T.H., and Ahmed, N.U.: Feedback Control of Chaotic Systems. Dynamics and Control, Vol. 4(1) (1994) 97-114
6. Zeng, Y., and Singh, S.N.: Adaptive Control of Chaos in Lorenz System. Dynamics and Control, Vol. 7(2) (1997) 143-154
7. Liao, T.H.: Adaptive Synchronization of Two Lorenz Systems. Chaos, Solitons, & Fractals, Vol. 9(9) (1998) 1555-1561
8. Becerikli, Y., Oysal, Y., and Konar, A.F.: On a Dynamic Wavelet Network and Its Modeling Application. Lecture Notes in Computer Science, Vol. 2714, Springer-Verlag (2003) 710-718
9. Becerikli, Y., Oysal, Y., and Konar, A.F.: Modeling of Nonlinear Systems with Dynamic Wavelet Networks. Proc. Turkish Symposium on Automatic Control, Vol. 1, METU, Ankara (2002) 71-79
10. Sparrow, C.: The Lorenz Equations. Bifurcations, Chaos and Strange Attractors. Springer-Verlag, Berlin (1982)
11. Becerikli, Y., Konar, A.F., and Samad, T.: Intelligent Optimal Control with Dynamic Neural Networks. Neural Networks, Vol. 16 (2003) 251-259
12. Mallat, S.: A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 11(7) (1989) 674-693
13. Konar, A.F., and Samad, T.: Dynamic Neural Networks. Technical Report SSDC-92-I 4152-2, Honeywell Technology Center, 3660 Technology Drive, Minneapolis, MN 55418 (1992)
14. Chapra, S.C., and Canale, R.P.: Numerical Methods for Engineers. McGraw-Hill, Inc. (1989)
15. Edgar, T.F., and Himmelblau, D.M.: Optimization of Chemical Processes. McGraw-Hill (1988)
16. Anderson, J.A., and Rosenfeld: Neurocomputing: Foundation of Research, Cambridge, MA: MIT Press (1988)
Intelligent Robot Control with Personal Digital Assistants Using Fuzzy Logic and Neural Network

Seong-Joo Kim, Woo-Kyoung Choi, and Hong-Tae Jeon

School of Electronic and Electrical Engineering, Chung-Ang University
221, Heuksuk-Dong, Donjak-Gu, Seoul, 156-756, Korea
ksj1212@ms.cau.ac.kr

Abstract. To control the mobile robot system with wired controller by manual is so easy but user must keep the status and movement of mobile robot all the time. It is more efficient to control the mobile robot with remote controller by automatically. The user does not need to know the current status of the robot or where it is. In this paper, we propose the intelligent robot control technique for mobile robot using the wireless and remote controller, personal digital assistants (PDA) that has an intelligent navigation algorithm such as fuzzy logic and neural network. With the proposed technique, the mobile robot can trace human at regular intervals by the remote control method with PDA without user’s manual command. The mobile robot can recognize the distances between it and human whom the robot must follow with both multi-ultrasonic sensors and PC-camera and then, can decide the direction and velocity of itself to keep the given regular distances. The proposed PDA control system, which is intelligent and remote, can make user be free from the observation of the mobile robot to control the robot properly.

1 Introduction

In this paper, the intelligent remote control system for the mobile robot system using PDA (Personal Digital Assistants) is proposed. Recently, the mobile robot system has been used for various applications. The operator of robot system must keep his eyes to the robot continuously and control it by manual. Therefore the man who can know the situation and control the robot is needed to have a special skill but it is difficult to control the robot against the various cases. To control the robot system without a person who controls the robot system, the PDA, small and portable system is adopted. To generate intelligent control, the PDA system learns the various control situation and rules using the soft computing method such as neural network and fuzzy logic.

To verify the performance of the proposed system, we select the mobile robot control problem. First, we decide the communication protocol between PDA system and remote robot system. Secondly, we set the wireless communication device both PDA and controller of remote robot system. At last, we program the intelligent algorithm for robot system to navigate without collision. The intelligent algorithm also makes the robot system follow the specific target without miss.
In the point of function, the robot system has several ultra-sonic sensors and USB-camera. The PDA and robot system communicate the data such as distance, color and shape, command of user and so on. With the several data, the PDA can decide the parameters used for control of the robot system instead of the person. A person only carry the PDA system and just call the command to control the robot system then, the PDA requests the data to robot system and the PDA controls robot system with the data to accomplish the order of a person.

Inside of the PDA system, there are neural network and fuzzy logic. The neural network decide whether the robot can operate or not by learning the status of the robot system and command. From the remote robot system, the PDA obtain the status data such as the distance between robot system and the around material, camera information to know whether the target is there or not and current parameter of the motor. The PDA infers the situation of robot system and decides whether the robot system can precede the command or not. If the robot system can proceed, then the PDA provides proper parameters to the robot system. In this point, the PDA uses the fuzzy logic to decide the parameters for control of robot system. For example, we assume that a person commands robot system to go right but there is an obstacle in the right hand of robot system. The PDA can know that the right side distance of robot system is very small and the color from the camera on robot system means an obstacle. The PDA also decides for robot system not to go right and command 'keep' to robot system. Really, the PDA sends the parameters such as drive motor speed, steering angle value and turret angle to make robot system keep the current status. The above parameters are decided with the fuzzy logic. We introduce several experimental results to verify the ability of the proposed system.

2 Fuzzy Control

A. Fuzzy logic control (FLC)
The fuzzifier has the effect of transforming crisp measured data (e.g., distance is 120cm) into suitable linguistic values (i.e., fuzzy sets, for example, distance is too far). The fuzzy rule base stores the empirical knowledge of the operation of the process of the domain experts. The inference engine is the kernel of a FLC, and it has the capability of simulating human decision making by performing approximate reasoning to achieve a desired control strategy.

As mentioned above, FLC has fuzzy rule base, control rule and is composed with inference engine. The control rule of fuzzy system is the form of “IF-THEN”. The example of “IF-THEN” rule is follows

- IF X is NB and DE is PB, THEN S is NB.

The notation in the above rule is follows. X means distance, DE means the gap of distance and S means velocity.

In this paper, we use triangular method as a fuzzifier and Mamdani’s min-max method as a fuzzy inference method and center of gravity method as a defuzzifier respectively.
B. MRS Control

MRS controller controls the velocity and rotate angle to trace the target using ultrasonic sensors and USB-camera. MRS takes the distance from ultrasonic sensors and position of the target from camera and then, inferences the set parameter of the MRS with FLC. Assume that MRS must trace the target at the 60cm interval. If the distance between MRS and the target is far from the set distance, 60cm, then MRS must move forward the target. If the distance is below then, MRS must move backward the target. There are also some limit conditions such that the maximum measure distance of ultrasonic sensor is 3m and the range of view of USB camera is ±20 degrees.

3 Neuro Network Control

The proposed algorithm is constructed by the one of features of the MRS. One of the main feature is that the MRS receives current distance values for the target with ultrasonic sensors. These values are stored in the memory with the values of next time and the MRS uses all stored data for deciding the velocity. Finally, the MRS can obtain current direction and velocity of the target from Eq. (1) using the past distance value, the current distance value and the velocity of robot.

\[ V = \frac{(D_A + 10 \times \pi \times Time \times V_B - D_B)}{(10 \times \pi \times Time)} \]  

where \( V \) is current velocity of trace target, \( D_A \) is past distance between the MRS and the target, \( V_B \) is past velocity of the MRS and \( D_B \) is current distance between the MRS and the target. The term ‘10 \times \pi’ is that can convert absolute velocity being measured with ‘rps’ dimension to companion velocity of the MRS. If the result of Eq.(1) shows negative or positive we make the value, 0 or 1, respectively and this is for calculation of the next data.

To decide driving velocity, we place emphasis on two decisions of compensation velocity by distance and velocity. First, compensation velocity by distance is for preserving distance with the target. Second, compensation velocity by velocity is for reduction of the velocity margin. In the MRS, the velocity is first compensated by the distance and if the MRS traces the target at regular intervals then, compensated by velocity to keep the current state. The equation 2 shows the compensation velocity by distance Output of algorithm displays whether the velocity is changing, whether the other velocity must be applied and whether the velocity of the MRS changes from current value to limit to trace the target.

\[ V_{d,c} = (D_n - D_c) \]  

where, \( V_{d,c} \) is the distance preservation compensation velocity, \( D_n \) is current the distance and \( D_c \) is before distance.

The block diagram for neural network learning is shown in figure 1.

The target of our work is to show whether the neural network algorithm can learn the amount of the velocity variations using 3 inputs.
The 3 inputs of neural network are current distance between the MRS and target, difference between the current velocity of the target and past velocity of the MRS, and the variable ‘change’.

In the neural network, the number of hidden neuron is 20, activation function is a hyperbolic tangent and learning rate is 0.016. The back-propagation algorithm is used to update weight and the data for learning are 1600 survey data and the iteration for learning is 100000 times.

![Flow-chart of learning algorithm using neural network]

**Fig. 1.** The flow-chart of learning algorithm using neural network

### 4 Simulations and Online Implementation

The membership functions and rule bases for the inputs such as distance and difference of between the target and MRS and the output of MRS, velocity, are shown in the figure 2-4 and table 1. The figure 2 and 3 are the membership functions for distance and the difference between the target and MRS, the figure 4 is the membership function for the velocity, output of the MRS. The variables of membership functions in the fuzzy logic is decided by knowledge of the professional operator. After the membership functions are constructed with their suitable variables, the trial-error tunes them. Like the figure 2, we set the distance of the pre-condition as 7 linguistic variables, the differences of distance as 5 and last, velocity of the post-condition as 9 variables. In the figure 4, the MRS decides the velocity of MRS for the inputs such as distance and difference of distance in the pre-condition.
In the above figures, the notation NB, NM, NS, ZO, PS, PM and PB mean the linguistic degree of each parameter. The linguistic degree that is decided by expert is basis of fuzz logic. For example, if the linguistic degree is NB, NB means that the value is negative big.

**Table 1. Rule Base for Velocity**

| \( \frac{d}{de} \) | NB   | NM   | NS   | ZO   | PS   | PM   | PB   |
|---------------------|------|------|------|------|------|------|------|
| NB                  | NVB  | NVB  | NB   | NS   | ZO   | ZO   | PS   |
| NS                  | NVB  | NB   | NS   | NVS  | ZO   | PVS  | PB   |
| ZO                  | NVB  | NS   | NVS  | ZO   | PVS  | PS   | PVB  |
| PS                  | NB   | NVS  | ZO   | PVS  | PS   | PB   | PVB  |
| PB                  | NB   | ZO   | ZO   | PS   | PB   | PVB  | PVB  |

The membership functions for the fuzzy control about the rotation angle of MRS are organized as a similar method of the case of velocity.

The following figure 5 shows comparison between the survey data and the data using neural network. The result shows the experimental results of movement of the mobile robot in the cases of PID control method and neural network control method, together. The velocity change rate of PID control method is sensitive for the change of velocity but it of the neural network control method is smooth and not sensitive for the change of velocity. Consequently, the neural network make the mobile robot system trace the target at regular intervals in smooth manner.
Above mentioned, the mobile robot system having both fuzzy logic and neural network as its controller shows good ability in the performance. The fuzzy logic takes one part and the neural network takes another part. The two methods come together and the mobile robot controller can be efficiently generated. Moreover, the two methods are constructed in the remote device (PDA). By using the remote device, the user can move freely anywhere and be irrelevant to the location of the mobile robot.

5 Evaluations

The problem to control the MRS with FLC and PID controller that are implemented in the remote system (PDA) for the same inputs and outputs of the controller such as velocity and rotation angle online is explained. Both FLC and PID controller take the control inputs such as distance and position from ultrasonic sensors and PC-camera and then outputs such as velocity and rotation angle of controller to the MRS. The distance and position at the time of before and after movement of the MRS are compared with each other. We move the target in the direction of forward and backward and get the velocity output of the FLC and PID controller. The MRS moves using this velocity value and we compare the distances after movement of the MRS using FLC, PID controller respectively and set distance. The results are shown in the figure 5. In the figure 5, we know that the distance error of the MRS using FLC is
more similar with the reference distance than it of the PID controller. The figure 6 shows the result of comparison about the rotation angle when the target moves right or left side. The result in the figure 6 also shows that the operation by MRS is more efficient than by PID controller.

6 Conclusions

In this paper, we design the control system of MRS with FLC and the MRS can trace the target at regular intervals using PDA. We use the values, distance and position using ultrasonic sensors and USB camera, respectively. The simulation shows the advanced ability of performance using FLC on the point of the error of distance and position being compared PID controller. However, if there is an obstacle, not the target, within sensing area, the inference of distance has some mistakes. The proposed algorithm using neural network shows that MRS keeps the distance while tracing the target with the speed above 40 percent of the driving capacity, as long as the velocity of the target is within the absolute velocity of the robot. The result of the simulation showed that the proposed MRS is more robust to the dynamic environments compared to the common MRS.

In the future works, it is demanded to use some another sensors for recognition of various circumstance more exactly.

Acknowledgement

This paper was supported by Korea Ministry of Science and Technology under the Brain Neuroinformatics Research Program.

References

1. O. Causse and L.H.Pampagnin, “Management of a multi-robot system in a public environment,” Proc. IEEE Int. Conf. Intelligent Robots and Systems, pp. 246-252, 1995.
2. S.J.Vestli and N. Tschichold-Gurman, “MOPS, a system for Mail Distribution in Office Type Building,” Proc. IEEE Int. Conf. Intelligent Robots and Systems, pp. 486-496, 1996.
3. Z.Bien, Fuzzy Logic Control, Hong-Reung Pub., 1997.
4. K.R.S.Kodagoda, W. S. Wijesoma, and E. K. Teoh, “Fuzzy Speed and Steering Control of a AGV,” IEEE Tran. On Control Systems Technology, Vol. 10, No. 1, Jan. 2002.
5. K. S. Narendra, “Neural networks for control: Theory and practice,” Proc. of IEEE, vol. 84, No. 10, pp.1387-1406, Oct. 1996.
6. S. Haykin, Neural Networks - A Comprehensive Foundation, Macmillian College Publishing Company Inc., 1994.
7. T. Tanaka, J. Ohwi, L. V. Litvintseva, K. Yamafuji and S. V. Ulyanov, “Intelligent Control of a Mobile Robot for Service Use in Office Buildings and Its Soft Computing Algorithms,” Journal of Robotics and Mechanics, Vol.8, No. 6, pp. 1-15, 1996.
Mobile Robot for Door Opening in a House

Dongwon Kim1, Ju-Hyun Kang2, Chang-Soon Hwang3, and Gwi-Tae Park1*

1Department of Electrical Engineering, Korea University, 1, 5-Ka Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea
{upground,gtpark}@korea.ac.kr
2LG Chem. Battery Research & Development, Korea
3Intelligent Robotic Research Center, Korea Institute of Science and Technology, Korea

Abstract. A new field of robot applications is emerging. Owing to the relevancy of the robot, people’s convenience can be enhanced. In this paper, we study on the task of opening a closed door by the home service robot which is mobile robot equipped with a manipulator. The objective of this research is to successfully accomplish door opening task using the home service robot. In addition, the robot has some advantages as the home service robot for the general use. The proposed robot is small and compact because double active universal joint is adopted for the robotics joint mechanism of the manipulator so it easy to move in a house through doorways. Moreover, cheap force sensors are employed for the information of door opening task instead of expensive JR3 sensor so we can expect cost effectiveness in the robot. Practical experiment has shown that the door opening task can be completed successfully by the proposed home service robot which is small, compact, inexpensive but efficient.

1 Introduction

A new field of robot applications is emerging. Robots are today moving towards applications beyond the structured environment of a manufacturing plant. They are making their way into the everyday world that people inhabit- hospitals, offices, homes, construction sites [1-3]. In particular, the home service robot with manipulator can enhance people’s convenience by realizing the tasks such as grabbing, lifting, and manipulating home appliances. In addition, if a mobile robot passes through a doorway in the house, a working area of the robot will be expanded widely. Therefore the particular task of opening a closed door by the mobile robot equipped with a manipulator in the house is chosen to be investigated in this paper.

Previous door opening approach [4] assumes that the position and radius of the door are well known. Another door opening system has been presented, called ROMAN [7], but suspiciously, almost no information is provided as to how ROMAN opens the door [6]. In addition, the size of the robot in [4, 6, 7] were large and some expensive force/torque sensors (JR3) are mounted on the wrist to get enough information. But the home service robot must move in the house through doorways using the
wheeled mobile robot with the manipulator. So the robot should be small and compact. In addition, the robot should have an advantage in cost effectiveness for the general use. The objective of this research is to successfully accomplish a human-oriented task like door opening using the home service robot, called Hombot, which is developed as the mobile robot equipped with an anthropomorphic manipulator arm. In addition, we employed cheap three-axis force sensor as contrast with the previous approach.

2 Home Service Robot

The developed home service robot is referred to as a ‘HomBot’ and shown in Fig. 1.

![Figure of the HomBot](image)

(a) Front view  (b) Side view

**Fig. 1.** Figure of the HomBot

The configurations of the locomotion module are as follows.

- The HomBot uses the locomotion system with two driving wheels driven by two DC motors
- Vision system for the object recognition is driven by stereo camera which are mounted on the locomotion system
- LCD panel is used for the commanding, monitoring, and commenting about the state of task execution and informing people in the work space
- Twelve ultrasonic sensors are mounted around the mobile robot for the navigation
- Laser slit and web camera are mounted in front of the mobile robot for the obstacle detection

Considering the manipulation module, the configurations are as follows.

- Three-axis force sensor is located at the gripper of the manipulator
- Force and moment of the gripper are measured by the force sensor
- Each joint torque sensor is mounted on the each joint of the manipulator.
- Double active universal joint (DAUJ) is mounted in the manipulator for the compact size of the manipulator.
2.1 Characteristics of the HomBot

The proposed robot has some advantages as home service robot which should be small, compact, and inexpensive. For the cost effectiveness, three-axis force sensor is employed instead of costly JR3. The three-axis force sensor at the gripper is shown in Fig. 2. Some information such as forces along the $x$, $y$ axes of the end-effector ($f_x, f_y$) and moment about the $z$ axe of the end-effector ($m_z$) can be obtained from the three-axis force sensor. Owing to the information obtained, the compliant control for the relaxation of the end-effector is performed and door opening task is accomplished successfully. For the small and compact size of the robot, it is possible to reduce the size by adopting DAUJ [5] as joint mechanism of the manipulator. DAUJ has an internal oblique joint which controls the two-dof joint rotations. It has two geared dc motors with reduction mechanisms that drive upper half sphere and lower half sphere, respectively. Each half sphere meet in inclined plane with angle $\phi$. The inner universal joint just transmits the torque of the motor to the upper-half sphere. The outer one is a sturdy structure to connect two links, and besides it prevent the relative motions between previous and next links. The oblique rotation mechanism is a joint which rotates around an inclined axis. Further discussion on the design and kinematics can be obtained in [5].

2.2 Compliant Control of the HomBot

Compliant control of the manipulator is important for the door opening task. When the robot moves on the planned path while pulling the door, position and orientation errors of the robot may result in large forces in the end-effector. Therefore, we have to control the end-effector compliantly along the vertical plane axis of the doorknob using the force sensor shown in Fig. 2. The relaxation of the end-effector is performed by the compliant control. The compliant control is realized as follows. The force sensor located at the gripper of the manipulator detects force of the end-effector. In proportion to the amount of the force, the reference position of the end-effector is shifted to detected force direction.
3 Experimental Results

The target environment is indoors such as a house. Rooms are connected with corridors by open-sided doors which are usually closed. Every door has a rotating doorknob. When the doorknob is rotated, a prop in the door is released and then door can be opened by pulling it. Using the home service robot described above, we implemented the proposed task which is opening a door in a house. Our approach consists of the following steps.

− Step 1: The mobile robot navigates to the front of the door.
− Step 2: The end-effector of the manipulator is posing toward the doorknob and opening its gripper. Fig. 3 depicts the position of the robot and its gripper.

![Position of the manipulator](image1)
![Figure of the gripper](image2)

Fig. 3. Position the robot in front of the door and its opened gripper

− Step 3: Visual servoing is to be done in this step for grasping the doorknob by using stereo camera which is mounted on the mobile robot. Fig. 4 shows the visual servoing for the doorknob and the aligning the gripper with doorknob, respectively.

![Visual servoing and aligning procedure of the gripper with doorknob](image3)

Fig. 4. Visual servoing and aligning procedure of the gripper with doorknob
Step 4: The robot grasps and rotates the doorknob. The results of the step are shown in Fig. 5.

![Grasping doorknob and rotating doorknob](image)

**Fig. 5.** Grasping and rotating the doorknob

Step 5: The robot pulls the door a little and releases the doorknob. Fig. 6 shows the results of this step.

![Pulling the door and releasing doorknob](image)

**Fig. 6.** Pulling the door and releasing the doorknob

Step 6: The sequence of pictures in Fig. 7 shows the robot executing the door opening task successfully.

Step 7: After door opening task, the robot folds the manipulator to go through a door way. Fig. 8 shows the results of this step.
The experiment above demonstrates the usefulness of the home service robot that is mobile robot equipped with a manipulator for execution of service task.
4 Conclusion

In this paper, we proposed a door opening behavior of the mobile robot equipped with a manipulator. The robot presented in this paper has some advantages as the home service robot. First, it has a small size. We can make this robot small using the double active universal joint. Another important advantage is a cost. As contrast with the previous approach, we employed cheap three-axis force sensor only. But the door opening task is completed successfully. So we can expect cost effectiveness and feasibility of the compactness for the general use of the home service robot.

References

[1] Engelberger, J.F.: Robotics in service. Biddles, Guildford (1991)
[2] Schmidt, G., Hanebeck, U.D., Fischer C.: A mobile service robot for the hospital and home environment. in: Proceedings of the IARP Second International Workshop on Service and Personal Robots: Technologies and Applications Genova Italy (1997)
[3] Schraft, R.D., Hagele, M.: Service Robots: Examples of the state-or-art in products, prototypes, and product vision. in: Proceedings of the IARP Second International Workshop on Service and Personal Robots: Technologies and Applications Genova Italy (1997)
[4] Nagatani, K., Yuta, S.: Designing Strategy and Implementation of Mobile Manipulator Control system for Opening Door. Proc. of Int. Conf. on Robotics and Automation. (1996) 2828-2834
[5] Ryew, S., Choi, H.: Double Active Universal Joint (DAUJ): Robotic Joint Mechanism for Humanlike Motions. IEEE Trans. Robotics and Automation. 17. 3. (2001) 290-300
[6] Petersson, L., Austin, D., Kragic, D.: High-level Control of a Mobile Manipulator for Door Opening. Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (2000) 2333-2338
[7] Hanebeck, U.D., Fischer, C., Schmidt, G.: ROMAN: A Mobile Robotic Assistant for Indoor Service Applications. Proc. IROS 97 (1997) 518-525
[8] Khatib, O.: Mobile manipulation: The robotic assistant. Robot. Auton. syst. 26. (1999) 175-183
[9] Nagatani, K., Yuta, S.: Designing a behavior of a mobile robot equipped with a manipulator to open and pass through a door. Robot. Auton. syst. 17. (1996) 53-64
[10] Kang, J.H., Hwang, C.S., Park, G.T.: A Simple Control Method for Opening a Door with Mobile Manipulator. ICCAS 2003. (2003) 1593-1597
[11] Niku, S.B.: Introduction to Robotics: Analysis, Systems, Applications. Prentice hall (2001)
[12] Craig, J.J., Introduction to Robotics: Mechanics and Control. 2nd ed. Addison-Wesley (1989)
Hybrid Fuzzy-Neural Architecture and Its Application to Time Series Modeling

Dongwon Kim\(^1\), Sam-Jun Seo\(^2\), and Gwi-Tae Park\(^1\)*

\(^1\) Department of Electrical Engineering, Korea University, 1, 5-ka, Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
\{upground, gtpark\}@korea.ac.kr
\(^2\) Department of Electrical & Electronic Engineering, Anyang University, 708-113, Anyang, 5dong, Manan-gu, Anyang-shi, Kyunggi-do, 430-714, Korea

Abstract. Modeling nonlinear systems in terms of fuzzy rules often encounters a few problems such as the conflict between overfitting and underfitting, and low reliability that increases the number of the necessary fuzzy rules. To deal with these problems, we propose a hybrid fuzzy-neural modeling technique. Performance of the proposed approach is compared to that of the conventional approach for the case of forecasting the time series. Result shows that the proposed method is more efficient and accurate in terms of the number of fuzzy rules and its generalization.

1 Introduction

Fuzzy systems have proven to be powerful in many applications, including medical imaging, pattern recognition and classification, remote sensing, control systems, etc. But there are still problems to be solved in the fuzzy systems. These problems are that the conflict between overfitting and underfitting, i.e., good approximation performance does not necessarily guarantee good generalization capability, bad repetition or low reliability, i.e., for the same model, the learning results are greatly dependent on the training data sets, and requiring too many fuzzy rules for accurate function approximation, particularly in the case of multidimensional input [1].

In this paper we present a hybrid fuzzy-neural architecture in order to alleviate the above-mentioned problems in nonlinear system modeling. The hybrid fuzzy-neural technique is a combined fuzzy model, introduced by Takagi and Sugeno [2] with polynomial neural networks (PNN) [3,9]. The TS fuzzy model [2] has attracted a lot of attention in the fuzzy modeling community because of its good results in different applications and due to its mathematical treatability. In the TS fuzzy model, the optimal consequence parameters are tested for the selected premise ones with fixed premise variables until the performance criterion is minimized and the optimal parameters both in premises and consequences are obtained. However, we do not use heuristic search method to find the optimal variables. Therefore, the fuzzy model used in this

*Corresponding author.
paper is not complex. The PNN is a group method of data handling (GMDH)-type algorithm [4-6]. The GMDH introduced by Ivakhnenko in the early 1970’s is an analysis technique for identifying nonlinear relationships between inputs and outputs of a given system. One of the GMDH-type algorithms is the PNN, which provides an automated selection of essential input variables and builds hierarchical polynomial regressions and partial description (PD) of required complexity. In addition, high-order regression often leads to a severely ill-conditioned system of equations. However, the PNN avoids this problem by constantly eliminating variables at each layer. Therefore, a certain complex system can be modeled without any specific knowledge or massive amount of data of the system.

To develop a model that has high approximation capability and good generalization capability, we have developed an intelligent method by integrating fuzzy system with PNN. This hybrid model is achieved by merging fuzzy systems and PNN in one unified framework. With this model, the problems stated above can be solved. The time series problem of Box-Jenkin’s gas furnace process studied in other literature is examined to validate the proposed method.

## 2 Hybrid Fuzzy-Neural Architecture

The architecture of the hybrid fuzzy-neural model proposed in this paper can be constructed as shown in Fig. 1. This architecture is obtained by combining the ANFIS architecture with the conventional PNN with cascade connection. The fuzzy system under consideration is ANFIS [7], which is a first-order TS model.

![Fig. 1. Hybrid fuzzy-neural model for the two input variables and two fuzzy labels for each input variable](image)

When we consider some fuzzy rules within the fuzzy model, the consequent part can be expressed by linear polynomials. Moreover, we can exploit various forms of membership functions (MFs), such as triangular and Gaussian types for fuzzy set in the premise part of the fuzzy rules. These are another factors contributing to the flexibility of the proposed approach.
For simplicity, the nonlinear system to be identified is assumed to have two input variables and each input variables has two fuzzy sets. For the fuzzy model, the \(j\)-th IF-THEN rules are

Rule 1: If \(x_1\) is \(A_1\) and \(x_2\) is \(B_1\), then \(y_1 = f_1(x_1, x_2)\)

Rule 2: If \(x_1\) is \(A_1\) and \(x_2\) is \(B_2\), then \(y_2 = f_2(x_1, x_2)\)

Rule 3: If \(x_1\) is \(A_2\) and \(x_2\) is \(B_1\), then \(y_3 = f_3(x_1, x_2)\)

Rule 4: If \(x_1\) is \(A_2\) and \(x_2\) is \(B_2\), then \(y_4 = f_4(x_1, x_2)\)

where \(A_i\) and \(B_i\) in the premise part of the rules are linguistic value (such as “small” or “big”) associated with input variable \(x_i\) and \(x_2\), respectively. \(f_j(x_1, x_2)\) is the first-order consequent polynomial functions for the \(j\)-th rule. Triangle and Gaussian types for the \(A_i\) and \(B_i\) were examined.

As the PNN is applied to the hybrid fuzzy-neural architecture, its fundamentals are briefly explained. Each polynomial in the PNN algorithm represents a PD and one best model is determined by selecting the most significant input variables and polynomial order. The design procedures are detailed in [3,9]. Here, the architecture and algorithm of PNN is briefly explained. The PNN algorithm is based on the GMDH method and utilizes a class of polynomials such as linear (Type 1), quadratic (Type 2), and modified quadratic (Type 3) types. For an example, specific forms of a PD in the case of two inputs are given as

Bilinear (Type 1) = \(c_0 + c_1x_1 + c_2x_2\)

Biquadratic (Type 2) = \(c_0 + c_1x_1 + c_2x_2 + c_3x_1^2 + c_4x_2^2 + c_5x_1x_2\)

Modified biquadratic (Type 3) = \(c_0 + c_1x_1 + c_2x_2 + c_3x_1x_2\)

where \(c_i\) is called the regression coefficients.

By choosing the most significant input variables and polynomial types among various types of forms available, we can obtain appropriate PDs in each layer. The PNN is developed to identify the model of nonlinear complex systems by the use of the input-output data set. This data set is divided into two parts: the first part is the training data set and the other the testing data set. The former is used to construct a PNN model, and the latter is used to evaluate the constructed PNN model. The total number of nodes is given by a combination of a fixed number of inputs among entire input variables. Both the number of input variables and the type of polynomial of each node are defined by designer in advance. By using the chosen input variables and type corresponding to each node, we construct a PD for each node. Then we determine the coefficients of PD by the least square method by using a given training data set, and finally we obtain the estimated output of each node. Furthermore, we evaluate each PD to check its predictive capability for the output variable using the testing data set. Then we compare these values and choose several PDs which give the best predictive performance. In the sequel, we construct the second layer in the same way, considering the output variable of each chosen node in the first layer as the
new input variables to the second layer. Afterwards, we repeat this procedure until the stopping criterion (usually the predetermined number of layers) has been satisfied. Once the final layer has been constructed, only the one node characterized by the best performance is selected as the output node. The remaining nodes in that layer are discarded. Furthermore, all the nodes in the previous layers that do not have influence on the selected output node are also removed by tracing the data flow path on each layer. And finally, the PNN model is obtained.

3 Simulation Results

The well-know problem of the Box-Jenkin gas furnace that has been studied previously in [10-15] is presented by using the proposed method in this section. The time series forecasting is nonlinear and it is difficult to describe by any ordinary method, so the hybrid fuzzy-neural model proposed in this paper is proposed. The experiment results show that this model has good generalization and high reliability, and can reduce the required number of fuzzy rules.

Box and Jenkin’s gas furnace is a famous example of system modeling. The well-known Box-Jenkins data set consists of 296 input-output observations, where the input $u(t)$ is the rate of gas flow into a furnace and the output $y(t)$ is the CO$_2$ concentration in the outlet gases. The total data set is split into two parts. The first one (consisting of 148 pairs) is used for training and the remaining part of the data set serves as a testing set for the generalization capability. Since the process is dynamical, there are different values of the variables that can be considered as candidates to affect the present output $y(t)$. In most cases, designers try to find the input variables among a set of candidates variables, which play a significant role in determining the output to get a good performance. So the once identified variables make up the input variables of premise part in the fuzzy rules. As a result, this procedure leads a difficult and nonlinear programming problem with heavy computation burden. But in our case we do not use any search method to find the optimal premise variables. We just take three inputs simple and easy which are the delayed terms of methane gas flow rate $u(t)$ and carbon dioxide density $y(t)$ such as $u(t-1)$, $y(t-1)$, $y(t-2)$. For comparison, the same performance index as in [10-15] is used, which is defined as

$$PI(EPI) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

(3)

where $y_i$ is the actual system output, $\hat{y}_i$ is the estimated one, and $m$ is the number of data. PI denotes a performance index of the model for the training data set while EPI for the testing data.

The results are summarized in the figures and tables. The results of fuzzy model listed in Table 1 show that many input candidates are chosen to achieve good results and many fuzzy rules are considered for model design. On the other hand, we have to deal with computationally intensive problem that normally requires a lot of time and conflict between approximation and good generalization capability, which causes improper results and poor generalization ability. As stated in the earlier, good ap-
proximation performance does not necessarily guarantee good generalization capability. As can be seen from the Table 1, the fuzzy model is overfitted to the training data set. So the model cannot be sound balanced model for Box and Jenkin’s gas furnace process. As a result, most of fuzzy model needs the structure identification that consists of finding the input candidates of the system, determining its actual variables which affect the output, and finding out the optimal number of rules. This significant drawback can be avoided by using hybrid fuzzy-neural architecture.

Table 1. Performance index of fuzzy model

| Inputs | MF for each input | No. of rules | Performance index |
|--------|------------------|--------------|-------------------|
|        |                  | Triangular   | Gaussian          |
|        |                  | PI | EPI | PI | EPI |
| y(t-1), u(t-1) | 2, 2 | 4 | 0.196 | 0.486 | 0.196 | 0.496 |
| y(t-1), y(t-2) | 2, 2, 2, 2 | 16 | 0.011 | 0.273 | 0.007 | 0.543 |
| u(t-1), u(t-2) | 2, 2, 3, 3 | 36 | 0.005 | 4e10 | 2e-10 | 2082 |
| y(t-1), y(t-2), y(t-3) | 2, 2, 2, 2, 2 | 64 | 3e-15 | 1512 | 3e-17 | 495 |

(a) Triangular MF (23rd case in Table 2)  (b) Gaussian MF (24th case in Table 2)

Fig. 2. Performance of the best hybrid fuzzy-neural model with 3 inputs and its error

The performance results of hybrid fuzzy-neural model are summarized in Table 2. When three inputs are considered simple and easy with two membership functions assigned to each input variables, the best model is identified at either 23rd case and 24th case. Triangular MF is used for the 23rd case and its MSE is equal to PI=0.027, EPI=0.123. Gaussian MF is used for the 24th case and its corresponding MSE is PI=0.037, EPI=0.117. The generalization capability for the 24th case is slightly better than the one for the 23rd case. The corresponding actual and estimated model output, and its error are depicted in Fig. 2. Table 3 provides a comparison of the proposed model with other techniques being already proposed in the literature. The comparison is realized on the basis of the same performance index for the training and testing data set. Additionally, PI denotes a performance index of the model for the training data
set while EPI for the testing data. It is obvious that the proposed hybrid model outperforms other models both in terms of their accuracy and higher generalization capabilities though we consider small number of input variables and fuzzy rules, and no structure identification procedure.

**Table 2.** Performance index of hybrid fuzzy model in case of three inputs-eight fuzzy rules are considered

| No. of input factors | No. of fuzzy rules | Cases | No. of layers | Polynomial type | Performance index |
|----------------------|--------------------|-------|---------------|----------------|-------------------|
|                      |                    |       | 1 2-5         | Triangular MF  |                    |
|                      |                    |       | 1 2-5         | Gaussian MF    |                    |
| 1                    | 2                  | 1     | 1             | 0.136          | 0.250             |
| 2                    | 2                  | 1     | 1             | 0.089          | 0.176             |
| 3                    | 3                  | 1     | 1             | 0.149          | 0.265             |
| 4                    | 4                  | 1     | 1             | 0.216          | 0.354             |
| 5                    | 5                  | 2     | 2             | 0.110          | 0.297             |
| 6                    | 6                  | 2     | 3             | 0.132          | 0.237             |
| 7                    | 7                  | 3     | 1             | 0.247          | 0.403             |
| 8                    | 8                  | 3     | 2             | 0.082          | 0.168             |
| 9                    | 9                  | 3     | 3             | 0.136          | 0.255             |
| 10                   | 10                 | 3     | 1             | 0.042          | 0.138             |
| 11                   | 11                 | 3     | 2             | 0.032          | 0.130             |
| 12                   | 12                 | 3     | 3             | 0.038          | 0.135             |
| 13                   | 13                 | 3     | 2             | 0.080          | 0.180             |
| 14                   | 14                 | 3     | 3             | 0.040          | 0.153             |
| 15                   | 15                 | 3     | 2             | 0.046          | 0.166             |
| 16                   | 16                 | 3     | 3             | 0.054          | 0.155             |
| 17                   | 17                 | 3     | 3             | 0.040          | 0.139             |
| 18                   | 18                 | 3     | 3             | 0.053          | 0.148             |
| 19                   | 19                 | 4     | 1             | 0.041          | 0.139             |
| 20                   | 20                 | 4     | 1             | 0.031          | 0.126             |
| 21                   | 21                 | 4     | 1             | 0.036          | 0.129             |
| 22                   | 22                 | 4     | 2             | 0.048          | 0.141             |
| 23                   | 23                 | 4     | 2             | 0.052          | 0.121             |
| 24                   | 24                 | 4     | 3             | 0.048          | 0.136             |
| 25                   | 25                 | 4     | 3             | 0.031          | 0.121             |
| 26                   | 26                 | 4     | 3             | 0.034          | 0.123             |
| 27                   | 27                 | 4     | 3             | 0.034          | 0.123             |

**Table 3.** Comparison of performance index of our model with other models

| Model                | Inputs                        | No. of rules | Performance index |
|----------------------|-------------------------------|--------------|-------------------|
|                      |                               |              | PI    | EPI    |
| Sugeno’s model [10]  | y(t-1), u(t -3), u(t -4)      | 6             | 0.190 |
| Xu’s model [11]      | y(t -1), u(t -4)              | 25            | 0.328 |
| Box’s model [12]     | y(t -1), y(t -2), u(t-3), u(t-4), u(t-5) | -  | 0.202 |
| Sugeno’s model [13]  | y(t -1), y(t -2), y(t -3), u(t-1), u(t-2), u(t-3) | 2  | 0.068 |
| Kim’s model [14]     | y(t-1), y(t-2), y(t-3), u(t), u(t-1), u(t-2) | 2  | 0.034 | 0.244 |
| Lin’s model [15]     | y(t-1), y(t-2), u(t-3), u(t-5), u(t-6) | 4  | 0.071 | 0.261 |
| our model            | y(t-1), y(t-2), u(t-1)        | 8             | 0.027 | 0.123 |
|                      |                               | 8             | 0.037 | 0.117 |
4 Conclusions and Future Research

In this paper, a hybrid fuzzy-neural architecture based on the fuzzy system and polynomial neural network in forecasting time series has been developed. We have described its potential architectural variations, and have applied the hybrid model to time series modeling such as Box-Jenkin’s gas furnace process. Result suggests that the proposed modeling technique performs well. The main advantages of the model are: (1) any conflict between overfitting and underfitting can be avoided, (2) high reliability can be achieved, and (3) the number of fuzzy rules can be reduced compared to the conventional fuzzy system. The future work will be focused on several factors such as simpler network architecture, an optimal network topology, a global optimization technique such as genetic algorithm.

References

1. Chen, J.Q. and Xi, X.G.: Nonlinear System Modeling by Competitive Learning and Adaptive Fuzzy Inference System. IEEE Trans. Syst., Man, Cybern. 28 (1998) 231-238
2. Takagi, T. and Sugeno, M.: Fuzzy Identification of Systems and Its Applications to Modeling and Control. IEEE Trans. Syst., Man, Cybern. 15 (1985) 116-132
3. Oh, S.K., Kim, D.W. and Park, B.J.: A Study on the Optimal Design of Polynomial Neural Networks Structure. Trans. KIEE. 49D (2000) 145-156
4. Ivakhnenko, A. G.: Polynomial theory of complex systems. IEEE Trans. Syst. Man Cybern. (1971) 364-378
5. Ivakhnenko, A. G., Krotov, G. I. And Ivakhnenko, N. A.: Identification of the mathematical model of a complex system by the self-organization method. in Theoretical Systems Ecology: Advances and Case Studies, E. Halfon, Ed. New York: Academic (1970)
6. Farlow, S. J.: Self-Organizing Methods in Modeling, GMDH Type-Algorithms, New York: Marcel Dekker (1984)
7. Jang, J. S.: ANFIS: Adaptive-Networks-Based Fuzzy Inference System. IEEE Trans. Syst., Man, Cybern. 23 (1993) 665-685
8. Jang, J.S., Sun, C.T. and Mizutani, E.: Neuro-Fuzzy AND Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall (1997)
9. Oh, S. K. and Pedrycz, W.: The design of self-organizing Polynomial Neural Networks. Inf. Sci. 141 (2002) 237-258
10. Sugeno, M. and Yasukawa, T.: A fuzzy-logic-based approach to qualitative modeling. IEEE Trans. Fuzzy Syst. 1 (1993) 7-31
11. Xu, C.W. and Zailu, Y.: Fuzzy model identification self-learning for dynamic system. IEEE Trans. Syst., Man, Cybern. 17 (1987) 683-689
12. Box, G.E.P. and Jenkins, F.M.: Time Series Analysis: Forecasting and Control 2nd ed. Holden-day (1976)
13. Sugeno, M. and Tanaka, K.: Successive identification fo a fuzzy model and its applications to prediction of a complex system. Fuzzy Sets Syst. 42 (1991) 315-334
14. Kim, E., Lee, H., Park, M. and Park, M.: A simple identified Sugeno-type fuzzy model via double clustering. Inf. Sci. 110 (1998) 25-39
15. Lin, Y. and Cunningham III, G.A.: A new approach to fuzzy-neural modeling. IEEE Trans. Fuzzy Syst. 3 (1995) 190-197
Accelerometer Signal Processing for User Activity Detection

Jonghun Baek\textsuperscript{1}, Geehyuk Lee\textsuperscript{2}, Wonbae Park\textsuperscript{1}, and Byoung-Ju Yun\textsuperscript{1}

\textsuperscript{1} Dept. of Information and Communications, Kyungpook National University, Daegu 702-701, South Korea
afqb@korea.com
\textsuperscript{2} School of Engineering, Information and Communications University, Daejeon 305-714, South Korea

Abstract. Estimation of human motion states is important enabling technologies for realizing a pervasive computing environment. In this paper, an improved method for estimating human states from accelerometer data is introduced. Our method for estimating human motion state utilizes various statistics of accelerometer data, such as mean, standard variation, skewness, kurtosis, eccentricity, as features for classification, and is expected to be more robust than other existing methods that rely on only a few simple statistics. A series of experiments for testing the effectiveness of the proposed method has been performed, and its result is presented.

1 Introduction

The concept of a wearable computer was invented to offer the user a portable and personalized user interface to computing resources, such as multimedia services and home automation services\textsuperscript{1,2}. One of the most distinguishing features of a wearable computer is that it should not expect to be attended to but should attend to the user. As a result, any sensible wearable computer should have the skill for sensing the user’s states and intentions. In this paper, we propose an improved method for probing the user’s activities by using an accelerometer. Information on the user’s activities can help a wearable computer provide a suitable service to the user at the right time. For instance, a PDA, which can provide a multimedia service, may pause and wait while the user is running\textsuperscript{3}. In addition to such a purpose, the same method can be used to estimate energy expenditure\textsuperscript{4}, or, by sensor fusion with other data, for estimation of the location and the health condition of the user.

Some of the related researches are Randell & Muller\textsuperscript{3}, Farringdon, et al.,\textsuperscript{5}, and Schmidt, et al.,\textsuperscript{6}. Most of the cases they used 1- or 2-axis accelerometers for sensing movements, average, RMS( Root Mean Square) and integrated values for feature values, and a neural network for classification. The accuracy of the activity estimators in their studies was around 85 $\sim$ 90\%, which may be good enough for some applications, but may not for others. Clearly there is
room for improvement in their results because they made use of the low-order
statistics only. We began our research in the hope to enhance the accuracy of
an activity detector by including higher order statistics of accelerometer signals
than RMS and integrated values. We investigated the characteristics of the output
from a two-axis accelerometer in different activity conditions by calculating
a histogram, and determined which features to use for the classification of the
states of the user. In the classification step, we used a feed-forward neural net-
work that accepts a set of statistics of an accelerometer signal and outputs a
motion state of the user. An accelerometer signal is thereby classified into one of
the eight motion states; standing, sitting, lying-back, lying-on, walking, running,
upstairs, and downstairs.

2 Signal Processing Steps

A single run of accelerometer data is a two-dimensional time series collected
for about two seconds at the sampling rate of 15 samples/s from the outputs
of a two-axis accelerometer (ADXL202EB) worn by the user on the waist. The
X and Y axes of the accelerometer are pointing the upward and forward di-
rection, respectively. As shown in Fig. 1, a feature extractor calculates a set
of statistics of the data, which in turn forms a feature vector to be input to a
classifier. The output of the classifier is one of the labeled classes shown in the
figure.

Fig. 1. Signal processing steps for motion state estimation

Fig. 2. (a) Orientation of the accelerometer in the four static states (standing, sitting,
lying back, and lying on), and (b) typical distribution of accelerometer data in the four
dynamic states (walking, running, upstairs, and downstairs)
5 Five Pairs of Features

As shown in Fig. 2a, the four static states can be easily distinguished because the orientation of an accelerometer is different. On the other hand, it is not easy to classify the dynamic states by using only the direction of the accelerometer because the orientations of the accelerometer are almost the same. As shown in Fig. 3, the histograms of acceleration data belonging to different classes exhibit unique characteristics in their shapes. From the observations, we selected the following statistics as signal features: mean, standard deviation, skewness, kurtosis, and eccentricity.

Mean and standard deviation are simple but the most useful features for distinguishing static states from dynamic states and also for the classification of static states from each other as shown in Fig. 4. The last two plots in Fig. 4 assert the difficulty of classifying dynamic states by using means and standard deviations alone.

Skewness is the degree of asymmetry in the distribution of acceleration data, and kurtosis is a measure that reflects how much a distribution is peaked at the center of a distribution.

As shown in Fig. 5, the skewness in X-acceleration can distinguish walking/running states from upstairs/downstairs states, the skewness in Y-acceleration can distinguish among walking/upstairs states from running state, and the kurtosis in X-acceleration can distinguish upstairs/downstairs states from walk-
ing/running states. It is not clear whether the kurtosis in Y-acceleration will be effective in classifying any states.

The statistics discussed so far are calculated for each of X and Y acceleration data independently while eccentricity is the feature of vector accelerations in the X-Y plane. As shown in Fig. 6, each of dynamic states exhibits different degree of eccentricity and also different major axes. As a means of expressing eccentricity mathematically, one can calculate the covariance matrix of X & Y acceleration data and use the two eigenvalues of the matrix, $e_1$ and $e_2$.

Eccentricity in the case of static states is not very meaningful because the trajectory of vector accelerations forms a point in the X-Y plane. As shown in Fig. 6b and c, eccentricity is expected to be effective in distinguishing walking/upstairs states from running states. Contrary to our expectation, eccentricity is not as useful as other simple statistics. We expect that eccentricity may be proved to be more useful in identifying walkers rather than distinguishing different states of the same user.

4 Multi-layer Perceptron Classifier

Figure 7 summarizes the expected role of the five pairs of features in the classification of acceleration signals into eight different classes. Although we used a neural network for classification in the current work, Fig. 7 suggests a possibility for a more efficient method for classification. For instance, a standard deviation

Fig. 5. Distribution of the skewness for (a) X, (b) Y and kurtosis for (c) X, (d) Y

Fig. 6. (a) Trajectories of vector acceleration data for the four dynamic states, and the distribution of the eigenvalues: (b) $e_1$ and $e_2$
Table 1. Definitions of features: \( N \) is a total number of data samples, \( c_i \) is the center value for interval \( i \) of the histogram, \( n_i \) is the number of data samples in interval \( i \), \( k \) is the number of intervals, \( x \) and \( y \) are the output values of the accelerometer, \( m_x \) and \( m_y \) are the means of \( X \) and \( Y \) acceleration, respectively.

| Features       | Definitions                                                                 |
|----------------|-----------------------------------------------------------------------------|
| Mean           | \( m = \frac{1}{N} \sum_{i=1}^{k} c_i \cdot n_i \)                         |
| Standard deviation | \( \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} (c_i - m)^2 \cdot n_i} \)   |
| Skewness       | \( \text{Skewness} = \frac{1}{N\sigma^3} \sum_{i=1}^{k} (c_i - m)^3 \cdot n_i \) |
| Kurtosis       | \( \text{Kurtosis} = \frac{1}{N\sigma^4} \sum_{i=1}^{k} (c_i - m)^4 \cdot n_i \) |
| Eccentricity   | Represented by the two eigenvalues of covariance matrix \( C \):             |
|                | \( C = \begin{bmatrix} \sum_{i=1}^{N} (x_i - m_x)^2 & \sum_{i=1}^{N} (x_i - m_x)(y_i - m_y) \\ \sum_{i=1}^{N} (x_i - m_x)(y_i - m_y) & \sum_{i=1}^{N} (y_i - m_y)^2 \end{bmatrix} \) |

is all that is needed to distinguish static states from dynamic states. If a standard deviation indicates the user is in a static state, there is no need to calculate any other statistics except a mean. This kind of reasoning leads naturally to a simple decision tree, which will be obviously more efficient but is less robust than the neural network classifier. Combination of a decision tree approach and a neural network approach may give a better solution but is not pursued in the current work.

The classifier is a multi-layer perceptron with 10 input nodes (5 pairs of features), a single hidden layer of 12 sigmoidal units, and an output layer of 8 sigmoidal units (8 activity states). The number of hidden layers and the number of units in the hidden layers were determined by a series of experiments. As usual, we had to consider trade-off between a training error and an overfitting problem in the determination of the number of units in the hidden layer.

![Fig. 7. Classification into the eight activity states by the five pairs of features](image-url)
Table 2. Tabular representation of the input and output relationship of the MLP classifier: Features are the input of the classifier, and user motion states are the output of the classifier, respectively. Numbers are mean feature values of the training data in the respective motion states.

| Features | MLP Classifier | User Motion States |
|----------|----------------|--------------------|
|          | F Mean X       | Stand Sit Lying-back Lying-on Walk Run Upstairs Downstairs |
| -0.92    | -0.8           | -0.19 -0.06 -0.94 -0.94 -0.92 -0.94 |
| e Mean Y | 0.8            | 1.2 -0.1 1.65 0.74 0.67 0.62 0.76 |
| a Standard Variation X | 0.01 | 0.01 0.01 0.01 0.26 0.07 0.28 0.43 |
| t Standard Variation Y | 0.01 | 0.01 0.02 0.02 0.23 0.28 0.24 0.24 |
| u Skewness X | 0 | 0 0 0 0.44 0.4 0.61 0.64 |
| r Skewness Y | 0 | 0 0 0 0.52 0.41 0.57 0.45 |
| e Kurtosis X | 0 | 0 0 0 1.43 1.35 1.69 1.65 |
| s Kurtosis Y | 0 | 0 0 0 1.4 1.39 1.56 1.5 |
| Eccentricity e1 | 0.02 | 0.03 0.05 0.05 9.84 17.2 8.65 10.8 |
| Eccentricity e2 | 0.24 | 0.1 0.28 0.19 20.4 85.7 26.5 51.7 |

5 Experimental Results

We restricted the goal of the proposed method to person-dependent estimation of activity, and gathered all the acceleration data from a single subject. We collected 30 runs of training data for each of the eight activity states, and 30 runs of testing data also for each of the eight activity states. Each run of data consists of a series of X and Y acceleration data recorded for two seconds. Training of the network with the training data was done without any problem (in 300 epochs), and the classification of the test data by so trained network was quite successful. As shown in Table 3a, the classification results were perfect; there was no misclassification.

Of course, we are aware of the limitation of the results. First of all, all the data for training and testing were from a single subject. Moreover, the data were collected on the same day and at the same place. In order to show the robustness of the proposed scheme, we had to collect more diverse data from different days and different places. As the minimal efforts to supplement the results, we collected another set of test data recently (about half a year after the first experiment) in a different building, and tested the same network that was used in the previous experiment (the network trained by the previous training data). The results are summarized in Table 3b. The correct classification rate was 100% for static, walking, and running states, and dropped to 87% and to 93% for downstairs and upstairs, respectively. The classifier misclassified upstairs cases as running cases possibly because the subject was moving upstairs faster than before.

In spite of such misclassifications, the overall rate of correct classification was about 97.5%. This is a clear improvement on the previous results by other research groups; the accuracy of inference for motion states of the user using a 2-axis accelerometer was around 85 ∼ 90%.
Table 3. Classification results by the proposed scheme: (a) results from the first experiment and (b) results from the second experiment using the same network that was trained in the first experiment but using a new set of test data obtained in a different building about half an year later. Symbols in the table: S1 for standing, S2 for sitting, S3 for lying-back, S4 for lying-on, S5 for walking, S6 for running, S7 for upstairs, and S8 for downstairs.

|     | Classification |
|-----|----------------|
|     | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 |
| O   | 30 | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| r   | 0  | 30 | 0  | 0  | 0  | 0  | 0  | 0  |
| i   | 0  | 0  | 30 | 0  | 0  | 0  | 0  | 0  |
| g   | 0  | 0  | 0  | 30 | 0  | 0  | 0  | 0  |
| i   | 0  | 0  | 0  | 0  | 30 | 0  | 0  | 0  |
| g   | 0  | 0  | 0  | 0  | 0  | 30 | 0  | 0  |
| n   | 0  | 0  | 0  | 0  | 0  | 0  | 30 | 0  |
| a   | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 30 |
| l   | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

6 Conclusion

We designed a user activity estimator that can classify acceleration data into eight motion states: standing, sitting, lying back, lying on, walking, running, upstairs, and downstairs. In the first experiment, we used data collected from a single subject on the same day and at the same place, and the classification results were perfect. In the second experiment performed about a half year later in a different building, the results were quite satisfactory though not perfect. Although more extensive and objective comparison is still desired, the results indicated an improvement on the other schemes proposed by other research groups. We believe that the use of higher order statistics such as skewness and kurtosis must have been the main reason for the improved performance.

Obviously, the current scheme has a lot of limitations that need further endeavors. First of all, the current experiments were aimed at person-dependent activity detection. Second, the second experiment revealed the fact that the acceleration data even from the same person can vary over time and at varying places. Third, the ideal assumption that the user will wear an accelerometer on the waist is not realistic or practical in the real application settings. We are currently working on improving the proposed scheme to better deal with such problems, and at the same time on practical implementation of the same scheme for use with small computers with limited computing resources.

References

1. S. Lee, A ubiquitous IT revolution strategy of a foreign country, http://www.etimesi.com, KOREA IT NEWS, 2002
2. Wearable computer, http://korea.internet.com, Korea.internet.com, 2001
3. C. Randell and H. Muller, Context awareness by analyzing accelerometer data, The Metadata International Symposium on Wearable Computers, 2000, pp. 175–176
4. M. Sekine, T. Tamura, T. Fujimoto and Y. Fukui, Classification of walking pattern using acceleration waveform in elderly people, Annual EMBS International Conference, In Engineering in Medicine and Biology Society, 2000, pp. 1356–1359
5. J. Farringdon, A.J. Moore, N. Tilbury, J. Church and P.D. Biemond, Wearable sensor badge & sensor jacket for context awareness. In Proceedings of The Third International Symposium on Wearable Computers, 1999, pp. 107–113
6. A. Schmidt, H.W. Gellersen and M. Beigi, A wearable context-awareness component. Proceedings of The Third International Symposium on Wearable Computers, 1999, pp. 176–177
Neural Network Models for Product Image Design

Yang-Cheng Lin¹, Hsin-Hsi Lai¹, and Chung-Hsing Yeh²

¹ Department of Industrial Design, National Cheng Kung University, Tainan, 701, Taiwan
   p3890104@ccmail.ncku.edu.tw
   hsinhsi@mail.ncku.edu.tw
² School of Business Systems, Monash University, Clayton, Victoria 3800, Australia
   ChungHsing.Yeh@infotech.monash.edu.au

Abstract. This paper develops four neural network models to help product developers work out a combination of product form elements for best matching a given product image. By applying four most widely used rules for determining the number of hidden neurons, these four models can be used to determine the value of the product image for a given combination of product form elements. An experimental study on mobile phones is conducted to evaluate the performance of these four models. The result of this study shows that there is no best rule for building the models and the performance of these models does not differ significantly. Although the mobile phones are chosen as the object of the experimental study, the approach presented is applicable to other products where a combination of form or other design elements is to be determined for matching a desirable product image.

1 Introduction

The image of a product plays an important role in consumers’ preference and choice of the product [1]. Whether consumers choose a product, depends largely on their perception of the product image [2]. Based on the relationship between the product form and the product image perceived by consumers, design support models [4, 5] and consumer-oriented technologies [11] have been developed to help designers design product form for a given product image. In particular, Kansei Engineering [12] has been developed as “translating technology of a consumer’s feeling (Kansei in Japanese) and image of a product into design elements”. However, the consumers’ preference and perception of product form is a complex problem. Moreover, the design information is difficult to obtain, particularly the consumers’ preference for a product. With their effective learning ability [13], neural networks can examine a product design without explicitly catching on to what the complex relationship between form elements and product image is. As such, in this paper we use neural network models to determine the best combination of form elements for a given design concept of product image. We use the concept of Kansei Engineering [7, 8] to collect numerical data for building neural network models.
In developing neural network models for product image design, we will address the following two research questions. Are neural network models really suitable for applying to the product design field? Which is the best rule to determine the number of hidden neurons in a hidden layer? To illustrate how the form elements of product design can be examined, we conduct an experimental study of mobile phones for their popularity as a consumer product and their wide variety of product form elements.

2 Morphological Analysis of Product Form Elements

In our experimental study, we used the concept of Kansei Engineering [7, 8] to extract the representative experimental samples, including representative mobile phones samples, and the morphological analysis. We focused on investigating and categorizing of various mobile phones in the consuming market. 33 mobile phones were used as representative experimental samples, as reported in our prior study [3, 9]. The 33 representative experimental samples included 28 training samples and 5 test samples for training and testing the neural network models.

We used the morphological analysis [5] to extract the form elements of mobile phones by surveying product design experts. In our study, the product form included not only the outline shapes, but also the product elements. The morphological analysis involved two primary processes. In the first process, 5 experts were asked to write down the influential form elements of the mobile phones individually according to their knowledge and experience. The survey result was separated into two parts, the “form feature” and the “form treatment” [1]. The “form feature” indicated the size and shape of outline components that make up the mobile phones, such as buttons, icons, a screen or a body shell. The “form treatment” described the relationship between the outline components, for example, the equidistance arrangement of the buttons or the size rate of the screen and the body shell.

In the second process, the same 5 experts formed a focus group [14] to discuss and examine the survey result. The similar components or opinions were combined or discarded. Table 1 shows the result of the morphological analysis. There are 9 form elements of mobile phones, including “Top Shape”, “Body Shape”, “Bottom Shape”, “Length and Width Ratio of Body”, “Function Buttons Style”, “Number Buttons Arrangement”, “Screen Size”, “Screen Mask and Function Buttons”, and “Outline Division Style”.

To collect numerical data about the relationship between the product image “Simple-Complex (S-C)” and 9 form elements of mobile phones, 15 subjects were asked to perform the experimental evaluation of Kansei Engineering. The “Simple-Complex (S-C)” image was adopted, because the “Simple-Complex” image had the highest predictive consistency, as compared to the “Handsome-Rustic” and “Leisure-Formal” images, based on the result of our prior study [3, 9]. The 15 subjects used seven scales (1-7) of the semantic difference (SD) method [15] to evaluate 33 small paper cards of representative mobile phones for their value of the S-C image, with 1 being the simplest and 7 being the most complex. The average evaluation values of the S-C image were used for the analysis of neural network models in the following section.
Table 1. Extracted form elements of mobile phones

| Items                          | Type 1            | Type 2            | Type 3            | Type 4            |
|-------------------------------|-------------------|-------------------|-------------------|-------------------|
| 1. Top shape                  | Line              | Curve             | Arc               | Irregular         |
| 2. Body shape                 | Parallel line     | Raised curve      | Concave curve     |                   |
| 3. Bottom shape               | Line              | Curve             | Arc               |                   |
| 4. Length and width ratio of body | Wide ratio 2:1 | Middle ratio 2.5:1 | Slender ratio 3:1 |                   |
| 5. Function buttons style     | With large button | Symmetry style    | Other style       |                   |
| 6. Number buttons arrangement | Regular           | Irregular         |                   |                   |
| 7. Screen size                | TV ratio 4:3      | Movie ratio 16:9  | Other ratio       |                   |
| 8. Screen mask and function buttons | Independence | Function buttons dependence on screen mask | Interdependence |                   |
| 9. Outline division style     | Normal division   | Rim division      | Special division  |                   |
3 Experimental Analysis of Neural Network Models

Neural networks (NNs) are non-linear models and are widely used to examine the complex relationship between input variables and output variables [13]. Due to the effective learning ability, NNs have been applied successfully in any amount of fields [6, 10, 16, 17]. NNs have several different kinds of algorithms. The back-propagation NN is an effective and most frequently used supervised learning algorithm [13]. This network belongs to "the feed forward neural networks". A multilayer network normally composes of many neurons in several layers. The typically back-propagation NN has an input layer, an output layer, and at least one hidden layer. There are two modes of operation for the back-propagation NN - the training process and the test process. During the training process, the NN takes the training samples and then checks the training process repeatedly. The test process is the prediction by evaluating the possible result of another sample set, based on the training result. The back-propagation NN is a gradient descent algorithm to improve its performance by reducing the total errors when changing the weights along its gradient [13].

Based on the notion that the process of product form design or the perception of the consumers is a “black box” or often involves “uncertain information”, we use NNs to evaluate the product image of form elements in product design. To this end, we develop four NN models to examine and predict the product image of form design, called NN-1, NN-2, NN-3, and NN-4 respectively.

The 27 types of the 9 form elements in Table 1 are used as the 27 input variables for the four NN models. For all input variables, we use the type value of 1 to indicate that the input variable is characterized by the type of the corresponding form element. If the input variable has no the type characterization, its type value is set to 0. The evaluation average values of the S-C image are used as the output neuron. The number of neurons of the hidden layer is dependent on the ideal solution of each model, and is different in each case. The four NN models are developed by using the following four most widely used rules [13] for determining the number of neurons of the hidden layer respectively:

\[(\text{The numbers of input neurons} + \text{the numbers of output neurons}) / 2\]  \hspace{1cm} (1)
\[(\text{The numbers of input neurons} \times \text{the numbers of output neurons})^{0.5}\]  \hspace{1cm} (2)
\[(\text{The numbers of input neurons} + \text{the numbers of output neurons})\]  \hspace{1cm} (3)
\[(\text{The numbers of input neurons} + \text{the numbers of output neurons}) \times 2\]  \hspace{1cm} (4)

Table 2 lists the neurons of the four NN models, including the input layer, hidden layer, and output layer. In addition, the learning rule used is Delta-Rule and the transfer function is Sigmoid [13].

Totally 28 experimental samples were used as the training samples. Each model trained 5,000 epochs at each run. When the cumulative training epochs were over 50,000, the training process was completed. Table 3 shows the training epochs of each
Table 2. Neurons of four NN models

| NN model | Input layer: 27 neurons for 27 types of the 9 form elements. Output layer: 1 neuron for the S-C image. |
|----------|-------------------------------------------------------------------------------------------------------------------------------------|
| NN-1     | Hidden layer: 14 neurons, \((27+1)/2=14\).                                                                                     |
| NN-2     | Hidden layer: 5 neurons, \((27*1)^{0.5}=5.20 \approx 5\).                                                                 |
| NN-3     | Hidden layer: 28 neurons, \((27+1)=28\).                                                                                       |
| NN-4     | Hidden layer: 56 neurons, \((27+1)*2=56\).                                                                                     |

Table 3. RMS errors of four NN models

| Learning epochs | NN-1   | NN-2   | NN-3   | NN-4   |
|----------------|--------|--------|--------|--------|
| 5000           | 0.0755 | 0.1181 | 0.1180 | 0.1135 |
| 10000          | 0.1115 | 0.1175 | 0.1423 | 0.0862 |
| 15000          | 0.1055 | 0.0924 | 0.0965 | 0.0767 |
| 20000          | 0.0617 | 0.0612 | 0.0681 | 0.1068 |
| 25000          | 0.0956 | 0.0701 | 0.0622*| 0.0985 |
| 30000          | 0.0575 | 0.0457*| 0.0740 | 0.0834 |
| 35000          | 0.0581 | 0.0676 | 0.1046 | 0.1026 |
| 40000          | 0.0707 | 0.0488 | 0.0975 | 0.0648*|
| 45000          | 0.0430*| 0.0662 | 0.0951 | 0.0939 |
| 50000          | 0.0658 | 0.0611 | 0.0709 | 0.1004 |

model run and their corresponding root of mean square (RMS) errors, and the lowest RMS error of each model is asterisked. As shown in Table 3, the NN-1 model has the lowest RMS error (0.0430) among all four models.

4 Verification of Neural Network Models

In order to examine if each model can be applied to new samples, the 5 test samples are used. The 15 subjects were involved in the test process, using the semantic difference method [15] with a 7-point scale (1-7). To choose among these four models, we calculate their corresponding RMS errors to compare the performance of each model. Table 4 shows the result.
Table 4. Prediction values and RMS Errors of four models

| Test sample no.         | 1     | 2     | 3     | 4     | 5     | RMS errors |
|-------------------------|-------|-------|-------|-------|-------|------------|
| Consumer perception     | 6.400 | 6.600 | 2.733 | 4.533 | 4.600 |            |
| The NN-1 model          | 6.529 | 7.158 | 4.227 | 5.364 | 2.184 | 0.1839     |
| The NN-2 model          | 5.960 | 6.969 | 4.107 | 5.275 | 1.967 | 0.1899     |
| The NN-3 model          | 6.437 | 7.299 | 4.310 | 5.262 | 2.655 | 0.1647*    |
| The NN-4 model          | 6.562 | 7.313 | 4.392 | 5.303 | 2.592 | 0.1715     |

Table 4 shows that the RMS error of the NN-3 model (0.1647) is the smallest among these models. In other words, the NN-3 model has the highest predictive consistency on the “Simple-Complex” image. This test result is not consistent with the training result, where the NN-1 model has the best performance. This suggests that there is no best rule for determining the number of neurons of the hidden layer for building NN models for product image design. In addition, the RMS errors of the NN-1 (0.1839), NN-2 (0.1899), NN-3 (0.1647), and NN-4 (0.1715) differ slightly. This seems to suggest that the number of neurons of the hidden layer do not have significant impact on the predictive ability of the model.

5 Conclusion

In this paper, we have presented a new approach for examining the relationship between product image and product form elements, with an experimental study of mobile phones. We have used the concept of Kansei Engineering to extract the representative experimental samples of mobile phones. We have developed four NN models to predict the value of the product image “Simple-Complex” for a given set of form elements. The verification of four NN models shows that there is no best rule for determining the number of neurons of the hidden layer. In practice, these four NN models should be built and tested, and the one with the best test result should be used for product image design. The chosen NN model would help the product designers better understand consumers’ perception of product form with respect to the corresponding product image. The result of this paper provides useful insights in designing form elements of a product for matching the image of the product. Although the mobile phones are chosen as the object of the experimental study, the approach can be applied to other products.
References

1. Chuang, M.C., Chang, C.C., Hsu, S.H.: Perceptual Elements Underlying User Preferences Toward Product Form of Mobile Phones. International Journal of Industrial Ergonomics 27 (2001) 247-258
2. Chuang, M.C., Ma, Y.C.: Expressing the Expected Product Images in Product Design of Micro-Electronic Products. International Journal of Industrial Ergonomics 27 (2001) 233-245
3. Guan, S.S., Lin, Y.C.: A Study on the Color and Style Collocation of Mobile Phones Using Neural Network Method. Journal of the Chinese Institute of Industrial Engineers 18 (2001) 84-94
4. Hsiao, S.W., Chen, C.H.: A Semantic and Shape Grammar Based Approach for Product Design. Design Studies 18 (1997) 275-296
5. Hsiao, S.W., Liu, M.C.: A Morphing Method for Shape Generation and Image Prediction in Product Design. Design Studies 23 (2002) 497-513
6. Hsu, C.H., Jiang, B.C., Lee, E.S.: Fuzzy Neural Network Modeling for Product Development. Journal of Mathematical and Computer Modeling 29 (1999) 71-81
7. Ishihara, S., Ishihara, K., Nagamachi, M.: An Automatic Builder for a Kansei Engineering Expert System Using Self-Organizing Neural Networks. International Journal of Industrial Ergonomics 15 (1995) 25-37
8. Kashiwagi, K., Matsubara, Y., Nagamachi, M.: A Feature Detection Mechanism of Design in Kansei Engineering. Human Interface 9 (1994) 9-16
9. Lin, Y.C., Lai, H.H., Guan, S.S.: A Study of Form and Color Collocation of Mobile phones Using Kansei Methods. 01’ International Conference on Advanced Industrial Design (ICAID), Taiwan, (2001) 137-142
10. Liu, P.X., Zuo, M.J., Meng, M.: Using Neural Network Function Approximation for Ideal Design of Continuous-State Parallel-Series Systems. Computers and Operations Research 30 (2003) 339-352
11. Nagamachi, M.: Kansei Engineering As a Powerful Consumer-Oriented Technology for Product Development. Applied Ergonomics 33 (2002) 289-294
12. Nagamachi, M.: Kansei Engineering: A New Ergonomics Consumer-Oriented Technology for Product Development. International Journal of Industrial Ergonomics 15 (1995) 3-10
13. Nelson, M.: Illingworth WT. A Practical Guide to Neural Nets. Addison-Wesley, New York (1991)
14. Nielsen, J.: Usability Engineering. AP Professional, New York (1993)
15. Osgood, C.E., Suci, C.J.: The Measurement of Meaning. Urbana: University of Illinois Press (1957)
16. Smith, K., Palaniswami, M., Krishnamoorthy, M.: A Hybrid Neural Approach to Combinatorial Optimization. Computers and Operations Research 23 (1996) 597-610
17. Smith, K., Gupta, J.: Neural Networks in Business: Techniques and Applications for the Operations Researcher. Computers and Operations Research 27 (2000) 1023-1044
Evaluation of Users’ Adaptation by Applying LZW Compression Algorithm to Operation Logs

Hiroshi Hayama and Kazuhiro Ueda

University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo, Japan
hiros-h@rogue.co.jp, ueda@gregorio.c.u-tokyo.ac.jp

Abstract. In this article we present a simple and easy-to-implement method for evaluating users’ adaptation to software. This method is based on the LZW compression algorithm. Novice users have tendencies to repeat redundant operations that have certain patterns. In the LZW compression algorithm, the compression rate represents the rate of repetition that consists of patterns, or sequences of some characters. By applying the LZW compression to users’ operation logs, effectiveness of operation or degree of acquirement can be evaluated from the compression rate. Moreover, it is possible to distinguish incorrect operations from purposive operations by investigating the string table that is used to replace input characters to a single code along the compression process.

1 Introduction

Over the past decades, many adaptive/adaptable user interfaces have been developed (e.g. Split Menus[1], PoBox[2], [3], etc.). In that sort of user interfaces, operation histories are often collected and analyzed to evaluate users’ adaptation or to predict users’ next operation. In order to make adaptive/adaptable user interfaces more effective, it is necessary to develop appropriate user modeling system. We present a simple method based on LZW compression algorithm. It has a great advantage because its implementation is extremely easy.

1.1 Background

Hayama & Ueda[4] showed the effects of presenting the differences between old version of software and new version of that one. They developed a schedule management system which presented small icons that indicated the differences between old and new versions on the left side of menu items (Fig.1). They conducted usability tests and it was shown that the difference presentation method facilitated users’ adaptation to the upgrade of the software. However, to erase the icons that show differences, it is necessary to monitor users’ operations and evaluate users’ adaptation. The reason is as follows. That is, as the software is upgraded again and again, its menus will become full of that sort of icons.
2 Evaluating Users’ Adaptation by Using LZW Compression Algorithm

2.1 Analysis of Operation Logs

To examine the characteristics of users’ adaptation, we drew state transition diagrams from users’ operation logs of the above experiments. From the diagrams, it was shown that low-performance subjects frequently went into “loops” that contained both correct operations and a lot of incorrect operations. They did not repeat random operations, but repeated redundant operations that have certain patterns. As they became expert, redundant operations decreased, and the “loops” had rarely appeared (Figs. 2-4).

We noticed that it would be able to use Lempel-Ziv (LZW) compression algorithm [5] to extract those patterns from the operation logs and to estimate the degrees of users' adaptation. There are two other kinds of compression algorithm: Run-Length method and Huffman method, and they are often combined with LZW algorithm to increase the compression rate. However, they do not extract the sequences that consist of various characters. In Run-Length method, short codes are given to the
sequences that consist of same characters. In Huffman method, short codes are given to the characters that appear frequently. So we adopt “pure” LZW compression algorithm. The compression rate reflects the redundancy, or the rate of repeated operations. We can also extract each sequence from operation logs by inspecting the string table (The string table is used to convert each sequence of strings to a single code). And we can extract repeated sequence by comparing string table and output (compressed data).

2.2 Evaluation of Adaptation by Compression Rate

Table 1 shows the procedure to make a string table and to output compressed data (a sample of LZW algorithm itself). Here, we define “compression rate” as (input character numbers) / (output character numbers) * 100. So, the larger the compression rate is, the smaller the redundancy is.

We encoded each command of the operation logs into a single byte code. For example, each “Copy” command in the logs was encoded into 0x07(7 in hexadecimal). Then, the LZW compression was applied to the encoded operation logs. The results from the third trials of version one of the software are shown in table 2 and 3. Table 2 shows the performances of typical 6 subjects. The numbers in the table represent the error steps. So, the smaller the number is, the better the performance is. And table 3 shows the compression rates. Two low-performance subjects’ compression rates were 58.3% and 61.2%, whereas two high-performance subjects’ compression rates were 79.6% and 77.6% (the smaller the percentage is, the higher the redundancy is).
These results suggest that it is possible to evaluate the degrees of users’ adaptation to software by using LZW compression rates in the controlled experimental situation.

Table 1. An example of LZW compression. In the output characters, repeated sequences are replaced with corresponding codes of string table. A code 257 (represents “ab”) repeated twice, and the compression rate is $\frac{9}{11} = 81.8\%$

| input          | string table | output                  |
|----------------|--------------|-------------------------|
| abcabdabcde    | 257: ab      | abc[257]d[257]cde       |
| (11 characters) | 261: da      |                         |
| 258: bc        | 262: abc     |                         |
| 259: ca        | 263: cd      |                         |
| 260: abd       | 264: de      |                         |

Table 2. Performances (error steps) of typical 6 subjects

| Use of the most effective functions | Performance | Completely | Partially | Not used |
|-------------------------------------|-------------|------------|-----------|----------|
|                                     | high        | 10         | 17        | 58       |
|                                     | Low         | 76         | 87        | 74       |

Table 3. Compression rate of the operation logs

| Use of the most effective functions | Performance | Completely | Partially | Not used |
|-------------------------------------|-------------|------------|-----------|----------|
|                                     | High        | 79.6%      | 68.5%     | 58.7%    |
|                                     | Low         | 69.5%      | 60.3%     | 58.3%    |

2.3 Analysis of String Table

In daily use of software, it is not always meaningful to adopt the compression rate as the index of redundancy, or the degree of adaptation to the software. The reason is that redundant operations include both incorrect operations and purposive operations (e.g. exploring new functions). Though, by comparing the string tables and output of
experts with the ones of novices, it is possible that we are able to distinguish incorrect operations from purposive ones, and are able to use compression rate as an index of

Table 4. Repeated sequences commonly detected in all subjects (third trials of version 1)

Repeated sequences in all subjects (4 subjects)
- double click -> finish editing
- edit -> cut
- edit -> paste
- edit -> paste -> edit

Table 5. Repeated sequences commonly detected in high performance subjects (third trials of version 1)

Repeated sequences in high performance subjects (2 subjects)
- edit -> copy

users’ adaptation. In fact, repeated sequences that were found in experts’ operations consisted of the subsets of the command tree, whereas novices’ operations logs also contained repeated sequences that were not the subsets of the command tree. Tables 4 to 6 show the repeated sequences in the operation logs.

Table 6. Repeated sequences commonly detected in low performance subjects (third trials of version 1)

Repeated sequences in low performance subjects (2 subjects)
- copy -> edit
- previous month -> previous month
- previous month -> previous month -> previous month
- previous month -> previous month -> previous month -> previous month
- paste -> edit
- edit -> format
- next month -> next month

It was also revealed that when subjects were not accustomed to the software, they tended to select two or more top level commands in the menu structure serially. Table 7 shows the repeated sequences that were extracted from the logs of the first trials of version two of the software. The numbers, which represent the level of the command in the menu structure, are added on the upper right of the command names.
As the subjects got accustomed, the repeated sequences of top level commands disappeared.

Table 7. Repeated sequences commonly detected in low performance subjects (first trials of version 2). The Numbers on the upper right of command names represent the level of commands in the menu structure

| Repeated sequences in low performance subjects (2 subjects) |
|----------------------------------------------------------|
| cell^2-> finish cell command                           |
| double click-> cancel editing                          |
| double click-> finish editing                          |
| tool^1-> help^1                                        |
| help^1-> tool^1                                        |
| search^1-> search^2                                    |
| search^1-> format^1                                    |
| format^1-> tool^1                                      |
| format^1-> line width^2                                |
| format^1-> search^1                                    |
| previous month -> previous month -> previous month     |
| previous month -> previous month -> previous month -> previous month |
| register idiomatic phrase^2-> finish registering       |
| edit^1-> copy^2                                        |
| edit^1-> search^1-> format^1                           |

3 Conclusions

It is expected that we are able to measure the degrees of users’ adaptation by using compression rate of LZW compression algorithm. However, in order to make a correct measurement, it is necessary to distinguish purposive operations from error operations. It is possible by extracting repeated operations that were the subset of the command tree. We would be able to use the compression rate as an index of adaptation by eliminating the purposive repetition. We also are able to evaluate the degrees of users’ adaptation by analyzing the level of commands in the operation logs. As a user becomes accustomed to the software, repeated sequences of top level commands will decrease.

The advantage of LZW compression algorithm is its simplicity. The essential part of LZW compression algorithm is able to written in several ten lines. There are many kinds of the user modeling method. Especially, Bayesian networks are broadly applied to user modeling (e.g. [6], etc.). They are, although, complex and difficult to understand for ordinary programmers. By applying LZW compression algorithm, it is possible to implement the method to evaluate the degrees of users’ adaptation easily into almost all software.
References

1. Sears, A, Shneiderman, B.: Split Menus: Effectively Using Selection Frequency to Organize Menus. ACM Transactions on Computer-Human Interaction (1994), Vol. 1, No. 1, pp. 27-51
2. Masui, T.: POBox: An Efficient Text Input Method for Handheld and Ubiquitous Computing. Proceedings of the First International Symposium on Handheld and Ubiquitous Computing (1999), No. 1707, pp. 289-300
3. Thomas, C. G., Krogsaeter, M.: An Adaptive Environment for the User Interface of Excel. Proceedings of the 1993 International Workshop on Intelligent User Interfaces (1993), pp. 123-130
4. Hayama, H. & Ueda, K.: Difference Presentation: A Method for Facilitating Users' Adaptation to Software Upgrade. Proceedings of HCI International 2003 (2003), Vol.3, pp.235-239
5. Welch, T. A.: A Technique for High-Performance Data Compression. IEEE Computer (1984), Vol. 17, No. 6, pp. 8-19
6. Conati, C., Gertner A., VanLehn K.: Using Bayesian Networks to Manage Uncertainty in Student Modeling. Journal of User Modeling and User-Adapted Interaction (2002), Vol.12, Issue.4, pp.371-417
Study on Segmentation Algorithm for Unconstrained Handwritten Numeral Strings

Zhang Chuang, Wu Ming, and Guo Jun

Lab of PRIS, Beijing University of Post and Telecommunication, Box186#, PRC,100876
zcbumpt@sina.com, wuming@sina.com, guojun@bupt.edu.cn

Abstract. In this paper, an integrated system of segmenting unconstrained handwritten numeral strings with unknowing number of digits is proposed, which consists of the extraction of connected components based on vertical projection and isolated components analysis, the length estimation of connected components using syntax analysis and waveform analysis and the segmentation of unconstrained connected handwritten numeral strings using innovative reverse “drop-falling” algorithm. This segmentation system which has promising results is then incorporated into a complete bank check character recognition system.

1 Introduction

The recognition of unconstrained handwritten numeral strings is applied widely in office automation, bank check processing, zip code recognition system, etc. Segmentation of connected handwritten digits is main bottleneck among several key techniques of recognition system. In the past years, many efforts have been devoted to the improvement of segmentation strategies, which strongly affects the recognition capabilities of systems in real environments. Specifically, Yi-Kai Chen and Jhing-Fa Wang [1] combine background and foreground analysis to segment single- or multiple-touching handwritten numeral strings; Bin Zhao and Hui Su[2] proposed a segmentation method based on recognition which also combining with dissection method and holistic method; G.Congedo and G.Dimauro[3,4] present a segmentation procedure used a hypothesis-then-verification strategy in which multiple segmentation algorithms based on contiguous row partition work sequential on the binary image until an acceptable segmentation is obtained. According to the targeted problems, these methods can be categorized into two classes, some for segmenting connected character strings with fixed numbers of characters [1] and the others with unfixed ones [2,3,4]. Both of them did not estimate the length of digit strings, which is helpful for the successful segmentation. In the recognition-based methods, the correct rate of segmentation depends too much on the robustness of recognizer and they are at the cost of the complexity of the algorithm.

In this paper, an integrated system of segmenting unconstrained handwritten numeral strings with unknowing number of digits will be proposed. This paper is organized as follows: in Section 2, the extraction of connected components will be
discussed; and we show how to estimate the length of digit strings based on syntax analysis in Section 3; we present an innovative reverse “drop-falling” algorithm for segmentation in Section 4; in Section 5, an experimentation platform, called MMSSToolbox (Multiple Method String Segmentation Toolbox) is developed, which enables the user to quickly change test data and variables of ‘drop-falling’ algorithms; in Section 6, we report our experimental results and draw some conclusions.

2 Extracting Connected Components

When the characters of the digit string are well spaced and written with some degree of neatness, segmentation is a straightforward process (Fig. 1). The vertical projection of the numeral strings consists of a simple running count of the black pixel in each column. It can serve for the detection of white space between successive characters letters. In that ideal situation, one just has to pick out each connected component. The problem becomes much more difficult when characters are touching, overlapping, or disjointed (Fig. 2).

A connected component is defined as a set of black pixels where each is a direct neighbor of at least one other black pixel in the component. After the vertical histogram analyzing, the split-blocks can be easily found including some dispersive noises. The split-blocks will comprise dispersive noises, part of a character, a whole character, or two or more characters. Each isolated connected component in the split-blocks will be represented by a list of component structures, which holds the component’s width and height, position of four borders, pixel area, black pixel count, and geometric center. If the isolated component is small enough to qualify as noise (width and height smaller than 1/3 of the average width and height, or black pixel count smaller than 1/3 of the average count, or the geometric center far from the center of split-blocks), it will be removed.

![Fig. 1. Vertical projection and vertical histogram of number strings](image1)

![Fig. 2. The split-blocks after vertical histogram analyzing (a) Touching (b) Overlap (c) Disjointed 5 (d) Touching, overlapping and noises](image2)
Most common occurrence of a numerical character being broken into two isolated components is when the top stroke of a ‘5’ is disjointed from the body of the numeral (Fig. 3a). Our segmentor uses the size and position information of the connected components to make a decision that whether a given connected component is a disjointed top stroke of ‘5’. We can judge a connected component to be the disjointed top stroke of a ‘5’ if the following criterions are met:

- The aspect ratio of the connected component is greater than 1.
- The height of the component is less than 1/3 the height of the previous character.
- The component is in the upper half of the image.
- The distance between the position of geometric center and the up/down border is less than 3.

When the disjoined top stroke of a ‘5’ was detected, move the smaller region to the left component to merge them as a whole region (Fig. 3b).

After connected component extracting, dispersive noises removed and disjointed region merged, there will only be the isolated connected component which is made up of one or more characters, and then how many digit strings compose of the isolated connected component should be estimated.

### 3 Length Estimation of Digit Strings Using Syntax Analysis and Waveform Analysis

Most isolated components are composed of one character; the others are composed of two or more characters. Accurate length estimation is very helpful for the successful segmentation of the connected numeral strings. The ratio of a connected component’s bounding box height to its width is the most common indicator of what fraction of characters the component is composed of, but it’s not enough to determine how many characters make up the component correctly.

Because the valley region in the upper contour of handwritten ‘4’ and ‘0’ will disturb the upper contour distribution of connected components (Fig. 4), we estimate the wave crest in the lower contour of connected components’ based on syntactical recognition instead of difference operator, which one’s coding chain is too long and less description of the concavity and convexity features in the wave. How many characters composed of the components can be determined based on the number of wave crest.
Because of the discontinuity contour of the components, we smooth the low contour of the connected numeral strings. The smooth algorithm is described as follows:

\[
dot[i] = \frac{(\dot{i-1} + \dot{i} + \dot{i+1})}{3}
\]  

(1)

Syntactic recognition is a good method to classify the waveform [5]. We construct the syntax description of the waveform, suppose \( w = w_1w_2...w_n \), \( w_i \in \{ p, n, 0 \} \) is the coding method of waveform, \( p \) stands for ascending stage, \( n \) stands for descending stage and \( 0 \) stands for the stage of neither ascending nor descending. In order to recognize the wave crest of the lower contour, a five elements deterministic finite automaton is defined in the following:

\[
A = \{ \Sigma, Q, \delta, z, F \}
\]

(2)

Where

\[
\Sigma = \{ n, p, 0 \}, Q = \{ z, p_1, p_2 \};
\]

(3)

\( z \) is the initial state, \( F = Q \) is the final state set; \( \delta \) is the mapping from \( Q \times \Sigma \) to \( Q \).

\[
\delta(z,n) = p_2 \quad \delta(z,0) = z \quad \delta(z,p) = p_1
\]

\[
\delta(p_1,n) = p_2 \quad \delta(p_1,0) = p_1 \quad \delta(p_1,p) = p_1
\]

(4)

\[
\delta(p_2,n) = p_2 \quad \delta(p_2,0) = p_2 \quad \delta(p_2,p) = p_1
\]

It is easy to recognize each crest and its initial point and final point by the deterministic finite automaton A: it is the position of the crest when \( p_1 \rightarrow p_2 \); it is the final point of current wave, and the initial point of next wave when \( p_2 \rightarrow p_1 \). It is necessary to identify the rationality of waveform, and eliminate the irrational crest caused by irregular handwriting (Fig. 5). We calculate each wave height and width according to the initial and final point, if one of the following constraints are satisfied, the wave is regarded as an irrational one and merged to the adjacent wave (Fig. 6).

- The height of the crest is smaller than 3.
- The width of the crest is less than 1/3 the average width of the rational crest.
- The height difference of two adjacent crests is greater than 3/4 the height of the tallest one.
The initial point and final point of the each wave can be a reference border of segmentation.

![Fig. 5. Irrational waveform caused by irregular handwriting](image)

![Fig. 6. The accurate length estimation based on syntax analysis and waveform analysis](image)

4 Improved “Drop-Falling” Algorithm for Connected Numeral Strings Segmentation

The basic idea of the “drop-falling” algorithm is to perform digit segmentation by simulating a “drop-falling” process. The main issue in its implementation is finding the starting point of drop-falling. There are four different algorithms of this type, up-left, up-right, down-left, down-right [3]. Salman [6] adopted hybrid “drop-falling” algorithms based on up-left and bottom-right which will miss the drop direction where there is the open region in the top of 4 and 0; Jibu [7] improved Salman’s work and chose segmentation path based recognition which is time consuming and much depend on the recognition engine. We propose an innovative reverse “drop-falling” algorithm which is detect the initials point by lower contour analyses and bottom-left /bottom-right next step searching strategy.

4.1 Reference Point and Initial Point

In section 3, we can mark the reference point of segmentation according to the wave crest in the lower contour. According to the reference point, the connected numeral strings will be segmented one by one, N numerals should be segmented N-1 times (Fig. 7). The initial point of drop-falling is very important since if the algorithm starts at the wrong place, the ‘drop’ could easily roll down the left side of the first digit or the right side of the next touching digit. We detect initial point according to the wave crest position in the lower contour and wave height.

If the wave height of the left digit is less than the right one, the initial point is marked by scanning the horizontal line (left wave crest located) from left to right to detect the first white “*” satisfying it’s left neighbor pixel is black (Fig. 8a), and if the wave height of the left digit is bigger than the right one, the initial point is marked by scanning the horizontal line (right wave crest located) from right to left to detect the
first white “*” satisfying its left neighbor pixel is black and a black pixel exists on the right of “*” (Fig. 8b).

Fig. 7. Reference points and segmentation times

Fig. 8. Detection of the initial point (a) wave_height (left) <= wave_height (right) (b) wave_height (left) > wave_height (right)
4.2 Next Step Searching Algorithm

After finding the initial point, the algorithm follows the next step searching algorithm until the bottom of the image is reached. The directions of the drop falling are according to the current pixel position and its five adjacent pixels. The stepwise movement rules are depicted graphically in Fig. 9.

Where \( x \) is the abscissa of the current pixel position, and \( X \) is the abscissa of the reference point. If all the pixels from 1 to 5 are black, there will be the touching field of the connected numbers, and it’s very easy to fall into the blind zone and lost the right way without any directions. We propose a “one of three” optimum searching algorithm, that means there are 3 directions, north, northeast, and northwest to choose when the “drop” fall into the blind zone (Fig. 10), the optimum one is which segmentation path cross over the least stokes. If the numbers of punched strokes are the same, the priority for the falling direction is north, northwest, and northeast. For example, in Fig. 11, both the 2nd path and 3rd path cross over 1 stroke, according to the priority, the 2nd path which drop to northwest is the optimum one. Fig. 12 shows the other two situations.

![Stepwise movement searching algorithm](image)

Fig. 9. Stepwise movement searching algorithm

5 Development and Design of the MMSSToolbox (Multiple-Method String Segmentation Toolbox)

In order to facilitate the experimentation, a graphical Multiple-Method String Segmentation Toolbox called MMSSToolbox was developed. The main purpose of building the toolbox was to provide an easy way to test various segmentation methods. So this tool enables the user to quickly change test data, display the test result, add on new routines, modify current routing, change the algorithms used for any step of the segmentation process, and easily incorporate the segmentation routines into another application.
Fig. 10. Different segmentation paths when the “drop” falls into the blind zone

Fig. 11. Punched stroke number of different segmentation path

Fig. 12. The correct segmentation path selection based on optimum searching algorithm
The interface consists of a main menu, a controls area, and a display area (Fig. 13). Using the controller in this area, the user is able to step through various segmentation steps, vertical histogram analysis, connected components extracting, dispersive noises removal, length estimation of digit strings, improved “drop-falling” algorithm for segmentation and so on. Finally, the display area is where all of the segmentation data are shown to the user.

![Fig. 13. Interface of MMSSToolbox1.0](image)

### 6 Experiment and Conclusion

The system was tested on the set account numbers of check images taken from a real business application (i.e. images of checks that were actually issued by China Bank of Construction). In order to gauge the accuracy of our segmentation algorithm, we must consider the performance on each segmentation step. The segmentation algorithm was tested using a set of 823 checks. But 16.8% of them (138 checks) are rejected, so 685 checks containing 5873 digits are segmented. The results are presented in Table 1.

|                      | Correct | Error | Total |
|----------------------|---------|-------|-------|
| Connected Extract    | 4702    | 66    | 4768  |
|                      | 98.6%   | 1.4%  |       |
| Length Estimate      | 4585    | 117   | 4702  |
|                      | 97.5%   | 2.5%  |       |
| Improved Drop-falling| 730     | 33    | 763   |
|                      | 95.7%   | 4.3%  |       |
Some correct examples shown as Fig. 14.

Some error examples show as Fig. 15, the main reason of the mistake is non-typical features in the touching field of the connected components, so it’s necessary to integrate other effective methods into our MMSSToolbox.
In this paper, we study on an innovative segmentation algorithm for unconstrained handwritten numeral strings. The improved performance of segmentation system has broader impact on improving the quality of automatic bank check recognition system. The significant experiment based on the actually bank check show our algorithm is more precise and robust for application.

Acknowledgments

This paper is supported by Chinese National 863 High-tech Project Foundation (No.2001AA114080).

References

[1] Yi-Kai Chen, Jhing-Fa Wang, “Segmentation of single- or Multiple-touching handwritten numeral string using background and foreground analysis”, IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, NO.11, November 2000, pp.1304-1317.
[2] Bin Zhao, Hui Su, Shaowei Xia, “A new method for segmenting unconstrained handwritten numeral string”, Document Analysis and Recognition,1995, Proceedings of the Fourth International Conference on, Vol. 2, 1997 ,pp.524 -527.
[3] Congedo. G, Dimauro. G, Impedovo. S, Pirlo. G, “Segmentation of numeric strings”, Document Analysis and Recognition, 1995, Proceedings of the Fourth International Conference on, Vol. 2, 1995, pp.1038 -1041.
[4] Dimauro. G, Impedovo. S, Pirlo. G and Salzo. A, “Automatic bank check processing: A new engineered system”, International Journal of Pattern Recognition and Artificial Intelligence,11 (4), 1997, pp.467-504.
[5] Yan Guohua, Gu jialiu, Ge Lichun, “Automatic classification of signal waveform useful in diagnosis”, Journal of Northwestern Polytechnic University, May 1996 Vol.14, NO.2, pp.294-298.
[6] Salman. Amin. Khan, “Character Segmentation Heuristics for Check Amount” Supervisor Dr. Amar Gupta MIT 1998 master dissertation.
[7] Jibu Punnoose, “An improved segmentation module for identification of handwritten numerals”, Supervisor Dr. Amar Gupta MIT 1999 master dissertation.
Wavelet-Based Image Watermarking Using the Genetic Algorithm

Prayoth Kumsawat¹, Kitt Attkitmongcol¹, Arthit Srikaew², and Sarawut Sujitjorn²

¹Signal and Image Processing Research Group, School of Electrical Engineering, Institute of Engineering, Suranaree University of Technology, 111 University Avenue, Muang District, Nakhon Ratchasrima, Thailand
{Prayoth, Kitt}@ccs.sut.ac.th

²Intelligent System Research Group, School of Electrical Engineering, Institute of Engineering, Suranaree University of Technology, 111 University Avenue, Muang District, Nakhon Ratchasrima, Thailand
{Ra, Sarawut}@ccs.sut.ac.th

Abstract. Image watermarking provides copyright protection of digital image by hiding appropriate information in the original image in such a way that it does not cause degradation of the perceptual image quality and cannot be removed. The watermarking methods for transform domains are usually achieved by using the discrete cosine transform or the discrete wavelet transform. In this paper, we develop a technique for optimizing the image watermarking using the genetic algorithm applied to the wavelet transform domain to improve the quality of the watermarked image and the robustness of the watermark. We then compare our experimental results with the results of previous works.

1 Introduction

Due to the advent of digital computer and subsequent development of digital multimedia technology, the demand for network distribution of images and video pictures has increased dramatically in the past decade. Although digital data has been shown to have many advantages over analog data, one of the potential problems on handling the digital data is that it can be easily duplicated. Thus, the importance of copyright protection becomes very crucial. As a solution to this problem, various digital watermarking techniques have been investigated to address the issue of ownership verification.

In general, digital watermarking can be performed in spatial domain or transform domain, where the properties of the underlying transform domain can be exploited. Previous works on digital image watermarking in spatial domain utilized the modified least significant bit of some pixels in an image. This scheme is fast and straightforward. Cox et al. [1] proposed a watermarking technique by embedding the watermark in the large discrete cosine transform coefficients using the concept of spread spectrum communication. Xia et al. [2] introduced a new multiresolution watermarking method based on the discrete wavelet transform. The watermark is embedded to the large wavelet coefficients at high and middle frequency bands of the
discrete wavelet transform of an image. In [3], Huang et al. proposed a watermarking method based on the discrete cosine transform (DCT) and the genetic algorithm. The genetic algorithm is applied to search for the locations to embed the watermark in the DCT coefficient block such that the watermarked image quality is optimized.

In this paper, we present the watermarking method using the genetic algorithm in the wavelet transform domain. The watermark is embedded to the wavelet coefficients larger than some threshold values based on [4] which does not require the original image in the detection process. We applied the genetic algorithm to search for optimal threshold values and the strength of the watermark to improve the quality of the watermarked image and the robustness of the watermark. We then compare our experimental results with the results of the previous works [4].

2 Proposed Method

In order to achieve an optimum performance of digital image watermarking, we always encounter three conflicting objectives. These are imperceptibility, robustness and data capacity. This work employs the genetic algorithm (GA) to optimally increase performance of the traditional watermarking methods. The diagram in Fig. 1 illustrates the optimization process in which GA is applied for the watermark embedding and the watermark detection processes, respectively.

2.1 Watermarking Method

We use the watermarking method based on the method proposed by Dugad [4]. The watermark insertion is performed in the discrete wavelet transform (DWT) domain by applying the 3-level wavelet decomposition with the Daubechies-8 wavelet. Fig. 2 shows the image subbands with their labels. Since the approximation subband (LL3) contains high-energy components of the image, we do not embed watermark in this subband to avoid visible degradation of image quality. Furthermore, the watermark is not embedded in the subbands of the finest scale (LH1, HL1 and HH1) due to the low-energy components to increase the robustness of the watermark. In other
subbands, we choose all coefficients which are greater than the embedding threshold $T_1$. These coefficients are named $V_i$ and applied to the following equation

$$V_i' = V_i + \alpha |x_i|$$

(1)

where $i$ runs over the wavelet coefficients $V_i$. $V_i'$ denotes the coefficients of the watermarked image. $x_i$ denotes the watermark which is the Guassian sequence of pseudo-random real number.

For watermark detection, we apply the same procedure but we pick the coefficients greater than the detection threshold $T_2 > T_1$ from all subbands except the approximation subband and the subbands of the finest scale. These coefficients are referred to as $\tilde{V}_i$. Then we compute the correlation $z$

$$z = \frac{1}{M} \sum_i \tilde{V}_i y_i$$

(2)

where $i$ runs over the wavelet coefficients $\tilde{V}_i$. $y_i$ is the different watermark generated similar to $x_i$. $M$ is the number of the $\tilde{V}_i$. We compare the correlation $z$ with the threshold $S$ described by

$$S = \frac{\alpha}{2M} \sum_i |\tilde{V}_i|$$

(3)

If the correlation $z$ is greater than $S$, the watermark has been detected.

![Wavelet decomposition of Lena image](image)

Fig. 2. Three level wavelet decomposition of Lena image

2.2 Genetic Algorithm for Improving the Watermarking Performance

Genetic algorithm is applied to search for optimal parameters in order to improve performance of watermarking. There are 3 parameters to search for: watermark strength (ALPHA or $\alpha$), embedding threshold ($T_1$) and detection threshold ($T_2$).
These parameters are searched for each subband with embedded watermark. Our optimization process employs a weighted objective function, $W$. $W$ composes of the terms reflecting the quality of the output images, and the robustness of the watermarks, respectively. This objective function is created based on the parameters obtained from the embedding and the detection of watermarks. Details of GA are described as follows:

Chromosomes in GA represent desired solutions to be searched. Number of chromosomes used in this work is 30. The encoding scheme is binary string with 32 bit resolutions for each solution. Since digital image processing can result in DWT coefficients of values $>T_1$. Care must be taken to ensure whether the watermark is embedded. Also for watermark detection, these DWT coefficients are not desirable in correlation $z$ computation. Thus, $T_2$ should be greater than $T_1$. The parameter $T_2$ can be computed from equation (4).

$$T_2 = T_1 + T_d$$

where $T_d$ is the optimal value for $T_2 > T_1$. The chromosomes then have the length of 96 bits representing $\alpha$, $T_1$ and $T_d$.

The objective functions use the mean-square-error ($MSE$) as the output image quality performance index, and the difference between correlation $z$ and the threshold $S$ ($DIFF$) as a watermark detection performance index respectively. An objective value $W$ can be computed from equation (5).

$$W = \delta_{MSE} \times MSE + \delta_{DIFF} \times DIFF$$

where $\delta_{MSE}$ and $\delta_{DIFF}$ are weighting factor of $MSE$ and $DIFF$, respectively. These weighting factors represent the importance of both indexes used in GA searching processes. If both indexes are both important, the values of these factors can both be 0.5 where the relationship $\delta_{MSE} + \delta_{DIFF} = 1.0$ must hold. The parameter $\alpha$, $T_1$ and $T_2$ are then searched for in order to achieve the best output image quality and watermark robustness.

Besides using the MSE in the objective function as the output image quality performance index, a mean-structural-similarity (MSSIM) [5] is also applied as in equation (5) with slight modification to compare the performance from both objective functions. In this work, the selection process employs ranking approach. The crossover and mutation probabilities are fixed at 0.7 and 0.05, respectively. The best chromosomes are then partially replaced (steady-state GA).

### 3 Experimental Results and Discussions

The output image quality and watermark robustness are two indexes for testing the performance. The images used in this test are grayscale of $512 \times 512$ pixels. The watermark is Gaussian sequence of uniformly-distributed pseudo-random real number with zero mean, and variance of one. The results obtained from the GA based on
MSE and MSSIM, which are called GA1 and GA2, respectively, are compared to the method proposed by Dugad [4], which is called METHOD1. In METHOD1, the parameter $\alpha$, $T_1$, and $T_2$ are fixed at 0.2, 40, and 50, respectively, while these parameters are varied during the GA searching process.

Fig. 3 shows the convergence of GA optimization of LH2 subband using MSE as an output image quality index at 50 generations of the tested image “Lena”. The resulting parameter $\alpha$, $T_1$, and $T_2$ are shown in Table 1. These parameters are optimally varied to archive the most suitable for the tested image.

![Fig. 3. MSE, DIFF and W from optimization of LH2 subband](image)

Table 1. Comparisons between $\alpha$, $T_1$ and $T_2$ from GA and results from METHOD1

| Watermarking Parameters | 6-Subbands in GA1(MSE) | METHOD |
|-------------------------|------------------------|--------|
| LH2                    | 0.105                  | 1      |
| HL2                    | 0.103                  |        |
| HH2                    | 0.116                  |        |
| LH3                    | 0.109                  |        |
| HL3                    | 0.113                  |        |
| HH3                    | 0.106                  |        |
| ALPHA($\alpha$)        | 0.105                  | 0.2    |
| T1                     | 31.61                  | 40     |
| T2                     | 46.58                  | 50     |

3.1 Invisibility Test

The output image quality is tested by watermarking the original image with the resulting parameters from GA. Next, standard values between the original image and the output image are computed and used to evaluate the quality of the output image. These standard values are peak signal to noise ratio (PSNR) and mean structural similarity (MSSIM). The tested results of output image quality from the optimization GA1 (MSE) and GA2 (MSSIM) are shown in Table 2, and our proposed method can improve the PSNR of the watermarked image about 6 dB.
Table 2. Comparisons of PSNR and MSSIM between GA1, GA2 and METHOD1

| IMAGES       | GA1(MSE) | GA2(MSSIM) | METHOD1 |
|--------------|----------|------------|---------|
|              | PSNR     | MSSIM      | PSNR    | MSSIM | PSNR | MSSIM |
| 1.Lena       | 46.55    | 0.9963     | 47.36   | 0.9969 | 41.33 | 0.9885 |
| 2.Baboon     | 42.09    | 0.9967     | 41.58   | 0.9973 | 34.68 | 0.9868 |
| 3.Gold Hill  | 46.49    | 0.9968     | 46.59   | 0.9970 | 40.41 | 0.9890 |
| 4.Boat       | 44.10    | 0.9956     | 44.54   | 0.9960 | 38.34 | 0.9851 |
| 5.Camera Man | 43.14    | 0.9932     | 43.23   | 0.9933 | 37.10 | 0.9771 |

**Fig. 4.** Correlation output using different attacks. (a) JPEG compression with various JPEG quality. (b) Lowpass filtering of 5 images (Lena, Baboon, Gold Hill, Boat and Camera Man), (c) Median filtering of Lena image, and (d) adding Gaussian noise to the Lena image.

### 3.2 Robustness Test

To verify the robustness of the watermark, we apply different attacks to the watermarked image and compare the results with METHOD 1. In the detection process, we also calculate the correlation output which is the ratio of the number of subbands that have the coefficients $> T_2$ and the number of all subbands which are searched for the watermark (in this case, 6). For example, if we detect the watermark...
in 5 different subbands from the total of 6 subbands, then the correlation output is $5/6$. Fig. 4(a) shows the results when the watermarked image is attacked by the JPEG compression. The results show that our method yields almost the same results as METHOD1. Then, we apply different attacks using common image processing such as lowpass filtering, and median filtering. In addition, we attack the watermarked image by adding Gaussian noise. From Fig. 4(b)–4(d), the results shows that our method yields better results than the METHOD1.

4 Conclusions

In this paper, the optimization for digital image watermarking using genetic algorithm has been presented. There are 3 parameters from watermarking process to be searched. These parameters are optimally varied to achieve the most suitable for the characteristic of each image. Our method is applicable to any images. The GA search guarantees the global optimum solution. The testing results of the output image quality and watermark robustness with various watermark attacks show that our proposed method can improve the performance of the watermarking such that the better output image quality and watermark robustness are achieved.

Acknowledgement. The financial support from Suranaree University of Technology is greatly acknowledged.

References

1. Cox, I.J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. In: IEEE Transactions on Image Processing. volume 6, no.12, pp. 1673-1687, Dec.1997.
2. Xia, X.-G., Boncelet, C.G., Arce, G.R.: A Multiresolution Watermark for Digital Images. In: Proceeding of the IEEE International Conference on Image Processing ICIP97, pp. 548-551, Santa Barbara, California (USA), Oct.1997.
3. Huang, C.-H., Wu, J.-L.: A Watermark Optimization Technique based on Genetic Algorithms. In: Proceeding of the SPIE-Visual Communications and Image Processing, pp.516-523, San Jose, California (USA), Feb. 2000.
4. Dugad, R., Ratakonda, K., Ahuja, N.: A New Wavelet-Based Scheme for Watermarking Images. In: Proceeding of the Conference on Image Processing ICIP98, pp.419-423, Oct. 1998.
5. Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P.: Image Quality Assessment: From Error Measurement to Structural Similarity. In: IEEE Transactions on Image Processing, volume 13, no. 4, Apr. 2004.
6. Srikaew, A.: Genetic Algorithms. In: Suranaree Journal of Science and Technology, volume 1, pp. 69-83, Jan. 2003.
Extraction of Road Information from Guidance Map Images

Hirokazu Watabe and Tsukasa Kawaoka
Dept. of Knowledge Engineering & Computer Sciences,
Doshisha University Kyo-Tanabe,
Kyoto, 610-0394, Japan

Abstract. A function of image recognition is indispensable to an intelligent robot which can coexist with a human being. Furthermore, the intelligent robot needs to understand the environment of their action range by getting information of characters and maps on advertisements and signboards in order to move autonomously. In this research, a method to search the route from the starting point to the destination on a guidance map, by extracting the road area on the map and revising degradation portions because of overlapping with characters or other figures is proposed. And the validity of this method is shown by the experiment using maps which were collected from advertisements or pamphlets.

1 Introduction

In order for an intelligent robot to carry out autonomous action, the robot has to understand information around him, such as characters, figures and guidance maps on a signboard or an advertisement. In this paper, the technique of extracting a road area from an image of the guidance map expressed abstractly such as a signboard or advertisement poster is proposed. And using this road area, the path planning from the start point to the destination on real time is possible.

Many effective techniques are proposed in search of the route on the map from the departure place to the destination in the research on the shortest path problem [1][2]. And, it is possible to obtain the shortest path at the high accuracy by utilizing GPS (Global Positioning system) by combining with the path planning technique in the current car navigation system. However, the car navigation system is not directly correspondent to image data, it requires topographical map called road network which is extracted from satellite image or map of the paper medium in manual operation by human beings.

This paper proposes the method of extracting the road area from map image which we humans use in everyday life for a direction map, such as the map on the signboard, advertisement posters and Web pages. In order to derive the optimal path by using extracted road area, the noise area by characters or symbols have to be eliminated. The validity of proposed method is shown by the experiment using maps which were collected from advertisement pamphlets and Web pages.
2 Extraction of Road Area

Generally, the guidance map image is constructed with a background area, a road area and the other area called a noise area (symbols, characters, etc.) (Fig. 1). And it has a characteristic that the number of pixels in the background area is more than the number of pixels in the road area, and the number of pixels in the noise area is the least. Therefore, the road area is extracted by selecting the second most brightness value of the brightness histogram (Fig. 2). However, the road area simply extracted by the second most brightness value of the brightness histogram has degradation since there are overlapped area on the road area by the noise area (Fig. 3). That is why it is necessary to restore the area to the road area without the noise area.

![Fig. 1. Guidance map](image1)

![Fig. 2. Brightness histogram](image2)
3 Removing Noises from Road Area

First of all, to restore the road area the map image is labeled into three areas, background area, road area and noise area (Fig. 4).

Fig. 3. Road area extraction from brightness histogram

Fig. 4. Labeling to three areas

(a) Inside road area  (b) Inside background area  (c) Both on road area and background area

Fig. 5. Types of noise areas
After the labeling, the road area restoring is possible by integrating the noise area to background area or road area. It is very easy if the noise area is completely inside of background area or road area (Fig. 5 (a), (b)). However, there are many cases that the noise area covers both of road area and background area (Fig. 5 (c)).

### 3.1 Integration of the Noise Area into the Road Area or the Background Area

Considering along with the line $y=y_i$ or $x=x_i$, there are three patterns of the noise area.

1) The noise area is completely inside of the road area (Fig. 6 (a)).
2) The noise area is completely inside of the background area (Fig. 6 (b)).
3) The noise area is both on the road area and the background area (Fig. 6 (c)).

![Fig. 6. Three noise patterns](image)

In order to eliminate the noise area, in the case of pattern 1, the noise area is integrated into the road area if the length of the noise area is short; otherwise the integration is not performed because the noise area may be connected to both of two
different roads (Fig. 7). And the noise area is integrated into the background area in the case of pattern 2. However, in the case of pattern 3, the noise area is integrated into neither road area nor background area in this line.

![Fig. 7. Exception case of noise pattern 1](image)

### 3.2 Road Area Extraction Algorithm

The following is the algorithm to extract the road area without the noise area.

1. Obtain the image A by scanning on the x-direction for performing the noise integration and by scanning on the y-direction for performing the noise integration.
2. Obtain the image B by scanning on the y-direction for performing the noise integration and by scanning on the x-direction for performing the noise integration.
3. The road area is the union set of the image A and B.

### 4 Construction of Road Network

Using the obtained road area image, the road network is constructed. Figure 8 shows the algorithm to generate the road network. At first, thinning operation is performed. And street intersections or terminal points are searched by determining the point which connects to more than two points or only connects to one point. These street intersections and terminal points become the nodes of road network.

![Fig. 8. Road network construction](image)
5 Experimental Results

To evaluate the proposed method, computer experiments are performed using 50 guidance map images, which are collected from advertisement pamphlets or Web pages. Among these images, the number of the success cases was 30, the number of almost success cases was 15, and the number of the fail cases was 5. The success means that all road areas are completely extracted for the purpose of road network generation. The almost success means that most road areas are extracted but a few road areas are not extracted.

Fig. 9. Experimental result (the success case)

Fig. 10. Experimental result (the fail case)
One of the success cases is shown in the figure 9. In this figures, (a) is a given gray scale guidance map image, (b) is the obtained image by scanning on the x-direction, (c) is the obtained image by scanning on the y-direction, and (d) is the final image by the proposed algorithm. Figure 10 shows the one of the fail cases. In this case, since the white boxes inside the road areas have the same brightness of the background area, these boxes are integrated into the background area, and the part of road areas are eliminated.

6 Conclusion

In this paper, the method of extracting the road area from map image for the purpose to generate the road network was proposed. Especially, the method of eliminating the noise area by characters or symbols was proposed. By the experiments using map images of advertisement pamphlets and Web pages, it is shown that the proposed method extracted the road area almost successfully from about 90% of map images.

Acknowledgements

This work was supported with the Aid of Doshisha University’s Research Promotion Fund.

References

1. E. W. Dijkstra: A note on two problems in connection with graphs, *Numerische Mathematik* 1, pp.269-271 (1959).
2. P. E. Hart, H. J. Nilsson and B. Raphael: A formal basis for the heuristic determination of minimum cost path, *IEEE Transactions on SSC ssc-4*, pp.100-107 (1968).
3. T. Hayakawa, T. Watanabe and N. Sugie: Extraction of road information from an urban map on the basis of the generation-verification paradigm of hypotheses, *IPSJ Journal*, Vol.35, No.1, pp.62-78 (1994).
4. W. Itonaga, I. Matsuda, N. Yoneyama and S. Itoh: Automatic extraction of road-networks from map images, *The IEICE Transactions on Information and Systems, PT.2* (Japanese Edition), Vol.J82-D-II, No.11, pp.1990-1999 (1999).
Dynamic Customer Profiling Architecture Using High Performance Computing

Qiubang Li, Rajiv Khosla, and Chris Lai

School of Business,
La Trobe University,
Victoria, Australia
{L.Qiubang, R.Khosla, C.Lai}@latrobe.edu.au

Abstract. Identifying profiles of customers based on their history trails of transactions and further providing personalization services to them are the best strategy to retain customers for online businesses. Currently profiling techniques have the defects of static offline operation, inexactness, and slow response, which do not meet the ever-changing demand of online transactions for e-commerce, e-banking, etc. This paper proposes a dynamic intelligent profiling architecture with the help of high performance computing. The architecture is tested and applied in an e-banking application.

1 Introduction

Today’s World Wide Web market is becoming more competitive. From the perspective of businesses, it’s more important than ever to understand exactly who are the best customers, and where to find more customers just like them. Customer profiling can help accomplishing this. It will help an e-business to get more customers, maximize the profitability of mid-potential customers, and identify low-potential customers so businesses can better reach them at minimum cost. It will also help e-commerce practitioner to position or feature their products on their shopping mall or website and also by knowing their online transaction behaviour they can formulate their marketing strategies. In addition, from the perspective of user of online businesses, customer profiling can identify customers from what they want or need without requiring them to ask for it explicitly and further deliver appropriate offers based on their profiles to them directly[1]. Knowing the buying habit of his peer group--a service provided by online businesses--is conducive to have the better choice of a new product he or she has not much knowledge about it. Many researchers have taken notice of this and have done a lot of work on it[2]. However, most existing approaches are static and get profiles in offline mode (Please refer to Section 2) owing to the huge computing workload of computers. Currently available commercial data increase in the order of a Gigabyte on a weekly or monthly basis. Some have reached Terabyte and even Petabyte [3]. Moreover, On-line customer profiling must respond to their users in a very short time—usually less than 30 seconds. High Performance Computing (HPC) becomes a necessity in these situations.
because of their super computing ability in terms of memory, multiprocessors, and secondary storage. For these reasons, we have implemented the customer profiling application using 128 processor Compaq Alpha Server (Alpha) SC, with 64 Gbyte of memory and 1.4 Terabyte of disk space on a Tru64 UNIX 5.1 Operating System.

This paper is organized as follows. Section 2 provides a background on customer profiling approaches. Section 3 describes the components of the multi-layered multi-agent customer profiling architecture. Section 4 illustrates the application and results in the banking domain. Section 6 concludes the paper.

2 Related Work in Customer Profiling

Currently, Most web-based companies use collaborative filtering technique which is heavily based on getting human intervention for customer profiling purpose. NetPerceptions tries to obtain or learns user preferences and discover Web information sources that correspond to their preferences, and possibly those of other individuals with similar interests using explicit rating from users [4]. The information inferences are mainly based on previous personal history and data accumulated from customers with similar attributes[5]. Similar cases can be found at GroupLens[6] and Firefly[7]. Some other examples using same approach includes WebWatcher ([8] and Syskill & Webert [9]. Syskill & Webert is a system that utilizes user profile and learns to rate Web pages of interest using Bayesian classifier and can suggest web pages based on their relevance to the user interest. Explicit rating has the fatal defects of losing quality control of profiling system in the beginning because of shortage of rating. Another profiling technique of Webmate [10] adopting a keyword vector to represent categories of user interest learns a user profile incrementally and facilitates user browsing and searching in the Web by using their profiles. Cooley et al. [11], and Buchner and Mulvenna [12] have applied data mining techniques to extract usage patterns from Web logs, for the purpose of deriving marketing intelligence. Shahabi et al. [13], Yan et al. [14], and Nasraoui et al. [15] have proposed clustering of user sessions to predict future user behaviour. These approaches have the problem that their predictions are all based on the previous user profiles offline, which is static and outdated to the customers. We propose a new customer profiling architecture using high performance computing facilities in VPAC (www.vpac.org). The machine is 128 processor Compaq Alpha Server (Alpha) SC, with 64 Gbyte of memory and 1.4 Terabyte of disk space on a Tru64 UNIX 5.1 Operating System. The dynamic online profiling architecture is discussed in the next section.

3 An Online Profiling Architecture

The architecture of customer profiling is a multi-agent, distributed, and complex system shown in Figure 1. From fine-grained to coarse-grained, there are three layers, which is sufficient to the domain[16], reactive agent layer, planning agent layer, and modelling agent layer[17]. Reactive agent layer focused on solving any specific problem, which needs to accept external sensor input and response immediately. It
basically includes three distinct agents of collaborative filtering, transaction frequency agent, and association agent. Some work of these agents in the layer can be found at [18]. The planning agent layer is in the middle layer of the customer profiling implementation hierarchy scheduling or optimising the lower-layered tasks in a knowledge-level view of the agent environment. High performance computing is used in this layer and the reactive layer to overcome the bottleneck of huge data exchange and realise the dynamic performance of customer profiling. The modelling agent layer, which is the uppermost implantation hierarchy of customer profiling system, interacts with users or customers via scheduling agent and reactive agent layers, which coordinates various modelling tasks in the layer and finally produces the global behaviours of the system.

The model layer is currently the least developed of the three-layer architecture in the present implementation. It includes a customer-profiling model consisting of rules and policies to guide a particular path or trajectory in a profiling process[19]. For example, it is no point for a customer to do transaction frequency analysis with only one or two transactions. As a result, a rule is established in the modelling agent to ignore the analysis in such a cases. The coordination model acts as a coordinator to coordinate different profiling agents representing different customers to solve possible conflicts once arisen as well as to help schedule the intelligent agents of different customers. For example, data reading and writing toward the same customer from two different agents at the same time is a conflict needing to be solved by coordination
agent. The communication of different agents is via message passing. This paper only further deals with the lower two levels of reactive agent layer and planning agent layer. More details of the architecture are in the next.

3.1 Parallel Scheduling Agent

The parallel scheduling agent has two parts: One is data partition; another is parallel processing using the high performance facilities in a multi-processor environment. Both parts are executed using MPI[20] commands.

![Fig. 2. Parallel Planning Scheduling Agent](image)

The necessity of data partition first is because the profiling result of a customer will be incomplete if we put the data from a customer into different processes, as most data mining algorithms rely on the whole data of customers by the consideration of data processing efficiency. From the data partition process shown in Figure 2, it is very easy to understand that the objective of data partition is to make the data of a customer integrated fractional and facilitate the profiling process of all customers. Therefore, in this case the data amount fed to each process is not necessarily equal.

In the parallel processing part, MPI commands are used to feed the available data to different processes after data partition and the Parallel Scheduling agent is used to collect data mining results from N-1 processes. The internal MPI implementation architecture used by the MPI Parallel Processing agent is shown in Figure 2. It is similar to a client/server mode. Clients, which are responsible to process raw data, are comparable to serial process. On the other hand, the server is in charge of collecting results from different processes. The performance of HPC will be shown in the experiment section.
3.2 Reactive Agent Layer

There are three agents in this layer: collaborative filtering, association rule agent, and transaction frequency agent. Their concepts and algorithms are shown next.

3.2.1 Collaborative Filtering
The concept of original collaborative filtering systems is to get personal recommendations by computing the similarity between the preferences of like-minded people. The following is the basic mechanism behind collaborative filtering systems:

- A large group of people's preferences are registered;
- Using a similarity metric, a subgroup of people is selected whose preferences are similar to the preferences of the person who seeks advice;
- (possibly weighted) average of the preferences for that subgroup is calculated;
- The resulting preference function is used to recommend options on which the advice-seeker has expressed no personal opinion as yet.

Cosine-based similarity, correlation-base similarity, and adjusted cosine similarity are typical similarity metrics[4]. We use the last one in this paper, which is suggested by Sarwar et al. 2001 at [4].

3.2.2 Association Rule
The concept definition at [21] is “Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items. In such a framework the problem is to discover all associations and correlations among data items where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items.” A good algorithm for large commercial database is suggested by Agrawal et al. 1993 at [22]. We will use it in this paper.

The algorithm is used for finding the association of two or more products and is used to calculate support and confidence. So if number of transactions for David to buy a ticket is 20; total number of transaction in database is 200; the number of transaction to buy air ticket and hotel booking is 16; then the confidence c=16/20=80% and support s = 16/200=8%. As a result we can say when David Buys ticket (antecedent), he has 80% confidence (or likelihood) to book accommodation (consequent), and the support for this allegation is 8% for all transactions.

3.2.3 Transaction Frequency
Transaction frequency agent tries to find the intrinsic acting frequency of customer purchasing/ selling in e-commerce or depositing/withdrawing in the banking domain. For example,

- John goes to but milk every two weeks.
- John goes to withdraw money after his deposit 20 days later.
By using this algorithm we will find out the transaction pattern of the customer—the average interval of transaction to the same product from the beginning to present.

4 Implementation of the Architecture and Partial Experiment Results

This section describes the online banking data model, results of customer profiling agents and the speed up achieved through use of high performance computing facilities.

4.1 Online Banking Data Model

The financial dataset available at the public domain archive for PKDD’99 Discovery Challenge (http://lisp.vse.cz/pkdd99/Challenge/chall.htm) is used to test the above-mentioned framework.

4.2 Profiling Association Rules of Banking Products

By implementing profiling agent and apply to banking products, we get following rules for our experiment.

- More transactions are done from the district where there are more inhabitants.
- More transactions are did by those clients whose average salary is also high.
- There are 10 clients who have both minus account balance and bad loans.
- We have 39 customers who have minus account balance out of them above 11 customers either have bad loans or credit card so that the reason they have negative account balance but we couldn’t find the reason that why 39-11=28 customers have minus account balance.
- Account_id 3948 used the most 212 times his credit card to pay his payment
- Most of the customers used credit card for household payments.
- In general it would be advisable to set business goal to promote credit card usage.

5 Conclusion

Customer profiling based on the history transactions trails of customers is of paramount importance for online businesses. In order to overcome the defects of static offline operation, inexactness, and delay response of current profiling approaches, this paper proposes a dynamic intelligent three-layer multi-agent distributed architecture with the help of high performance computing. The three layers are reactive layer, planning layer, and modelling layer. The reactive layer agents are composed of collaborative filtering agent, transaction frequency agent, and association agent. The planning layer agents include data partition agent and parallel processing agent. The modelling layer agents are in charge of coordinate different profiling agents or solve the conflicts once arisen in the system. The architecture is tested in an online banking domain and the response time is accord with the requirement of less than 30 seconds after 12 CPUs is used.
References

1. Ha, S.H., *Helping online customers decide through Web personalization*. Intelligent Systems, IEEE, 2002. 17(6): p. 34-43.
2. Li, Q., *Profiling Customer Buying Habit Project*. 2002, La Trobe University. p. 30.
3. Düllmann, D., *Petabyte databases*. in SIGMOD'99. 1999: ACM.
4. Sarwar, B., et al. *Item-based collaborative filtering recommendation algorithms*. in *The tenth international World Wide Web conference on World Wide Web*. 2001.
5. Li, Q. and R. Khosla. *Multi-agent Architecture for Automatic Recommendation System in E-commerce*. in *5th International Conference on Enterprise Information Systems*. 2003. Angers, France.
6. A. Konstan, J., et al., *GroupLens: applying collaborative filtering to Usenet news*. Communications of the ACM, 1997. 40(3): p. 77-87.
7. Shardanand, U. and P. Maes. *Social information filtering: algorithms for automating “word of mouth”*. in *Conference on Human Factors and Computing Systems*. 1995. Denver, Colorado, United States: ACM Press/Addison-Wesley Publishing Co. New York, NY, USA.
8. Armstrong, R.C., T. Joachims, and T. Mitchell *Webwatcher: A learning apprentice for the world wide web*. in AAAI Spring Symposium on *Information Gathering from Heterogeneous, Distributed Environments*. 1995.
9. Pazzani, M., J. Muramatsu, and D. Billsus. *Syskill & Webert: Identifying interesting web sites*. in *AAAI Spring Symposium*. 1996. Stanford, CA.
10. Chen, L. and K. Sycara. *WebMate: A Personal Agent for Browsing and Searching*. in *the 2nd International Conference on Autonomous Agents and Multi Agent Systems, AGENTS ’98*. 1998: ACM.
11. Cooley, R., B. Mobasher, and J. Srivastava, *Data Preparation for Mining World Wide Web Browsing Patterns*. Journal of Knowledge and Information Systems, 1999. 1(1).
12. Buchner, A. and M.D. Mulvenna, *Discovering internet marketing intelligence through online analytical web usage mining*. SIGMOD Record, 1998. 27(4): p. 54-61.
13. Shahabi, C., et al. *Knowledge discovery from users Web-pages navigation*. in *Workshop on Research Issues in Data Engineering*. 1997. Birmingham, England.
14. Yan, T., et al. *From user access patterns to dynamic hypertext linking*. in *5th International World Wide Web Conference*. 1996. Paris, France.
15. Nasraoui, O., et al. *Mining Web access logs using relational competitive fuzzy clustering*. in *Eight International Fuzzy Systems Association World Congress*. 1999.
16. Sycara, K., *Multiagent Systems*. AI Magazine, 1998. 19(2): p. 79-92.
17. Khosla, R., E. Damiani, and W. Grosky, *Human-Centered e-Business*. 2003: Kluwer Academic Publishers, MA, USA. 315.
18. Li, Q. and R. Khosla. *Intelligent Agent-Based Framework for Mining Customer Buying Habit in E-Commerce*. in *Fourth International Conference on Enterprise Information Systems*. 2002.
19. Nowostawski, M., M. Purvis, and S. Cranefield. *A Layered Approach for Modelling Agent Conversations*. in *the First Int'l Conference on Knowledge Discovery and Data Mining*. 1995. Montreal, Quebec, Canada.
20. Gropp, W., E. Lusk, and A. Skjellum, *Using MPI Portable Parallel Programming with Message-Passing Interface*. second ed. 2000, Cambridge Massachusetts London England: The MIT press.
21. Cooley, R., J. Srivastava, and B. Mobasher. *Web Mining: Information and Pattern Discovery on the World Wide Web*. in *the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’97)*. 1997.

22. Agrawal, R., T. Imielinski, and A. Swami. *Mining Associations between Sets of Items in Massive Databases*. in *the ACM-SIGMOD 1993 Int’l Conference on Management of Data*. 1993. Washington D.C.
Predicting Business Failure with a Case-Based Reasoning Approach

Angela Y.N. Yip
School of Business Information Technology, RMIT Business, RMIT University, GPO Box 2476V, Melbourne, Victoria 3001, Australia
angela.yip@ems.rmit.edu.au

Abstract. Accurately identifying potentially failing companies is beneficial to stakeholders of the companies. This paper uses a case-based reasoning (CBR) approach to predict business failure. CBR is a problem-solving paradigm that uses past experiences to solve new problems. Nearest neighbor (NN) is a common CBR algorithm for retrieving similar cases, but its similarity function is sensitive to irrelevant attributes. To ensure the effective retrieval of similar cases, statistical evaluations are used for automatically assigning relative importance of the attributes for the NN retrieval. The results of this study indicate that this approach is an effective and competitive alternative to predict business failure in a comprehensible manner.

1 Introduction

The failure of a company affects its entire existence and it is costly to the owners, investors, and creditors as well as workers. Accurately identifying potentially failing companies prevents or reduces the number of failures and is beneficial to stakeholders involved. Numerous approaches have been developed for predicting business failure and they have evolved from statistical to artificial intelligence approaches. Statistical approaches such as discriminant analysis (DA) and logit analysis are techniques used traditionally [2, 13]. Artificial intelligence approaches, in particular artificial neural networks, have been widely used in the last decade [9,15,16,18].

Until now, there is no general agreement upon which approach is consistently superior. Contradictory comparative results are often present [3,11,15,18]. Moreover, explanation and justification for a prediction is often neglected. Neither statistical models nor neural networks can explain their results [5].

Case-based reasoning (CBR) is a problem-solving and reasoning paradigm that is intuitively similar to the cognitive process humans follow in problem solving [10]. People often recall past similar experiences, and reuse or modify solutions of these experiences to generate a plausible answer for a problem. Applications of CBR can be found in planning, design, diagnosis and classification, and legal reasoning [12].

The strength of CBR lies in its capability to provide an explanation for its decision based on previous cases. Citing relevant previous experiences or cases is a way to justify a position [10]. Lawyers often use precedent cases for constructing and justifying arguments in new cases. Comprehensibility is often crucial in solving financial problems. To support financial decision making, simply an ‘accept’ or ‘reject’ format decision is not enough. For example, a company will demand an explanation when its
corporate bond is being rated as ‘C’ by a rating agency or when its loan application is
denied by a bank. When a company is identified as failing, CBR is able to give exam-
pies of similar companies that failed in the past as a justification for its prediction.
The reasoning behind a decision or a solution is therefore clear and can be explained.

Few studies have applied CBR to bankruptcy or business failure prediction. Among
them are Jo et al. [9], Bryant [5], and Park and Han [14]. The research of Park
and Han [14] does not compare performance of CBR with any other approaches and
therefore cannot inform as to the superiority or otherwise of CBR. The study of Jo
et al. [9] uses a case-based forecasting system (CBFS) to predict bankruptcy and the
results are inferior to that of DA and neural networks. Their study only gives the
overall predictive accuracies but does not inform how well these models performed in
correctly predicting failed or healthy companies. Bryant [5] states CBR is not useful
in predicting bankruptcy when compared with the logit model. This may be caused by
an inappropriate set of predictors leading to misspecification of the model. The appli-
cability of CBR to bankruptcy or business failure prediction remains an open research
question.

This paper applies nearest neighbor (NN), a common
CBR retrieving algorithm,
for predicting business failure in Australia. To guide effective matching and retrieval
of similar companies, relevancy of the attributes is taken into consideration. This pa-
per uses statistical evaluations for assigning the relevancy of the attributes in the NN
retrieval. The results of this study show that NN using weights derived from statistical
evaluations outperforms DA, and is particularly helpful in providing early warnings
of potentially failing companies in a comprehensible manner.

This paper is organized as follows: Section 2 describes the case matching and re-
trieving process, Section 3 reports the experiment and the results, and Section 4 con-
cludes the paper.

2 Case Matching and Retrieving

In CBR, knowledge is represented by cases. A case is a conceptualized piece of
knowledge representing an experience [10]. A case usually includes a problem de-
scription, its corresponding solution and/or outcome. Depending on the specific prob-
lem to be solved, the case may not include all these parts. A representative set of
cases forms a case base for a problem domain. The main tasks in a CBR cycle involve
retrieving the most similar cases, reusing the solutions of the retrieved cases, revising
the proposed solution if necessary, and retaining the new case [1]. Among the tasks,
retrieving the most similar cases is the first and crucial step because without this sub-
sequent steps cannot take place.

NN is a non-parametric algorithm that assesses similarity between a target case and
a stored case based on their attribute resemblance. A total of $K$ most similar cases to
the target case, which is called the $k$-NNs, are to be found and their outcomes are re-
used as the solution for the target case. These similar cases can be cited when needed
to justify for the solution. A pure NN algorithm assumes that all attributes are equally
important. This is not necessarily the case because some attributes are inherently more
relevant than others. The performance of the algorithm will be distorted by the pres-
ence of irrelevant attributes if all the attributes are treated equally. A good matching
function should take into account the relevancy of the attributes to guide the effective
matching and retrieving of similar cases. A weight that expresses the significance of an attribute is needed so that more important attributes are assigned with higher weights whereas lower weights are utilized for less important attributes.

Formally, given that $X$ and $Y$ are the target and stored case respectively with $n$ number of attributes and $x_i$ and $y_i$ are the values for their $i$th attribute, the similarity measure typically used for NN assessment is the inverse of the weighted normalized Euclidian distance. A similarity score between $X$ and $Y$ is calculated by,

$$SIM(X,Y) = 1 - DIST(X,Y) = 1 - \frac{1}{\sqrt{\sum_{i}^{n} w_i \cdot dist^2(x_i, y_i)}}$$  \hspace{1cm} (1)

where

$$dist(x_i, y_i) = \frac{|x_i - y_i|}{max_i - min_i} \hspace{1cm} (2)$$

By Equation (2), every attribute in the target case is matched to its corresponding attribute in the stored case. For symbolic attributes, $dist(x_i, y_i) = 0$ if $x_i = y_i$; otherwise $dist(x_i, y_i) = 1$. For numerical attributes, $max_i$ and $min_i$ are the maximum and minimum values of the $i$th attribute respectively. A weight $w_i$, being normalized, is assigned to every attribute representing its relative importance. As such, the similarity score is normalized within the range of 0 and 1, where 0 is a total dissimilarity and 1 is an exact match. This calculation is repeated for every stored case in the case base. Those cases with higher scores are more similar to the target case and will be ranked before the lower score cases. In this study, NN is chosen for matching and retrieving cases due to its explicit measure of finding the most similar cases.

Finding the set of weights that can improve the overall accuracy is an important as well as a challenging task. Domain experts can resolve this problem by determining the relevancy of the attributes [6]. However, experts are not always available and their subjective preferences are prone to inconsistency. Statistical evaluation is a method for assigning the relative importance of the attributes [7, 10]. In this study, four sets of values generated from statistical analysis are used as attribute weights. They include (a) F values (FVs), (b) structure matrix (SM) values, (c) Pearson’s correlation coefficients (PCs), and (d) standardized coefficients (SCs). The potential of each attribute can be measured by examining the absolute size of the significant FV. Values in the SM show the correlation of each attribute with the discriminant function. PCs measure the linear association of the attributes and the outcome. SCs allow the comparison of attributes measured on different scales and coefficients with larger absolute values imply greater discriminating ability of the attributes. They all tell the relevancy of the attributes and thus can be assigned as weights in the case matching and retrieval process.

3 The Experiment

3.1 Data

The research data used in this study is obtained from CD-Financial Analysis Publication 2002, in which financial data for the non-failed group only starts from 1991. The
data sample selected includes 44 Australian listed companies in which half of them failed and half of them survived. The 22 failed companies represent 5 randomly selected industries including food, media, property, retail and distribution, and services. The date of failure is taken to be the date of the appointment of an administrator, a receiver or a liquidator, whichever is earlier. The last financial statements of the failed companies are used. A similar sized company, measured by total assets, is randomly selected from the same industry to match with every failed company and its financial statement for the same year is used. The period under investigation is from 1991 to 2001 inclusively. Equal number of failed and non-failed companies is used because unique features of the companies in each group can be sufficiently recognized only with a nearly identical numbers of training samples [8].

Australian data on failed companies are quite restricted [4]. With a relatively small sample size, the leave-one-out resampling technique is used in order to obtain an almost unbiased estimation of the true error rate of classification [17]. This technique involves holding one case out of the whole sample and using the other cases to predict the extracted case. This is repeated for every case in the sample, and the proportion of misclassification in each class of outcome, failed or non-failed, is the misclassification rate.

3.2 Attribute Selection and Relevancy

Financial ratios derived from the financial statements are used as inputs in this study. Attributes in the models are selected out of a total of 44 ratios by stepwise selection using SPSS to reduce dimensionality. Table 1 lists the 5 selected attributes. X1 and X2 measure the profitability, X3 measures the liquidity of a company, X4 measures how well dividend payments are supported by profits, and X5 measures the leverage of a company.

| Definitions                                      | FVs | SM  | PCs | SCs |
|--------------------------------------------------|-----|-----|-----|-----|
| X1 Pre-tax profit over total assets              | 0.43| 0.33| 0.31| 0.47|
| X2 Earnings after abnormals over total assets    | 0.36| 0.30| 0.29| 0.42|
| X3 Working capital over total assets             | 0.13| 0.18| 0.19| 0.05|
| X4 Payout on operating profit before abnormals and tax | 0.05| 0.11| 0.12| 0.03|
| X5 Debt over gross cash flow                     | 0.02| 0.08| 0.09| 0.03|

These 5 selected attributes are not equally important. Their relevancies are determined by statistical evaluations, and 4 sets of weights are derived. Absolute values of these weights are normalized to fall between 0 and 1 and shown in Table 1. Although the relative importance of the attributes varies among the statistical evaluations, the order of the relevancy is consistent. Table 1 shows that X1 is the most important attribute indicating that profitability is most important to the viability of a company in this study.
3.3 The Prototype System

A prototype system was developed for predicting business failure. It is written in Java with part of its code adapted from Weka 3.2 [19]. The system parses the selected input file of cases and the associated file that contains sets of weights. The cases are in fixed format of attribute-value pairs. The case outcome emerges as a value of either failed or non-failed. The system utilises Equation (1) to calculate the similarity between a target case and a stored case. The number of \( k \) in the \( k \)-NNs is 1 in this study. In other words, only the stored case with the highest similarity score is retrieved. This is due to the fact that the expectation of the \( k \)-nearest cases being similar may not hold for a large \( k \) with a limited sample size [17]. In the context of business failure prediction, no revision or adaptation of the outcome is needed. The outcome of the retrieved case can be directly reused as the outcome of the target case. The predicted outcome is compared with the actual outcome of the target case to decide whether or not the classification is correct. An overview of the business failure prediction system is provided in Fig. 1. Using the 4 derived sets of weights, there are 4 sets of classification results. The system also implements a pure NN algorithm with equal weighting of the selected attributes as a cross reference and therefore, it generates 5 sets of results in total.

![Diagram of the business failure prediction system](image)

**Fig. 1.** An overview of the business failure prediction system

3.4 Empirical Results

The 5 sets of generated results are compared with the results of DA, the benchmark model. Fig. 2 shows that SCs weighted NN yields the highest overall accuracy of 90.9%. Pure NN with equal weights has the lowest accuracy of 79.5%. This implies that the selected attributes are actually not equally important and to assume that deteriorates accuracy. A weighted NN algorithm is therefore more accurate than a pure NN one.
Although the other weighted NNs have the same level of overall accuracy as DA does, they are still better than DA because business failure prediction is not neutral between Type I and Type II errors. Type I error is the misclassification of a failed company as non-failed whereas Type II error is the misclassification of a healthy company as failed. Type I error is normally considered to be more costly than Type II error to most stakeholders. For example, a creditor is likely to be more concerned with lending to a potentially failing company that may lead to bad debt, but is less worried about rejecting business opportunities with companies that turn out to be successful. Therefore, it is important to maintain a low Type I error to correctly predict failure. In this study, all weighted NNs are good at maintaining a low Type I error. SCs weighted NN can accurately predict all the failed companies while the others have a lower Type I error than DA does.

4 Conclusion

CBR is a problem-solving and reasoning paradigm that can provide a justification for its solution. This study applies CBR to business failure prediction, using NN as the retrieval algorithm. Weights obtained by statistical evaluations are used in the NN algorithm for effective retrieval of similar cases. The results of this study show that this approach outperforms DA in terms of accuracy and is particularly useful in identifying failing companies. To this end, CBR is a competitive alternative to provide early warning of failing companies for the benefits of stakeholders.

Acknowledgement

The author wishes to thank Robert Brooks and John Byrne for helpful comments on earlier versions of this paper.
References

[1] Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AICOM 7 (1994) 39-59
[2] Altman, E. I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance 23 (1968) 589-609
[3] Altman, E. I., Marco, G., Varetto, F.: Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (The Italian Experience). Journal of Banking and Finance 18 (1994) 505-529
[4] Altman, E.I., Narayanan, P.: Business failure classification models: an international survey. In: Choi, F. (ed.): International Accounting and Finance Handbook. Wiley, New York (1997) 35.1-35.50
[5] Bryant, S. M.: A Case-Based Reasoning Approach to Bankruptcy Prediction Modeling. Intelligent Systems in Accounting, Finance and Management 6 (1997) 195-214
[6] Gonzalez, A. J., Laureano-Ortiz, R.: A Case-Based Reasoning Approach to Real Estate Property Appraisal. Expert Systems with Applications 4 (1992) 229-246
[7] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C.: Multivariate Data Analysis with Readings. Prentice-Hall, NJ (1995)
[8] Jain, A. K., Chandrasekaran, B.: Dimensionality and sample size considerations in pattern recognition practice. In: Krishnaiah, P., Kanal L. (eds.): Handbook of Statistics. North-Holland, NY (1982) 835-855
[9] Jo, H., Han, I., Lee, H.: Bankruptcy Prediction using Case-Based Reasoning, Neural Networks, and Discriminant Analysis. Expert Systems with Applications 13 (1997) 97-108
[10] Kolodner, J. L.: Case-based Reasoning. Morgan Kaufmann, CA (1993)
[11] Lennox, C.: Identifying Failing Companies: A Re-evaluation of the Logit, Probit and DA Approaches. Journal of Economics and Business 51 (1999) 347-364
[12] Marling, C., Sqalli, M., Rissland, E., Munoz-Avila, H., Aha, D.: Case-based reasoning integrations. AI Magazine 23(1) (2002) 69-86
[13] Ohlson, J. A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research 18 (1980) 109-131
[14] Park, C., Han, I.: A Case-Based Reasoning with the Feature Weights Derived by Analytic Hierarchy Process for Bankruptcy Prediction. Expert Systems with Applications 23 (2002) 255-264
[15] Varetto, F.: Genetic Algorithms Applications in the Analysis of Insolvency Risk. Journal of Banking and Finance 22 (1998) 1421-1439
[16] Tan, R.S.K., Poh, H.I., Predicting Corporate Distress: Logit Analysis vs Artificial Neural Networks. Accounting Research Journal 15 (2002) 146-158
[17] Weiss, S. M., Kulikowski, C. A.: Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann, CA (1991)
[18] Wilson, R. L., Sharda, R.: Bankruptcy Prediction Using Neural Networks. Decision Support Systems 11 (1994) 545-557
[19] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)
Capturing and Applying Lessons Learned During Engineering Equipment Installation

Ian Watson

Dept. of Computer Science,
University of Auckland,
Auckland,
New Zealand
ian@cs.auckland.ac.nz
www.cs.auckland.ac.nz/~ian/

Abstract. This paper describes the implementation of a knowledge management tool to capture and reuse the lessons learned from the installation of engineering equipment. It has been developed as an adjunct to an existing system that uses case-based reasoning to reuse previous engineering installation specifications and designs. The system described lets engineers recall details of installation, commissioning and operational problems with systems. The paper discusses how lessons learned support the reuse and revision processes of the traditional CBR cycle.

1 Introduction

Several papers have been presented describing the author’s work in collaborating on the development of a case-based reasoning (CBR) system, called Cool Air, that supports the installation of HVAC (heating ventilation and air conditioning) equipment [Watson & Gardingen 1999a & b]. This system has been successfully fielded and made a significant return on its investment. It was designed to meet several goals:

- to reduce the installation specification and quotation time from five days or more to two days,
- to reduce the margin of error built-in to pricing and thereby produce more competitive quotations, and
- to reduce the burden on head office engineers in checking every detail of every specification.

The fielded system met these goals and generated a good return on its investment. However, although it was capturing and reusing engineering designs and specifications it was not adequately or consistently enabling the lessons learned.

1 A lesson learned is only a lesson learned if it has been applied to help prevent a mistake or error. Otherwise it is merely a lesson stored [Aha et al., 1999].
during design and installation to be applied. This was because the system did not proactively offer relevant lessons learned (LL) to engineers. Instead, like most LL repositories it relied on engineers actively searching for and retrieving relevant LL knowledge.

This paper describes enhancements to the Cool Air system to support the proactive delivery of LL knowledge to engineers when that knowledge is likely to be relevant, thus enhancing its knowledge management (KM) role. The status of the enhancements described are currently that of a research demonstrator.

2 System Architecture

Cool Air is a distributed client server system operating on the Internet. On the engineers (client) side a Java applet is used to gather the customer’s requirements and send them as structured XML to the server. On the server side another Java applet (a servlet) uses this information to query the database (approx. 14,000 records) to retrieve a set of similar records. This process takes the original query and relaxes terms in it to ensure that a useful number of records are retrieved from the database. This is similar to the query relaxation technique used by Kitano & Shimazu [1996] in the SQUAD system at NEC, although as is discussed in [Watson 2000] we have improved its efficiency using an introspective learning heuristic.

The Java servlet then converts the set of records into XML and sends them to the client side applet that uses a simple k-nearest neighbour algorithm to rank the set of cases. Once a matching case is retrieved the engineer obtains the installa-
tion specification files from the company FTP server. These files include CAD drawings, technical specifications, bills of quantities, contracts and notes (or trouble tickets) made by previous engineers describing problems with installation, commissioning and operation of the HVAC system. This requires that the engineer proactively downloads, via FTP, the appropriate trouble tickets and reads them. The engineer is not presented with the file name and location of trouble tickets from other similar installations, which may also be relevant. Consequently, the lessons learned from previous similar installations are not being transmitted and therefore cannot be learned.

Fig. 2. A Portion of a Symbol Hierarchy for Mechanical Heating & Cooling Systems

There are few publications referring to KM specifically in the construction industry, for example Cser et al., [1997], but there is a growing body of work about the application of CBR to KM. In particular a AAAI workshop on Exploring Synergies of Knowledge Management and Case-Based Reasoning [Aha et al., 1999], a workshop at ICCBR’99 on Practical Case-Based Reasoning Strategies for Building and Maintaining Corporate Memories [Gresse von Wangenheim & Tautz, 1999] and a recent book Applying Knowledge Management: techniques for building corporate memories [Watson, 2003]. This growing interest is not surprising since the recognition of similar problems and their solutions are central to both CBR and KM. Moreover the use of the Internet as a vehicle for supporting distributed KM is becoming more common [Caldwell et al., 1999].

Figure 3 shows parts of three example trouble tickets; one describing the need to reduce noise when installing a system in a residential nursing home, another describing the actual installed diameter of some ducting and the third describing a problem with a thermostat when located too far from a controller. Trouble tickets are indexed by Code (this refers to the job type), Location, Client (including a reference for client type) and a list of the equipment and contractors used (not shown in Figure 3). In each trouble ticket the problem and its solution are re-
corded. The trouble tickets are indexed in Cool Air’s database by these key features, along with a file reference to the trouble ticket itself.

3 Lessons Learned

The LL system offers a proactive two stage reminding. In the first stage when the set of similar installation records (typically between 10 & 20) is sent to the client all associated trouble tickets of these installations are also sent to the client as XML. Since these installations are similar it is reasonable to assume that any problems encountered with these installations may be relevant. Engineers can peruse these and use the information gained to improve the resulting design.

Fig. 3. Three Sample Trouble Tickets

In CBR terms the trouble tickets are being used to inform the case reuse and case revision or adaptation processes. Once a specification for the job is finalised the details for this new project are used to re-search the knowledge repository to obtain trouble tickets that might be relevant to the proposed job type, location, client type, equipment and contractors. This is relevant because the final adapted project specification may include significant variations from the cases upon which it was based and consequently it is valid to check its proposals for potential installation, commissioning or in-use problems.

Retrieval of trouble tickets uses CBR and the same abstraction hierarchies used by the query relaxation algorithm of the Cool Air system. An example hierarchy for mechanical heating and cooling equipment is shown in Figure 2. Using this hierarchy it is easy to see that U31A Athol and U32A Athol are both types of fan coil, are similar and hence may share similar problems and trouble tickets.

Searching is not performed on the body of the trouble ticket itself. Like many CBR systems Cool Air is featured based and it does not perform textual case-based reasoning [Ashley 1999]. Neither are trouble tickets retrieved using an iterative conversational CBR process [Aha & Breslow 1997]. However, textual CBR, conversational CBR or a combination of the two is being looked at as a possibility for future research.
Lessons Learned & the CBR-cycle

During the installation and commissioning of the HVAC system engineers will be encouraged to create trouble tickets using simple web-based forms. Once the project reference number is known all the relevant indexed features can be automatically added to the trouble ticket. Leaving the engineer free to concentrate on the body of the trouble ticket. Through the forms interface they will be encouraged to consider both the trouble encountered and the eventual solutions.

4 Conclusions

The first stage of the LL enhancements to the Cool Air system have undergone some limited testing and received qualified support. The second stage has not been field tested yet (March 2004) although it has performed satisfactorily in the laboratory. However, I do not underestimate the significant management problems associated with the successful operation of an LL system. Primarily these centre not upon the technology itself, which performs satisfactorily, but upon the management of the process [Davenport, 1997]. Put simply, not all engineers take the time to record problems and their solutions regarding this activity as a non-value adding task or at worst a threat to their experience and consequent, value to the company.

These issues, as many commentators have noticed, are as important to KM as the technology itself. Several methods are being suggested to overcome the reluc-
tance of engineers to create and use trouble tickets. These range from a system of rewards to encourage compliance to disciplinary penalties to enforce compliance. However, CBR has proven itself useful in the retrieval of LL knowledge and moreover, in an interesting synergy, the LL knowledge is useful in guiding both the reuse and the revision or adaptation processes of CBR. This is illustrated in Figure 4, which shows how LL knowledge first informs the selection of a past case upon which to base the subsequent solution and then secondly can be used to anticipate problems with that that solution.

References

Aha, D., Becerra-Fernandez, I., Maurer, F. & Munoz-Avila, H. (1999). Exploring Synergies of Knowledge Management and Case-Based Reasoning. AAAI Workshop Technical Report WS-99-19. AAAI Press

Aha, D. & Breslow, L.A. (1997). Refining conversational case libraries. In, Proc. 2nd. Int. Conf. On Case-Based Reasoning, pp. 267-78. Springer-Verlag.

Ashley, K. (1999). Progress in Text-Based Case-Based Reasoning. Invited talk at the 3rd. Int. Conf. On Case-Based Reasoning. Online (March 2000): www.lrdc.pitt.edu/Ashley/TalkOverheads.htm

Gresse von Wangenheim, C. & Tautz, F. (1999). Practical Case-Based Reasoning Strategies for Building and Maintaining Corporate Memories. Online (March 2000): www.eps.ufsc.br/~gresse/call_ws2.html

Caldwell, N.M.H., Rodgers, P.A. and Huxor, A.P. (1999). Web-Based Knowledge Management for Distributed Design, IEEE Intelligent Systems, September, 1999.

Cser, J., Beheshti, R. and van der Veer, P. (1997). Towards the development of an integrated building management system, Proceedings of the Portland International Conference on Management & Technology Innovation in Technology Management - The key to Global Leadership, PICMET.

Davenport, T. (1997). Information Ecology: Mastering the Information and Knowledge Environment: Why Technology is not enough for Success in the Information Age, Oxford University Press, 1997

Kitano, H., & Shimazu, H. (1996). The Experience Sharing Architecture: A Case Study in Corporate-Wide Case-Based Software Quality Control. In, Case-Based Reasoning: Experiences, Lessons, & Future Directions. Leake, D.B. (Ed.) pp.235-268. AAAI Press/The MIT Press Menlo Park, Calif., US.

Watson, I. (2000). A Case-Based Reasoning Application for Engineering Sales Support using Introspective Reasoning. In Proc. IAAI 2000. Austin Texas. AAAI Press. Forthcoming.

Watson, I. (2003). Applying Knowledge Management: techniques for building corporate memories. Morgan Kaufmann Publishers Inc., San Francisco, CA, US

Watson & Gardingen, D. (1999). A web-based CBR system for heating ventilation and air conditioning systems sales support. Knowledge Based Systems Journal, Vol. 12 Nos. 5-6, pp.207-214.

Watson, I. & Gardingen, D. (1999). A Distributed Case-Based Reasoning Application for Engineering Sales Support. IJCAI'99 31st. July - 6th August 1999, Vol. 1: pp. 600-605. Morgan Kaufmann Publishers Inc.
Case-Based Adaptation for
UML Diagram Reuse

Paulo Gomes, Francisco C. Pereira, Paulo Carreiro, Paulo Paiva, Nuno Seco,
José L. Ferreira, and Carlos Bento

CISUC - Centro de Informática e Sistemas da Universidade de Coimbra,
Departamento de Engenharia Informática, Universidade de Coimbra,
3030 Coimbra, Portugal
{pgomes, camara}@dei.uc.pt

Abstract. In this paper we present an approach to software design reuse based
on Case-Based Adaptation. We show how this approach is integrated in a CASE
tool suggesting solutions to the software designer. This approach generates new
UML designs based on previous ones, which are stored in a central repository.
Two different strategies are described and evaluated experimentally.

1 Motivation

The complexity of software systems has increased along the past decades. Nowadays
user interfaces are more sophisticated, data structures and system functionalities are
more complex. Software development companies have to build new systems in less
time and with limited resources. One possible solution for this situation is the reuse
of software [1]. There are several types of knowledge involved in the software devel-
opment process that can be reused. System specifications, software designs, and code,
are only part of the knowledge that can be reused. Code reuse has been the most com-
mon type of reuse, but it is not the most efficient in the sense that it works at a lower
level, not taking advantage of reuse of bigger software constructs. Decisions made at
the design level have a much stronger influence in the system development than de-
cisions made at the implementation level. This is one reason why we think that de-
sign reuse can be a solution for developing software faster and better. Designers need
Computer Aided Software Engineering (CASE) tools capable of helping them in this
task.

Software designers tend to reuse parts (or ideas) from different previous designs,
integrating them into a coherent design. Most of the times the generated design is novel
and bears characteristics that do not appear in previous designs. From the cognitive
point of view, design composition can be regarded as a cognitive process that can gen-
erate new designs. This process is a natural way of synthesizing new designs, and the
quality of the output greatly depends on the designer’s experience. The more experi-
ence, the easier is for the designer to reuse previous designs, and to reuse them in a
better way. We are interested in these two aspects of reuse: design composition and
experience. This paper describes how we have modelled these aspects into a CASE
tool.
2 Our Approach

In general, designers do not start a new project from scratch, they reuse old solutions that they developed earlier [2]. This is why inexperienced designers have a lower performance than experienced ones. The goal is to systematize this type of procedure, providing a cognitive framework capable of storing old designs and reuse them. Case-based reasoning [3] is a reasoning paradigm that uses experience in the form of cases to solve new problems. It enables the reuse of previously stored cases, seen as experiences, in new situations, which is the same cognitive process that designers often use in their job. The main entity in case-based reasoning (CBR) is a case, which represents a specific situation. In the design domain, it represents an artifact (a software design for example). Cases are stored and indexed in a case library. Usually associated with this case library, there is an indexing structure, enabling a more efficient retrieval of cases from memory.

The approach that we propose, is to develop an intelligent CASE tool for software design based on CBR. The software engineering process assumed in our approach, must have a design phase, in which the system being developed is modelled using Unified Modelling Language (UML) [4] class diagrams. There are two key ideas in our work: CBR as the reasoning framework for intelligent support, and the use of a general ontology as the conceptual basis for the knowledge used by CBR. Having these two issues in mind we developed a system named REBUILDER implementing our approach.

The main goal of REBUILDER is to provide the software designer with a design environment capable of promoting the reuse of software designs. It comprises four different modules: Knowledge Base (KB), UML Editor, KB Manager and CBR engine (see figure 1).

![Fig. 1. REBUILDER’s architecture](image)

The UML editor is the front-end of REBUILDER for the software designer, and the environment where s/he develops the system’s designs. The KB Manager module is used by the administrator to manage the KB, keeping it consistent and updated. The CBR engine performs all the inference work in REBUILDER. It comprises five submodules: retrieval, composition, analogy, verification, and learning. The retrieval module searches the case library for designs or design objects similar to the query. The most similar ones are presented to the user, allowing the user to reuse these designs or part of them. The analogy module maps designs from the case library, to the query design. The resulting mapping establishes the knowledge transfer from the old design to the query design. The composition module can be used to adapt a past design (or part of
it) to the query using design composition. The verification module checks the current design for inconsistencies. The learning module acquires new knowledge from the user interaction, or from the system reasoning. This paper describes how we have used design composition as a CBR adaptation strategy.

The next section presents the architecture of REBUILDER. Section 3 details the Knowledge Base used for REBUILDER’s reasoning. Section 4 presents our approach to case adaptation using design composition. Section 5 describes some experimental results obtained from user interaction. Section 6 presents related work by other researchers, and section 7 concludes this paper.

## 3 Knowledge Base

The KB comprises a case library, the WordNet ontology, the case indexes and the data type taxonomy. We start this section with the case representation, and then we describe the various sub modules of the KB.

In REBUILDER, a case describes a software design, through the use of UML Class Diagrams (figure 2 presents a simple example). Conceptually, a case in REBUILDER comprises: a name used to identify the case within the case library; the main package, which comprises all the objects that describe the case’s class diagram; and the file name where the case is stored. UML class diagram objects considered in REBUILDER are: packages, classes, interfaces and relations. A package is an UML object used to group other objects. A class describes an entity in UML and it corresponds to a concept described by attributes at a structural level, and by methods at a behavioral level. Interfaces only have method declarations, since they describe a protocol of communication for a specific class. Relations describe relationships between objects.

WordNet [5] is used in REBUILDER as an ontology. It uses a differential theory where concept meanings are represented by symbols that enable a theorist to distinguish among them. Symbols are words, and concept meanings are called synsets. A synset is a concept represented by one or more words. WordNet comprises a list of word synsets, and different semantic relations between synsets. The first part of WordNet is a list of words, each one with a list of synsets that the word can represent. The second part, is a set of semantic relations between synsets, like *is-a* relations, *part-of* relations, and other relations. In REBUILDER, we use the word synset list and four semantic relations: *is-a*,
part-of, substance-of, and member-of. Synsets are used for categorization of software objects. Each object has a context synset, which represents the object meaning. The object’s context synset can then be used for computing object similarity (using the WordNet semantic relations), or it can be used as a case index, allowing the rapid access to objects with the same classification.

As cases can be large, they are stored in files, which makes case access slower compared to cases stored in memory. To solve this problem, we use case indexes. Indexes provide a way to access relevant case parts for retrieval, without having to read all the case files from disk. Each object in a case is used as an index. REBUILDER uses the context synset of each object to index the case in WordNet. With this indexing scheme, REBUILDER can retrieve a complete case, using the case root package, or it can retrieve a subset of case objects. This allows REBUILDER to provide the user with the possibility to retrieve not only packages, but also classes and interfaces.

The data type taxonomy is a hierarchy of data types used in REBUILDER. Data types are used in the definition of attributes and methods. The data taxonomy is used to compute the conceptual distance between two data types.

4 Case-Based Composition

The input data used in the composition module is a class diagram, in the form of a package. This is the user’s query, which usually is a small class diagram in its early stage of development. The goal of the composition module is to generate new diagrams that have the query objects, thus providing an evolved version of the query diagram. Generation of a new UML design using case-based composition involves two main steps: retrieving cases from the case library to be used as knowledge sources, and using the retrieved cases (or parts of them) to build new UML diagrams. The following subsections describe these phases.

4.1 Diagram Retrieval

The retrieval process comprises two distinct phases. In the first phase it uses the context synsets of the query diagram to get $N$ objects from the case library, where $N$ is the number of objects to be retrieved ($N$ is user defined). This search is performed using the WordNet semantic relations that work like a conceptual graph, and case indexes that relate the case objects with WordNet synsets. The second phase ranks the set of retrieved objects using object similarity metrics.

The first phase uses the context synset of the query object as an entry point in the WordNet graph. Then it gets the objects that are indexed by this synset using the case indexes. Only objects of the same type as the query are retrieved. For instance, if the query is a class, then only classes are retrieved. If the objects found do not reach $N$, then the search is expanded to the neighbour synsets navigating in the is-a relations. Then, the algorithm gets the new set of objects indexed by these synsets. If there are still not enough objects, the system keeps expanding until it reaches the desired number of objects, or till there are no more objects to expand.

The result of the previous phase is a set of $N$ objects. The second phase ranks these objects by similarity with the query. Ranking is based on object similarity, and there
are three types of object similarities: package similarity, class similarity, and interface similarity.

Package similarity is based on four aspects: type similarity, sub-package similarity, diagram similarity, and dependencies similarity. Type similarity evaluates the distance between the query context synset and the retrieved context synset. This distance is computed as the number of semantic relations between the synsets in WordNet. The second criterion returns the similarity between the query sub-packages and the retrieved sub-packages, which is a recursive call to the package similarity metric. Diagram similarity is based on the similarity between the UML objects of both packages. As both diagrams are graphs, this metric computes the graph similarity between diagrams, taking into account the types of nodes (classes or interfaces) and structural constraints represented by the UML associations. The last component of the package similarity is the external similarity computed using the package dependencies, which are UML relations between packages.

Class similarity is based on three features: type similarity, intra-class similarity, and inter-class similarity. The type similarity is computed in the same way as the package similarity, evaluating the objects’ synset distance. The intra-class similarity is based on the similarity of attributes and methods of the classes. The inter-class similarity is based on the classes’ relations. Interface similarity is equal to class similarity, except that the intra-interface similarity is only based on the interface methods, since interfaces do not have attributes.

4.2 Diagram Composition

The diagram composition phase uses one or more retrieved cases to build a new case through the use of splitting and merging operations. Two adaptation strategies are used: best case composition, and best complementary cases composition.

In the best case composition, the adaptation module starts from the case most similar to the problem, mapping the case objects to the problem objects. The case mapped objects are copied to a new case. If this new case maps successfully all the problem objects, then the adaptation process ends. Otherwise it selects the retrieved case, which best complements the new case (in relation to the problem), and uses it to get the missing objects. This process continues while there are unmapped objects in the problem definition. Note that, if there are objects in an used case that are not in the problem, they can be transferred to the new case generating new objects. The complete algorithm is:

1. **RetrievedCases** ← Cases retrieved from the Case Library
2. **SelectedCases** ← ∅
3. **BestCase/Mapping** ← Select the best case and map it to **Problem**
4. **NewCase** ← Use **BestCase, Mapping** and **Problem** to generate a new case
5. Remove **BestCase from RetrievedCases**
6. WHILE **NewCase** does not map all the **Problem** objects AND **SelectedCases** ≠ ∅ DO
   (a) **SelectedCases** ← Search **RetrievedCases** for cases with unmapped problem objects
   (b) **SelectedCase** ← Select the best one, the one with more unmapped problem objects
   (c) **NewCase** ← Complete **NewCase** with **SelectedCase**
   (d) Remove **SelectedCase from RetrievedCases**
7. Return **NewCase**
The best complementary cases composition starts by matching each retrieved case to the problem, yielding a mapping between the case objects and the problem objects. This is used to determine the degree of problem coverage of each case, after which several sets of cases are constructed. These sets are based on the combined coverage of the problem, with the goal of finding sets of cases that globally map all the problem objects. The best matching set is then used to generate a new case. The algorithm is:

1. `RetrievedCases ← Retrieve cases from the Case Library`
2. `FOR RetrievedCase in RetrievedCases DO`
   (a) `Mapping ← Map RetrievedCase to Problem`
3. `CaseSets ← Create the sets of complementary cases based on each case mapping`
4. `BestSet ← Select the best set from CaseSets, this is the set whose mapping has the best coverage of problem objects`
5. `NewCase ← Generate a new case using the BestSet`
6. Return NewCase

5 Experiments

Evaluation of the two composition strategies used in REBUILDER, was done using user evaluation experiments. We have used a Knowledge Base with 60 cases describing software designs. These cases are from four different domains: banking information systems, health information systems, educational institution information systems, and store information systems (grocery stores, video stores, and others). Each design comprises a package, with 5 to 20 objects (total number of objects in the knowledge base is 586). Each object has up to 20 attributes, and up to 20 methods. These designs are defined at a conceptual level, so the design is at an early stage of development having only the fundamental objects.

Twenty five problems were defined, each one having one package with several objects (between 3 and 5), which were related to each other by UML associations or generalizations. These problems are distributed by the four case domains in the following way: banking information systems (6), health information systems (7), educational institution information systems (3), and store information systems (9).

The design composition mechanism uses a parameter that establishes if an object can be mapped or not with a problem’s object. If a mapping candidate object is conceptually closer to the problem’s object than the threshold value, then this object can be mapped with the problem. To assess the conceptual distance, REBUILDER uses WordNet is-a links, so the threshold value is the maximum mapping distance. In the experiments, we also evaluate the influence of this parameter in the mechanism evaluation. Six configurations were defined: Best Case Composition with threshold values 5 (C1), 10 (C2) and 15 (C3), Best Complementary Cases Composition with threshold 5 (c4), 10 (C5) and 15 (C6).

For each problem REBUILDER generated six solutions, one for each configuration. The problems and their respective solutions were then presented to the test users (software designers and software engineers) for evaluation.

Six test users were inquired to evaluate the solutions, giving their evaluation about the number of objects that they considered inadequate or incorrectly defined, regarding the problem that it was supposed to solve. Most of the designers made this judgment
based on what they would delete from the suggested solution in order to solve the presented problem. The results obtained are presented in table [1] *Time* - is the average computation time (in seconds) for generating one solution. *T.Objs* - is the average number of transferred objects by solution. *U.Cases* - is the number of cases used for generating a solution. *%M.Objs* - is the percentage of objects mapped to problem objects. *%W.Objs* - is the number of objects in the solution considered wrong by users.

|     | Time | T.Objs | U.Cases | %M.Objs | %W.Objs |
|-----|------|--------|---------|---------|---------|
| C1  | 72.08| 18.20  | 4.40    | 70.27   | 29.18   |
| C2  | 73.20| 18.52  | 4.36    | 70.27   | 29.24   |
| C3  | 74.96| 18.52  | 4.36    | 70.27   | 29.44   |
| C4  | 52.48| 5.00   | 1.32    | 64.60   | 11.86   |
| C5  | 59.68| 5.84   | 1.28    | 71.07   | 21.57   |
| C6  | 60.48| 5.84   | 1.28    | 71.07   | 22.93   |

The results of table [1] show that solutions generated by the Best Complementary Cases strategy were considered more accurate, since the test subjects considered them with less incorrect objects. One characteristic of this strategy that may have been decisive, is the number of objects and relations that the solutions generated have. As can be seen from the table, the Best Complementary Cases strategy generates few objects and relations, but the ones that are selected are taken to be more relevant for the problem being solved.

To compare the computational performance of both strategies, we have tested the computation time that each strategy uses to produce solutions. We have used the same 25 problems and performed several runs with each one. We obtained the average of the computation time for the generation of two solutions, using each strategy. Overall values show that the Best Complementary Cases strategy is better, outperforming the Best Case strategy by 22%.

From the test users evaluations and from the computational performance, it can be inferred that the Best Complementary Cases strategy performs better than the Best Case strategy in most of the problems used.

### 6 Related Work

We have divided the related works in two main types of systems: CBR design systems that perform adaptation, and CBR systems for software reuse. The remaining of this section explores these types of systems.

There are several approaches to case adaptation in CBR design systems. For instance, CADSYN [6] uses constraint satisfaction algorithms for adaptation. Composer [7] goes further, and combines design composition with constraint resolution. Other approaches to design adaptation involve the application of rules (FABEL [8]), or model-based reasoning (KRITIK [9]). FABEL is a system that uses several adaptation methods such as constraint satisfaction or generate and test. Most of these systems involve domain knowledge, which
might not be available in some domains. In software design, part of the world is modelled as a software system. So, the domain knowledge is the domain of application, which can be anything. This makes an approach to adaptation using domain knowledge limited to the domains for which the system was developed. In REBUILDER, the goal is to build a generalist system that could be used in almost every type of software systems.

Most of the research work on CBR for software reuse is on code reuse systems. Nevertheless there are some systems that can be considered related to REBUILDER. González et. al. [10] presented a CBR approach for software reuse based on the reuse and design of Object-Oriented code. Déjà vu [11] is a CBR systems for code generation and reuse using hierarchical CBR. The main improvement of this system is the adaptation-guided retrieval, which retrieves cases based on the case adaptation effort instead of the similarity with the target problem. CAESER [12] is another code reuse CBR tool. It works at the code level and uses data-flow analysis to acquire functional indexes. Althoff [13] have a different approach to software reuse and design. Instead of reusing code, they reuse system requirements and associated software development knowledge.

7 Conclusions and Future Work

In this paper, we present an approach to case adaptation of software designs in the form of UML class diagrams. This approach is based on composition of pieces of design cases, and is complemented by WordNet as an indexing structure. The main advantage of REBUILDER adaptation mechanism is that it can generate designs from old cases, thus extending the system’s solving abilities. As REBUILDER is a generalist approach to software design reuse, it does not use knowledge about specific domains, instead it uses abstract operations that allow the adaptation mechanism to generate new solutions only from cases. One of the limitations of this approach is due to the lack of domain specific knowledge, but the time and resources that have to be used to acquire this domain knowledge are too expensive to justify its usage.

Acknowledgments

This work was partially supported by POSI - Programa Operacional Sociedade de Informação de Fundação Portuguesa para a Ciência e Tecnologia and European Union FEDER, under contract POSI/33399/SRI/2000, by program PRAXIS XXI. REBUILDER homepage is http://rebuilder.dei.uc.pt.

References

1. Coulange, B.: Software Reuse. Springer Verlag, London (1997)
2. Tong, C., Sriram, D.: Artificial Intelligence in Engineering Design. Volume I. Academic Press (1992)
3. Kolodner, J.: Case-Based Reasoning. Morgan Kaufman (1993)
4. Rumbaugh, J., Jacobson, I., Booch, G.: The Unified Modeling Language Reference Manual. Addison-Wesley, Reading, MA (1998)
5. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: an on-line lexical database. International Journal of Lexicography 3 (1990) 235 – 244
6. Maher, M.L.: Casecad and cadsyn. In Maher, M.L., Pu, P., eds.: Issues and Applications of Case-Based Reasoning in Design, Mahwah, NJ, Lawrence Erlbaum Associates (1997) 161–185
7. Pu, P., Purvis, L.: Formalizing the adaptation process for case-based design. In Maher, M.L., Pu, P., eds.: Issues and Applications of Case-Based Reasoning in Design, Mahwah, NJ, Lawrence Erlbaum Associates (1997) 221–260
8. Voss, A.: Case design specialists in fabel. In Maher, M.L., Pu, P., eds.: Issues and Applications of Case-Based Reasoning in Design, Mahwah, NJ, Lawrence Erlbaum Associates (1997) 301–336
9. Goel, A., Bhatta, S., Stroulia, E.: Kritik: An early case-based design system. In Maher, M.L., Pu, P., eds.: Issues and Applications of Case-Based Reasoning in Design, Mahwah, NJ, Lawrence Erlbaum Associates (1997) 87–132
10. Fernández-Chamizo, C., González-Calero, P., Gómez-Albarrán, M., Hernández-Yáñez, L.: Supporting object reuse through case-based reasoning. In Smith, I., Faltings, B., eds.: Third European Workshop on Case-Based Reasoning (EWCBR’96). Volume 1168., Lausanne, Suiss, Springer-Verlag (1996) 150–163
11. Smyth, B., Cunningham, P.: Deja vu: A hierarchical case-based reasoning system for software design. In Neumann, B., ed.: 10th European Conference on Artificial Intelligence (ECAI’92), Vienna, Austria, John Wiley and Sons (1992)
12. Fouqué, G., Matwin, S.: Compositional software reuse with case-based reasoning. In: 9th Conference on Artificial Intelligence for Applications (CAIA’93), Orlando, FL, USA, IEEE Computer Society Press (1993)
13. Althoff, K.D., Birk, A., Wangenheim, C.G.v., Tautz, C.: Case–based reasoning for experimental software engineering. Technical Report 063.97/E, Fraunhofer IESE (1997)
Harmonic Identification for Active Power Filters Via Adaptive Tabu Search Method

Thanatchai Kulworawanichpong, Kongpol Areerak, Kongpan Areerak, and Sarawut Sujitjorn
School of Electrical Engineering, Suranaree University of Technology
Nakhon Ratchasima, 30000, Thailand
sarawut@ccs.sut.ac.th

Abstract. Harmonic identification by using Adaptive Tabu Search (ATS) Method embedded in the active power filter is proposed in this paper. The use of the ATS identifies harmonic components more accurately and precisely. Besides the accuracy and precision, it is able to select only some particular harmonic orders that cause severe consequences to the system for elimination. This principle thus leads to the reduction in size and cost of hardware implementation + active power filters. In this paper, two test current waveforms are simulated to validate and verify the performance of the proposed algorithm. The satisfactory results obtained by this identification method are also compared against those obtained by the d-q axis based harmonic identification method. As a result, the ATS based method has better performance for eliminating only selective harmonic orders over the d-q method. Furthermore, the compensated current from the proposed method has a good transient response while there is the first-cycle delay due to the use of the d-q method.

1 Introduction

To date, power electronic converters have been successfully developed and widely used in various industrial applications. Their principle is based on the high-speed operation of power switching devices in such a way that their waveforms are characterized in order to control power transfer between the two sides of the converters. It is the fact that the use of such high-speed devices operating at higher frequency than the fundamental inevitably causes undesired harmonic components. These components are troublesome to power supplies especially ones with highly sensitive electronic loads. Harmonic as “pollution” in electrical networks may cause unpredictable events to harm electric appliances such as cogging and crawling in motors [1], reduction in accuracy and precision of protective devices which could damage a power system or even harm the system operators [2, 3].

In general, harmonic power filters can be categorized into passive and active classes. For passive filters, they are limited to fixed orders for compensation. A design must set up the filter to compensate some selected harmonic components only, while the others

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 687–694, 2004.
© Springer-Verlag Berlin Heidelberg 2004
are still there. Unlike passive ones, active power filters [4] are more efficient due to
the ability of their switching devices to generate adjustable compensating current.
Most active power filters have their structures similar to that shown in figure 1. The
operation of the active power filter requires harmonic identification via so-called
“harmonic identifier”. This identifier must have an ability to detect most harmful
harmonics in an unhealthy power supply for elimination. Thus, this paper is focused
on study and comparison of two harmonic identification methods: 1) d-q axis (DQ) [5]
harmonic identification method and Adaptive Tabu search (ATS) [6] harmonic
identification method. To assess the effectiveness of each identification, their prin-
ciples are briefly reviewed, method-by-method, in Section 2 while two test waveforms
and results are situated in Section 3.

2 Algorithms for Harmonic Identification

2.1 d-q Axis Harmonic Identification Method

This identification is based on the concept of transforming three-phase current (u, v
and w phases) to two-phase current (α and β phases), which can be written in a com-
 pact matrix as in equation (1).

\[
\begin{bmatrix}
i_\alpha \\
i_\beta 
\end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix}
1 & -\frac{1}{2} & -\frac{1}{2} \\
0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2}
\end{bmatrix} \begin{bmatrix}
i_u \\
i_v \\
i_w
\end{bmatrix}
\] (1)

From the two-phase model, the current is then transformed again but this time to
the rotated d-q axis rotating with the speed of a selected harmonic frequency as de-
scribed by equation (2).

\[
\begin{bmatrix}
i_d \\
i_q 
\end{bmatrix} = \begin{bmatrix}
\cos(\omega t) & \sin(\omega t) \\
-\sin(\omega t) & \cos(\omega t)
\end{bmatrix} \begin{bmatrix}
i_\alpha \\
i_\beta
\end{bmatrix}
\] (2)
When the selected harmonic to be eliminated is transferred to the d-q axis, it appears to be stand still, this harmonic is simply sifted by using a low-pass filter. Let $i_{dh}$ and $i_{qh}$ are the $h^{th}$ harmonic components on the d-q axis, $i_{ah}$ and $i_{bh}$ are the corresponding components transferred to the $\alpha\beta$ axis, and $i_{uh}$, $i_{vh}$ and $i_{wh}$ are the selected harmonic components to be eliminated that are transferred back to the three-phase system. Moreover, the DQ detection method can be schematically summarized in figure 2.

\[ \hat{i}(t) = \sum_{h=1}^{\infty} I_h \sin(\omega t + \phi_h) \]  

where $I_h$ and $\phi_h$ are the magnitude and phase of the $h^{th}$ harmonic order respectively, while $\omega$ is the fundamental frequency in rad/s. With the ATS method applied to this scheme, $I_h$ and $\phi_h$ of not-passed orders are adjusted in order to minimize the objective

---

**Fig. 2.** Block diagram for the DQ harmonic detection method

### 2.2 Adaptive Tabu Search Harmonic Detection Method

Adaptive Tabu Search (ATS) Method [6] is a modified form of the original Tabu search proposed by Glover [7] in 1986 especially for combinatorial optimization problems. The modified version was developed according to the need for a powerful search method to solve non-linear continuous optimization problems. The essence of this method, which distinguishes itself from the original is that 1) a continuous search space must be discretized and 2) back-tracking and adaptive radius features are employed to enhance the overall performance of the search process [6]. Its effectiveness has been proved and verified by some intensive works [6,8]. In this paper, the ATS method is selected to be an alternative for the application of active power filters. It is used to identify harmonic contents of a one-period-sampled non-sinusoidal waveform as described by the flow diagram in figure 3. The proposed algorithm begins with examining harmonic components, order-by-order, against the IEEE519-1992 standard [9]. Thus, only orders that do not pass the standard are compensated for. This leads to the less compensation from the harmonic power filter.
function written in equations (4) and (5). The proposed method is initialized by using the harmonic components obtained from the FFT to speed up the process. At the end, an obtained solution that best fits the objective is used to command the active power filter.

![Flowchart](chart.png)

**Fig. 3.** Harmonics identification via ATS

\[ J = \sum e^2(kT); kT = 0,1,2,... \]  
\[ \cos t \text{ function} = \min(J) \]  

3 Results and Discussions

The two harmonic detection algorithms reviewed in the previous section are challenged by two waveforms as shown in figures 4 and 7 to situate the tests. The results obtained from each method are compared to evaluate their compensation performance. The comparison is illustrated, case-by-case, as follows.

![Example Waveforms](waveforms.png)

(a) time-domain waveform  
(b) spectrum

**Fig. 4.** Test current for Test Case 1
3.1 Test Case 1

Table 1 presents the error on the harmonic identification performed by each method in comparison with the actual one. It reveals that the error from the ATS method is less than that from the DQ method. Note that the 11th and the 13th harmonic orders are not taken into account because they satisfy the standard.

Table 1. Case-1 comparison between the harmonic components detected by the two methods

| Order | Actual (magnitude) | DQ | ATS |
|-------|--------------------|----|-----|
|       |                    | magnitude | error | magnitude | error |
| 5     | 8.00 (40%)         | 8.089    | 1.11% | 8.00      | 0.03% |
| 7     | 2.86 (14.3%)       | 2.88     | 0.70% | 2.86      | 0.05% |
| 11    | 0.67 (3.33%)       | satisfy std. | satisfy std. | satisfy std. | satisfy std. |
| 13    | 0.44 (2.22%)       | satisfy std. | satisfy std. | satisfy std. | satisfy std. |
| rms error |                  | 0.93%    |       | 0.04%    |

When the compensating current from the active power filter based on each method is injected to the system, the source current is therefore shaped to be a purely sinusoidal waveform as shown in figures 5 and 6 for the DQ and ATS methods, respectively.

Fig. 5. Corresponding waveforms and spectrum due to the DQ harmonic detection for Test Case 1

Fig. 6. Corresponding waveforms and spectrum due to the ATS harmonic detection for Test Case 1
Clearly, figure 5 shows that the DQ method for harmonic detection has the delay (failure to perform the compensation) within the first cycle of the operation. This delay is because the appearance of the poles in the LPF. In contrast, there is no delay when the ATS harmonic detection is used. In addition, as a result, the ATS detection method has better ability to compensate the fifth and seventh harmonic orders of the test waveform. Nevertheless, the results obtained from both methods satisfy the IEEE standard.

(a) time-domain waveform  
(b) spectrum

Fig. 7. Test current for Test Case 2

Table 2. Case-2 comparison between the harmonic components detected by the two methods

| Order | Actual (magnitude) | DQ magnitude | DQ error | ATS magnitude | ATS error |
|-------|--------------------|--------------|----------|---------------|----------|
| 5     | 0.80 (4%)          | satisfy std. | satisfy std. | satisfy std. | satisfy std. |
| 7     | 8.57 (42.85%)      | 8.61         | 0.47%    | 8.57          | 0.01%    |
| 11    | 3.64 (18.2%)       | 3.67         | 0.83%    | 3.64          | 0.03%    |
| 13    | 1.54 (7.7%)        | 1.55         | 0.65%    | 1.54          | 0.07%    |
| rms error |          | 0.67%        |          | 0.04%        |

Fig. 8. Corresponding waveforms and spectrum due to the DQ harmonic detection for Test Case 2
3.2 Test Case 2

In this test case, only the fifth-order harmonic passes the standard and is not needed to be identified. The results of this test give the similar conclusion to the previous test case. The ATS detection method is noticeably better in the harmonic detection in such a way that a compensating current from the active power filter is able to suppress any chosen harmonic orders completely, and also the transient response of the compensated current as shown in figures 8 and 9. Note again that the results obtained from both methods satisfy the standard.

4 Conclusions

This work proposes a new harmonic detection algorithm based on the ATS method. It gives less errors than the DQ method does when comparing the compensated current to the fundamental. In addition, it has the special feature to select some orders of the harmonic waveform to be eliminated. This may lead to the reduction of cost and size of the active power filter. Furthermore, the proposed method is better in transient responses. Although the DQ detection algorithm is not as good as the ATS method in this consideration, it still passes the IEEE 519-1992 standard as its THDi and spectrum.

Acknowledgement

The financial support from Suranaree University of Technology is greatly acknowledged.

References

1. IEEE Standard 141-1993: Harmonics in Power Systems.
2. Brozek, J.P.: The Effects of Harmonics on Overcurrent Protection Devices, IEEE Industry Applications Society Conference, 2 (1990) 1965-1967
3. Ho, J.M. and Liu, C.C.: The Effects of Harmonics on Differential Relay for a Transformer, IEE International Conference and Exhibition on Electricity Distribution (CIRED), 2 (2001)
4. Jung, G.H. and Cho, G.H.: New Active Power Filter with Simple Low Cost Structure without Tuned Filters, IEEE Power Electronics Specialists Conference (PESC’98), 1 (1998) 217-222
5. Takeda, M., Ikeda, K., Teramoto, A. and Aritsuka, T.: Harmonic Current and Reactive Power Compensation with an Active Filter, IEEE Power Electronics Specialists Conference (PESC ’88), 2 (1988) 1174-1179
6. Puangdownreong, D., Areerak, K-N., Srikaew, A., Sujitjorn, S. and Totarong, P.: System Identification via Adaptive Tabu Search, IEEE Int. Conf. on Industrial Technology (ICIT’02), 2 (2002) 915-920
7. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers and Operations Research 13 (1986) 553-549
8. Kulworawanichpong, T., Puangdownreong, D. and Sujitjorn, S.: Finite Convergence of Adaptive Tabu Search, ASEAN. Journal (accepted).
9. IEEE Standard 519-1992.
Active Power Filter Design by a Simple Heuristic Search

Thanatchai Kulworawanichpong, Kongpol Areerak, and Sarawut Sujitjorn

School of Electrical Engineering, Suranaree University of Technology
Nakhon Ratchasima, 30000, Thailand
sarawut@ccs.sut.ac.th

Abstract. This article proposes a new design method based on a simple heuristic technique to eliminate harmonic in power systems. It is called modified least compensation current control (MLC) method. The effectiveness of the method is verified by comparison studies among the d-q axis, the synchronous-detection, and the sliding-window Fourier analysis methods, respectively. All are regarded as group-harmonic identification methods. The results confirm the effectiveness of our proposed MLC method.

1 Introduction

At present, electronic converter technology is vital for industries. No one can deny the fact that the converter’s switching function pollutes the power system quality. This kind of pollution is often termed harmonic. It causes degraded performance of machines [1], failure of protective devices [2,3], short life-time of lamps [4], etc. Various types of filters have been used to eliminate harmonic [5,6]. Among those, active filters having a common structure as shown in fig. 1 are very efficient.

![Fig. 1. Structure of active power filters](image)

To provide proper current injection to the system to eliminate harmonic, the active filter needs the information of harmonic contents. It thus requires a harmonic identifier as its essence.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3215, pp. 695–702, 2004.
© Springer-Verlag Berlin Heidelberg 2004
To date, the following group-harmonic identification methods have been accepted about their effectiveness: the d-q axis (DQ), the synchronous-detection (SD), and the sliding-window Fourier analysis (SWF) methods [7-9], respectively. We thus conducted our comparison studies among them as our results reported herein. Section 2 of this article describes our proposed new method for harmonic elimination so called MLC method. Its structure possesses a nonlinear element whose characteristic needs to be identified by a simple heuristic search. Sections 3 and 4 provide the results, discussions, and conclusions, respectively.

2 MLC Method for Harmonic Elimination

Zhou and Li, 2000 [10], introduced the approach of injecting the filter’s current into the system to compensate for harmonically contaminated load current. It results in a sinusoidal supply current of 50 Hz frequency. To achieve this, the prediction of the fundamental component must be accurate. It may further require an accurate phase detection for unity power factor control. We propose a modified method based on Zhou and Li’s approach so called modified least compensation current control (MLC) method for harmonic elimination. Our method is represented by the diagram in figure 2. Fundamentally, it can be viewed as the simplest structure shown in figure 3.

![Fig. 2. Power active filter using MLC method](image)

![Fig. 3. Simple diagram representing a power system](image)
Equation (1) describes the relationship of the currents, in which $i_S$ is supply current, $i_L$ is load current, and $i_F$ is filter current. It is expected that the supply current contains only the fundamental component as given in equation (2), where $I_s$ is the magnitude, and $\omega_S$ is the frequency of the fundamental, respectively. We accept the fact that the prediction error can occur. So, the equation (1) when the prediction error ($\Delta i_S$) is taken into account can be rewritten as equation (3).

\[
\begin{align*}
 i_S &= i_L - i_F \quad (1) \\
 i_S &= I_s \sin \omega_S t \quad (2) \\
 (i_S + \Delta i_S) &= i_L - (i_F + \Delta i_F) \quad (3)
\end{align*}
\]

Hence,

\[
\Delta i_S = \Delta I_S \sin \omega_S t = -\Delta i_F \quad (4)
\]

The amount of compensating current injected into the system at any instant ($t=T_{samp}$) is

\[
\Delta i_{F,T_{samp}} = -\Delta I_S \sin \omega_S T_{samp} \quad (5)
\]

that is

\[
\Delta i_{F,T_{samp}} \propto \Delta I_S \quad (6)
\]

Control of compensating current injection can be achieved by using a P-controller having saturation characteristic as shown in figure 4. Due to the non-linear loads, and the stability requirement, the search for the proper characteristic of the controller is an important issue to be discussed in the following section.

- **Simple heuristic parameter tuning for the MLC method**

$I_{S,max}$ and $k_p$ are two key parameters resulting in the shape of the compensated current. In this work, only $k_p$ is varied while $I_{S,max}$ is kept constant. The adjustment is based on a simple heuristic search technique, which can be described by the following rule.
IF $i_s$ is not purely sinusoidal \hspace{1cm} \textbf{THEN} \hspace{0.5cm} k_p$ is reduced by the factor of 2

To demonstrate this technique, the compensated waveforms resulting from the above IF-THEN rule are shown in figure 5.

![Compensated waveforms with $k_p$ adjustment](image)

**Fig. 5.** Compensated waveforms with $k_p$ adjustment

### 3 Simulation Results

This section describes two test case scenarios through computer-based simulation as follows.

#### 3.1 Test1

Figure 6 gives a waveform for the first test together with its spectrum. The four candidates are employed to eliminate harmonic components. The results show that the total

![Current waveform for Test 1](image)

(a) time-domain waveform \hspace{1cm} (b) spectrum

**Fig. 6.** Current waveform for Test 1
Fig. 7. Compensated currents for Test 1
Harmonic distortion of the compensated current, THDi, is 1.46%, 0.29%, 0.67% and 0% for the SD, DQ, SWF and MLC methods, respectively. In addition, the spectrum of each case is shown in figure 7.

As can be seen, the disadvantage of the SD, DQ and SWF methods is that they have the delay within the first cycle, the current is fully compensated for without any delay via the proposed method.

### 3.2 Test2

Figure 8 gives a waveform for the second test together with its spectrum. The four candidates are employed as the first test. The results show that the total harmonic distortion of the compensated current, THDi, is 3.13%, 0.67%, 8.04% and 0% for the SD, DQ, SWF and MLC methods, respectively. For this case, the compensated current resulting from the SWF method does not pass the IEEE19-1992 standard [11]. In addition, the spectrum of each case is shown in figure 9.

![Fig. 8. Current waveform for Test 2](image)

As the first test, the disadvantage of the SD, DQ and SWF methods is the delay for compensating the current at the first cycle. In contrast, when the MLC method is used, the current is compensated for with no delay and the compensated waveform is purely sinusoidal.

### 4 Conclusions

The MLC method proposed in this paper has the ability to fully eliminate all harmonic components. It is fast and simple to be used in real-time. Also, when compared with the responses compensated by the SD, DQ and SWF methods, the proposed method gives the best transient characteristic. However, the use of this method depends upon the P-controller design. It requires a careful parameter setting in such a way that the compensated current must be purely sinusoidal as its reference. Therefore, in this paper, the parameter setting is tuned through a simple heuristic search technique. With this technique, the proportional gain of the P-controller can be adjusted appropriately to guarantee the perfect compensation.
Fig. 9. Compensated currents for Test 2
5 Acknowledgement

The financial support from Suranaree University of Technology is greatly acknowledged.

References

1. IEEE Standard 141-1993: Harmonics in Power Systems.
2. Brozek, J.P.: The Effects of Harmonics on Overcurrent Protection Devices, IEEE Industry Applications Society Conference, 2 (1990) 1965-1967
3. Ho, J.M. and Liu, C.C.: The Effects of Harmonics on Differential Relay for a Transformer, IEE International Conference and Exhibition on Electricity Distribution (CIRED), 2 (2001)
4. Wakileh, G.J.: Power Systems Harmonics, Springer-Verlag, 2001
5. Lin, B.R., Yang, B.R. and Tsai, H.R.: Analysis and Operation of Hybrid Active Filter for Harmonic Elimination, Electric Power Systems Research, 3 (2002) 191-200
6. Jung, G.H. and Cho, G.H.: New Active Power Filter with Simple Low Cost Structure without Tuned Filters, IEEE Power Electronics Specialists Conference (PESC’98), 1 (1998) 217-222
7. Takeda, M., Ikeda, K., Teramoto, A. and Aritsuka, T.: Harmonic Current and Reactive Power Compensation with an Active Filter, IEEE Power Electronics Specialists Conference (PESC ’88), 2 (1988) 1174-1179
8. Chen, C.L., Lin, C.E., Huang, C.L.: The Reference Active Source Current for Active Power Filter in an Unbalanced Three-Phase Power System via the Synchronous Detection Method, IEEE Instrumentation and Measurement Technology Conference (IMTC/94), 2 (1994) 502-505
9. El-Habrouk, M. and Darwish, M.K.: Design and Implementation of a Modified Fourier Analysis Harmonic Current Computation Technique for Power Active Filter Using DSPs, IEE Proc.-Electr. Power Appl., 1 (2001) 21-28
10. Zhou, L. and Li, Z.: A Novel Active Power Filter Based on the Least Compensation Current Control Method, IEEE Trans. on Power Electronics, 4 (2000) 655 – 659
11. IEEE Standard 519-1992.
Stochastic Local Search for Incremental SAT and
Incremental MAX-SAT

Malek Mouhoub and Chonghai Wang

Department of Computer Science,
University of Regina,
3737 Waskana Parkway,
Regina SK, Canada, S4S 0A2
{mouhoubm,wang203c}@uregina.ca

Abstract. The boolean satisfiability problem (SAT) is stated as follows: given a boolean formula in CNF, find a truth assignment that satisfies its clauses. The optimization variant of SAT, called MAX-SAT, consists of finding a truth assignment that satisfies the maximum possible number of clauses in a given formula. In this paper, we present methods based on stochastic local search for solving the incremental SAT and incremental MAX-SAT problems. Given a satisfiable boolean formula in CNF, incremental SAT (respectively incremental MAX-SAT) consists of checking whether the satisfiability (respectively maximum satisfiability) is preserved when new clauses are added to the current clause set and if not, look for a new assignment that satisfies all the clauses (respectively the maximum number of clauses). Experimentation has been conducted on randomly generated incremental SAT and incremental MAX-SAT problems. The results show the efficiency of our methods to deal with large size dynamic SAT and dynamic MAX-SAT problems.

1 Introduction

A boolean variable is a variable that can have one of two values: \textit{true} or \textit{false}. If \(x\) is a boolean variable, \(\neg x\) is the negation of \(x\). That is, \(x\) is true if and only if \(\neg x\) is false. A literal is a boolean variable or its negation. A clause is a sequence of literals separated by the logical \textit{or} operator (\(\lor\)). A logical expression inconjunctive normal form (CNF) is a sequence of clauses separated by the logical \textit{and} operator (\(\land\)). For example, the following is a logical expression in CNF: \((x_1 \lor x_3) \land (\neg x_1 \lor x_2) \land \neg x_3\). The CNF-Satisfiability Decision Problem (called also SAT problem) is to determine, for a given logical expression in CNF, whether there is some truth assignment (set of assignments of true and false to the boolean variables) that makes the expression true. For example, the answer is “yes” for the above CNF expression since the truth assignment \(x_1 = \text{true} , x_2 = \text{true} , x_3 = \text{false}\) makes the expression true. SAT and MAX-SAT problems are fundamental to the theory of NP-completeness. Indeed, using the concept of “polynomial-time reducibility” all NP-complete problems can be polynomially reduced to SAT\(^1\). This means that any new technique for SAT or MAX-

\(^1\) We will refer the reader to the paper published by Cook [6] proving that if CNF-Satisfiability is in P, then P = NP.
SAT problems will lead to general approaches for thousands of hard combinatorial problems. One important issue when dealing with SAT problems (respectively MAX-SAT problems) is to be able to maintain the satisfiability (respectively the maximum satisfiability) of a propositional formula anytime a conjunction of new clauses is added. That is to check whether a solution to a SAT problem (respectively optimal solution of a MAX-SAT problem) continues to be a solution anytime a set of new clauses is added and if not, whether the solution can be modified efficiently to satisfy the old formula and the new clauses.

In this paper we will discuss the applicability of stochastic local search (SLS) algorithms for solving SAT and MAX-SAT problems in an incremental way. This choice is motivated by the fact that the underlying local search paradigms are well suited for recovering solutions after local changes (addition of constraints) of the problem occur. Experimental evaluation of the different methods based on SLS, we propose for solving SAT and MAX-SAT problems, have been performed on randomly generated dynamic SAT and dynamic MAX-SAT instances. The results show the efficiency of our methods to deal with large size SAT and MAX-SAT problems. Note that related work on solving SAT problems in an incremental way has already been reported in the literature. These methods however rely solely on stochastic local search [1,2] or branch and bound [3,4,5]. To our best knowledge, we are not aware of any proposed method for solving incremental MAX-SAT problems.

In the next section, we define the dynamic SAT and dynamic MAX-SAT problems and present a general framework for solving these problems. In section 3, we present the implementation of the SLS based method we use for solving dynamic SAT and dynamic MAX-SAT. Section 4 is dedicated to the empirical experimentation evaluating our methods. Concluding remarks and possible perspectives are finally presented in section 5.

2 Solving Dynamic SAT and Dynamic MAX-SAT

2.1 Solving Dynamic SAT

We define a dynamic SAT problem (DSAT) as a sequence of static SAT problems \(\text{SAT}_0, \ldots, \text{SAT}_i, \text{SAT}_{i+1}, \ldots, \text{SAT}_n\) each resulting from a change in the preceding one imposed by the “outside world”. This change can either be a restriction (adding a new set of clauses) or a relaxation (removing a set of clauses because these later clauses are no longer interesting or because the current SAT has no solution). In this paper we will focus only on restrictions. Let us assume that we have the following situation:

\[ \text{SAT}_{i+1} = \text{SAT}_i \land \text{NC} \]

where \(\text{SAT}_i\) is the current SAT formula, \(\text{NC}\) is a new set of clauses to be added to \(\text{SAT}_i\), and \(\text{SAT}_{i+1}\) is the new formula obtained after adding \(\text{NC}\). Both \(\text{SAT}_i\) and \(\text{NC}\) (and by consequence \(\text{SAT}_{i+1}\)) are defined on a set \(X\) of boolean variables. Asuming that \(\text{SAT}_i\) is satisfiable, the goal here is to check the consistency of \(\text{SAT}_{i+1}\) when adding the new set of clauses denoted by \(\text{NC}\). To do so, we have defined the following procedure:

1. If \(x \land \neg x\) is contained in \(\text{NC}\), return that \(\text{NC}\) is inconsistent. \(\text{NC}\) cannot be added to \(\text{SAT}_i\).
2. Simplify NC by removing any clause containing a disjunction of the form \(x \lor \neg x\).

3. Let \(NC = NC_1 \land NC_2\), where \(NC_1\) is the set of clauses, each containing at least one variable that appears in \(SAT_i\), and \(NC_2\) the set of clauses that do not contain any variable that appears in \(SAT_i\) or \(NC_1\). Let \(SAT_i = S_1 \land S_2\) where \(S_1\) is the set of clauses, each containing at least one variable that appears in \(NC\) and \(S_2\), the set of clauses that do not contain any variable that appears in \(NC\) or \(S_1\). \(S_2\) will be discarded from the rest of the procedure since any assignment to the variables of \(NC\) will not affect the truth assignment already obtained for \(S_2\).

4. Assign the truth assignment of \(SAT_i\) to \(NC_1\). If \(NC_1\) is satisfiable go to 7.

5. Using SLS flip the variables of \(NC_1\) that do not appear in \(S_1\). If \(NC_1\) is satisfied go to 7. Else record the minimum number of non satisfied clauses found for \(NC_1\). We call \(\text{min}_{NC_1}\) this number.

6. Using SLS, look for a truth assignment for both \(S_1\) and \(NC_1\). If no such assignment is found return \(NC\) cannot be added as it will affect the consistency of \(SAT_i\).

7. Using SLS look for a truth assignment for \(NC_2\). If no such assignment is found return \(NC\) cannot be added as it will affect the satisfiability of \(SAT_i\).

2.2 Solving Dynamic MAX-SAT

The MAX-SAT problem consists of finding a truth assignment that satisfies the maximum possible number of clauses in a given formula in CNF form. Let us assume that we have the following situation: \(MSAT_{i+1} = MSAT_i \land NC\), where \(MSAT_i\) is the current MAX-SAT formula, \(NC\) is a new set of clauses to be added to \(MSAT_i\), and \(MSAT_{i+1}\) is the new formula obtained after adding the new set of clauses. Assuming that \(\text{min}_i\) is the minimum number of non satisfied clauses of \(MSAT_i\), the goal here is to find the minimal number of unsatisfied clauses of \(MSAT_{i+1}\) when adding the new set of clauses denoted by \(NC\). This can be done using the following steps:

1. Simplify NC by removing any clause containing a disjunction of the form \(x \lor \neg x\).

2. Let \(NC = NC_1 \land NC_2\), where \(NC_1\) is the set of clauses, each containing at least one variable that appears in \(MSAT_i\) and \(NC_2\) the set of clauses that do not contain any variable that appears in \(MSAT_i\) or \(NC_1\). Let \(MSAT_i = S_1 \land S_2\) where \(S_1\) is the set of clauses, each containing at least one variable that appears in \(NC\) and \(S_2\), the set of clauses that do not contain any variable that appears in \(NC\) or \(S_1\).

3. Assign the truth assignment of \(MSAT_i\) to \(NC_1\). If \(NC_1\) is satisfied go to 6.

4. Using SLS flip the variables of \(NC_1\) that do not appear in \(S_1\). If \(NC_1\) is satisfied go to 6 else record the minimum number of non satisfied clauses found for \(NC_1\). We call \(\text{min}_{NC_1}\) this number.

5. Using SLS look for the minimum number of non satisfied clauses for both \(MSAT_i\) and \(NC_1\) and update the value of \(\text{min}_i\) if necessary. The search will start with a \(\text{min}\) value equal to \(\text{min}_i + \text{min}_{NC_1}\).

6. Using SLS look for the minimum number of non satisfied clauses for \(NC_2\). Add the value obtained to \(\text{min}_i\).
3 Implementation of SLS for Dynamic SAT and MAX-SAT

One of the well known randomized local search algorithms for solving SAT problems is the GSAT procedure [7,8]. GSAT is a greedy based algorithm that starts with a random assignment of values to boolean variables. It then iterates by selecting at each step a variable, flips its value from false to true or true to false and records the decrease in the number of unsatisfied clauses. The algorithm stops and returns a solution if the number of unsatisfied clauses is equal to zero. After MAX-FLIPS iterations, the algorithm updates the current solution to the new solution that has the largest decrease in unsatisfied clauses and starts flipping again until a solution satisfying all the clauses is found or MAX-TRIES is reached. In the following, we will introduce how we implement the GSAT search method to solve the dynamic SAT problems. The procedure is described below.

```
For i=1 until MAX-TRIES do
begin
Randomly generate a truth assignment
// Initialization of the structure af
for i1=1 until numberOfClauses do
begin
pos ← the number of vars with satisfied value for this clause
if pos=0 then
   for every variable xj in this clause do
      af[j].diff ← af[j].diff + 1
   else if pos=1 and xk is the variable with satisfied value then
      af[k].diff ← af[k].diff - 1
end
for i2=1 until MAX-FLIPS do
begin
   if the assignment is satisfied then return(af)
   flip x_j // xj is the variable with the maximal decreasing number of unsatisfied clauses when flipped
   af[j].diff ← 0
   for each clause containing xj do
      pos ← the number of vars with satisfied value for this clause
      for every var xk in this clause do
         af[k].diff ← af[k].diff - 1
      else if pos = 1 and xj has unsatisfied value and xk, a variable in this clause, has satisfied value then
         af[k].diff ← af[k].diff - 1
      else if pos=0 then
         for every var xk in this clause do
            af[k].diff ← af[k].diff+1
      else if pos=2 and xj has satisfied value and xk, a variable in this clause, has satisfied value then
         af[k].diff ← af[k].diff + 1
end
end
```

To make a flip, we choose the variable that minimizes the number of unsatisfied clauses when it is flipped. af[i].diff is used to maintain the number of clauses that become satisfied if the variable xi is flipped. af[i].value stores the value of the variable xi. To solve MAX-SAT problems, we use the implementation presented in the pseudo code below with the following modification. After each try we update the best solution obtained so far (solution minimizing the maximum number of clauses) if we obtain a better one for the current try.
4 Experimentation

In this section we present the experimental tests we have conducted in order to evaluate the performance of our method for solving dynamic SAT and dynamic MAX-SAT problems. All tests are performed on a 2GHz Pentium 4 computer under Linux. All procedures are coded in C language. Since we did not find libraries providing dynamic SAT problems, we take 3-SAT instances from the well known SATLIB library\(^2\). Each dynamic SAT instance is generated from a 3-SAT one in a series of stages. At each stage, a random number of clauses is taken from the 3-SAT instance and added to the dynamic SAT one until there are no more clauses to take. In our experimentation, the number of stages is fixed to 10. The number of clauses \((N_1 \ldots N_{10})\) taken at each stage are generated as follows. \(N_1\) and \(N_2\) are randomly chosen form \([1, N/5 - 1]\). \(N_3\) and \(N_4\) will then be generated from \([1, (N - N_1 - N_2)/4 - 1]\), and \(N_5\) and \(N_6\) from \([1, (N - N_1 - N_2 - N_3 - N_4)/3 - 1]\). \(N_7, N_8, N_9\) and \(N_{10}\) will be generated in the same manner. This will ensure that the average number of clauses in each stage is almost equal to \(N/10\).

Table 1 presents the total running time in seconds needed by the dynamic method we propose (that we call dynamic GSAT) and the standard GSAT method (that we call static GSAT) that consists of looking for a solution from scratch anytime a new set of clauses is added. The results obtained (see table 1) by our method are very appealing comparing to those of the static method. This is mainly due to steps 3 to 7 of our resolution method. Indeed, anytime a new set of clauses is added to a given formula in CNF, the goal of these steps is to maintain the satisfiability of the new formula by checking the satisfiability of the new set of clauses and only the clauses related to them from the initial set. Static GSAT, however, considers the entire set of clauses anytime we add new ones.

| Test Set     | # of variables | # of clauses | Static GSAT | Dynamic GSAT |
|--------------|----------------|--------------|-------------|--------------|
| Uf50-218     | 50             | 218          | 0.004       | 0.00174      |
| Uf75-325     | 75             | 325          | 0.18        | 0.00285      |
| Uf100-430    | 100            | 430          | 0.273       | 0.00476      |
| Uf125-538    | 125            | 538          | 0.652       | 0.0066       |
| Uf150-645    | 150            | 645          | 4.021       | 0.00705      |
| Uf175-753    | 175            | 753          | 5.302       | 0.00154      |
| Uf200-860    | 200            | 860          | 2.901       | 0.0191       |
| Uf225-960    | 225            | 960          | 5.348       | 0.0213       |
| Uf250-1065   | 250            | 1065         | 14.194      | 0.0211       |

Table 2 presents the total running time in seconds needed by our method and static GSAT to maintain the maximum satisfiability of randomly generated incremental MAX-SAT problems. As mentioned in subsection 2.2 (steps 2 to 6), anytime new clauses are added, our method considers only the subset of clauses that can be affected

\(^2\)http://www.intellektik.informatik.tu-darmstadt.de/SATLIB.
by this addition (instead of the total number of clauses). This saves a lot of effort in maintaining the optimal solution anytime clauses are added (as we can see in table 2).

Table 2. Performance of dynamic GSAT on random Incremental MAX-SAT problems

| Test Set    | # of vars | # of clauses | Max #sat | Static GSAT | Dynamic GSAT |
|-------------|-----------|--------------|----------|-------------|--------------|
| Uf50-218    | 50        | 218          | 217      | 0.00335     | 0.002        |
| uf75-325    | 75        | 325          | 324      | 0.00109     | 0.00454      |
| uf100-430   | 100       | 430          | 429      | 0.0394      | 0.00812      |
| uf125-538   | 125       | 538          | 537      | 0.948       | 0.0621       |
| uf150-645   | 150       | 645          | 644      | 0.244       | 0.0235       |
| uf175-753   | 175       | 753          | 752      | 1.53        | 0.136        |
| uf200-860   | 200       | 860          | 859      | 1.132       | 0.0319       |
| uf225-960   | 225       | 960          | 959      | 2.932       | 0.109        |
| uf250-1065  | 250       | 1065         | 1064     | 31.561      | 0.139        |

5 Conclusion

In this paper we have presented a general framework based on stochastic local search for maintaining the satisfiability of SAT and MAX-SAT problems in an incremental way. Our work is of interest to a large variety of applications that need to be processed in an evolutive environment. This can be the case of applications such as reactive scheduling and planning, dynamic combinatorial optimization, dynamic constraint satisfaction and machine learning in a dynamic environment.

One perspective of our work is to deal with retraction of clauses in an efficient manner. Assume that during the search, a given clause (or a set of clauses) is removed. Would it be worthwhile to reconsider any decision made because of these clause(s) or would it be more costly than just continuing on with search. Another idea we will investigate in order to improve the performance of our general procedure for solving incremental SAT consists of processing steps 4 until 6 and step 7 of our procedure in parallel. If any of these two parallel phases fails then the main procedure will stop as the new added clauses are inconsistent with the current CNF formula. In the same way, steps 3 to 5 and step 6 of the procedure for solving incremental MAX-SAT can be processed in parallel.

References

1. H.H. Hoos and K. O'Neil. Stochastic Local Search Methods for Dynamic SAT - an Initial Investigation. In AAAI-2000 Workshop on Leveraging Probability and Uncertainty in Computation, pages 22-26, 2000.
2. J. Gutierrez and A.D. Mali. Local Search for Incremental Satisfiability, In International Conference on Artificial Intelligence, pages 986-991, 2002.
3. J.N. Hooker. Solving the Incremental Satisfiability Problem. Journal of Logic Programming, vol. 15, pages 177-186, 1993.
4. H. Bennaceur, I. Gouachi and G. Plateau. An incremental Branch-and-Bound Method for Satisfiability Problem. INFORMS Journal on Computing, vol. 10, pages 301-308, 1998.
5. Jesse Whittemore, Joonyoung Kim and Karem A. Sakallah. SATIRE: A New Incremental Satisfiability Engine. DAC 2001, pages 542-545, 2001.
6. S.A. Cook. The complexity of theorem proving procedures. In 3rd Annual ACM Symposium on the Theory of Computing, pages 151-158, 1971.
7. B. Selman and H.A. Kautz. An empirical study of greedy local search for satisfiability testing. In AAAI’93, pages 46-51, 1993.
8. B. Selman, H.A. Kautz and B. Cohen. Noise Strategies for Improving Local Search. In AAAI’94, pages 337-343. MIT Press, 1994.
Finite Convergence and Performance Evaluation of Adaptive Tabu Search

Deacha Puangdownreong†, Thanatchai Kulworawanichpong, and Sarawut Sujitjorn

School of Electrical Engineering, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand
† Seconded from Department of Electrical Engineering, Faculty of Engineering, South-East Asia University, Bangkok, 10160, Thailand

Abstract. The naïve tabu search (NTS) has been enhanced with two adaptive mechanisms namely back-tracking and adaptive search radius. The proposed search is called adaptive tabu search (ATS). The paper provides convergence and performance analyses of the ATS.

1 Introduction

The tabu search (TS) method was proposed in 1986 by Glover to solve combinatorial optimization problems [1]. Two principles of the TS method are the neighborhood search approach and the tabu list (TL), respectively. The method is often applied in the simplest form referred to as naïve tabu search (NTS) that is usually trapped by local solutions. The method has found a variety of applications such as [2-7] although a dead-lock by a local solution can occur.

We propose an enhanced version of the NTS that composes of two mechanisms: namely back-tracking, and adaptive search radius mechanisms. Moreover, the method possesses a random movement of solution findings in the preset neighborhood. These additional features have made the method more efficient and powerful. The search method has been named the adaptive tabu search (ATS) and successfully applied to identify linear and nonlinear system models [8].

The convergence analysis of the conventional TS method has been proved [9,10]. The proofs were based on the deterministic recency and frequency approaches. In this paper, a new proof is provided for the ATS method to ensure its convergence. In addition, the performance evaluation was conducted through many-thousand search trials on three nonlinear mathematical functions. These are the unsymmetrical trigonometric sum, one of Bohachevsky’s functions, and the circle function. This paper reports the finite convergence analysis, and the performance evaluation of the ATS.

2 Convergence of the ATS

2.1 Definitions

Definition 1. Let $\Omega$ be a finite search space having $n$ members ($n < \infty$).
Definition 2. Let the finite search space $\Omega$ have $k$ strictly local minima and be divided into $k$ regions denoted by $\Lambda_i$ ($i = 1, 2, \ldots, k$). Each region having a total of $w$ members contains only one local minimum, which must not be located at the boundary.

Definition 3. Let $\Psi$ be a randomly-created finite sub-space of $\Omega$, $\Psi \subset \Omega$, having $m$ members ($m < n$).

Definition 4. Let a finite sequence $S = \{x_0,i\}$, $i = 1, 2, \ldots, p$, be a collection of solution movements, $x_0$, consisting of $p$ solutions to reach the global minimum ($k < p$).

Definition 5. Let $Time(x)$ be a time consumed to visit a solution $x$ in the search space $\Omega$ and it assumes to be constant for visiting any $x \in \Omega$. That is $Time(x_i) = Time(x) > 0$ for $i = 1, 2, \ldots, n$.

Definition 6. Let $Iteration$ be a cumulative number of iterations indicating how many solutions in $\Psi$ were already visited. $Iteration$ is initialized at the start of a new sub-space exploration. After exploring all generated solutions in $\Psi$, the updated $Iteration$ is equal to $m$ and the time is $m \cdot Time(x)$.

Definition 7. Let $Count$ be a cumulative search round of sub-space explorations indicating how many sub-spaces in $\Omega$ were already explored entirely. $Count$ is initialized only once at the beginning. $Count$ is updated when all solutions in any $\Psi$ have been visited. After an entire exploration, $Count$ is equal to $p$ and the overall time consumed is $p \cdot m \cdot Time(x)$.

Definition 8. Let BT denote the back-tracking mechanism to allow the use of any previously visited local minimum recorded in the TL for generating a new starting point rather than the one just obtained.

Definition 9. Let AR denote the adaptive search radius mechanism that reduces the accessing time to a local minimum. Given that $\rho = \mu \cdot r$ is the adaptive radius where $r$ is a nominal radius and an arbitrary constant while $0 < \mu \leq 1$. The radius is used to define a neighborhood around a current solution.

2.2 ATS Algorithms

Step 1) Initialise the Tabu List (TL= $\emptyset$), $Iteration = 0$ and $Count = 0$.

Step 2) Randomly select an initial solution $x_{0,Count}$ from the search space $\Omega$ and assign it as an initial global minimum $x^*$. The time used for visiting the initial solution is $Time(x)$.

Step 3) Update $Count$ by 1, then create a sub-space $\Psi_{Count}$. Evaluate the objective function of $\forall x \in \Psi_{Count}$. Update $Iteration$ by 1 when an $x$ is examined. A solution with the minimum objective function is $x'$. When the exploration of the subspace is finished ($Iteration = m$), the cumulative time consumed is $m \cdot Time(x)$.

Step 4) If $x' < x_{0,Count}$, keep $x_{0,Count}$ in the TL and set $x_{0,Count} = x'$. Otherwise put $x'$ in the TL instead.

Step 5) Update the global minimum. $x^* = x_{0,Count}$ if $x_{0,Count} < x^*$. 
Step 6) Evaluate the termination criteria (TC) and the aspiration criteria (AC).
- Go to step 7 if TC is satisfied, otherwise repeat step 3.
- Activate the AR mechanism to speed up the searching process.
- Activate the BT mechanism if a local minimum trap occurs. Reset Iteration and repeat step 3.

Step 7) Terminate the search process. Accept the last updated \( x^* \) as the global solution.

Only a few numbers of solutions in \( \Omega \) would be randomly visited and it is sufficient to locate the global minimum by \( \text{Count} = p \) and \( p\cdot m\cdot \text{Time}(x) \) of the overall time consumed.

2.3 Proof of Global Convergence

**Theorem A.** If a total number of members, \( m \), in a sub-space \( \Psi \) is large enough to give good representatives of a neighborhood, a local minimum nearby can be found by generating a sequence of some successive sub-spaces.

**Proof.** Let \( \hat{x} \) be a strictly local minimum in a considered region, \( \Lambda(x_0) \), of \( x_0 \). That is \( f(\hat{x}) < f(x) \) for \( \forall x \in \Lambda(x_0) \) and also for \( \forall x \in N_\rho(x_0) \). This implies that both \( N_\rho(x_0) \) and sets of solutions nearby lie on the same region, \( \Lambda(x_0) \). In a similar manner as the Hill-climbing algorithm, updating a current solution leads descent direction to reach a nearby local minimum.

Given an initial solution \( x_{t=0} \) in a finite sub-space \( \Psi_t \subset \Omega \). To generate a sequence of \( x_{t+1} \), the descent property must be held to guarantee that a next move leads to a local minimum. At any current solution, there are only two possible outcomes that are either i) the solution is improved, \( f(x_{t+1}) < f(x_t) \), or ii) the solution is not improved, \( f(x_{t+1}) \geq f(x_t) \). In the ATS method, given that the neighborhood, \( N_\rho(x_t) \), of the current solution \( x_t \) is created and has a total of \( N \) members. The sub-space \( \Psi_{t+1} \subset N_\rho(x_t) \) is then randomly generated with \( m \) finite members where \( m < N \), and \( m \) is constant. This process is based on the assumption that not all members in the neighborhood give better cost than \( x_t \) does, but only \( u \) members of \( N_\rho(x_t) \) satisfy \( f(x) < f(x_t) \) where \( x \in \Psi_{t+1} \). The probability to improve the solution \( P = P(f(x) < f(x_t)) \) is given as follows.

**Case 1:** \( m > N - u \)

\[ P = 1, \text{ in this case, at least one of } m \text{ satisfies the condition.} \]

**Case 2:** \( m \leq N - u \)

In this case, there are \( \binom{N}{m} = \frac{N!}{(N-m)!m!} \) of the possible combination for randomly selecting \( m \) members out of \( N \). In addition, \( \binom{N-u}{m} = \frac{(N-u)!}{(N-u-m)!m!} \) is a total of ways that the solution is not improved. Thus, the probability of the sampling, which cannot improve the current solution, is shown as follows.
\[ P = \frac{(N-u)!(N-m)!}{N!(N-u-m)!} \] 

When \( m \) and \( N \) are both fixed, Eq. (1) depends on \( u \) only. \( u \) is large when the current solution is close to the local minimum. This search process updates the current solution with the best member in each iteration. Therefore, the solution will move towards the local minimum when the time increases. That is \( \lim_{t \to \infty} u(t) = 0 \). From Eq. (1), the probability of the event that the solution cannot be improved anymore (local minimum found) is expressed below.

\[ \lim_{t \to \infty} P(t) = \lim_{t \to \infty} \frac{(N-u(t))!(N-m)!}{N!(N-u(t)-m)!} = 1 \] 

When the process is repeatedly performed with a considerable amount of time, the probability of finding the local eventually global minimum is close to unity.

**Theorem B.** The BT mechanism leads the search process to obtain multiple local minima. Among them, one is the global minimum.

**Proof.** As previously mentioned, the random search process might fail to escape from a trap due to ineffectiveness of the algorithms. The use of some solution stored in the TL as an initial solution for the next search round enables various search directions. It increases possibility to run away from the already visited local minimum. Given \( n_{re} \) be a counter for a solution cycling. “Solution cycling” means that the search cannot escape the entrapment of the just visited local minimum, so the movement of solutions will return to the just visited local minimum at the end of the next search round. The counter is updated every time a new final solution of any search round being equal to the one previously visited and already stored in the list. Let \( n_{re, Max} \) be the maximum number allowance of the solution cycling. Therefore, the BT mechanism is activated by the following condition. If \( n_{re} < n_{re, Max} \), then continue the search whether it can eventually escape from the solution lock, otherwise, performing the BT process. Once \( n_{re} \geq n_{re, Max} \), one of the solutions recorded in the TL is selected to be a new initial solution for creating the next sub-space \( \Psi \). \( n_{re} \geq n_{re, Max} \) is an Aspiration Criteria. The BT mechanism will select a solution \( x_h \in \text{TL} \) in such a way that \( x_h = \max_{x \in \text{TL}} \| x - x_0 \| \) and the condition \( f(x_h) < f(x_0) \) must hold. After selecting the solution, set \( x_0 = x_h \) as a new initial solution for the next search round. Therefore,

i) If the local minimum \( \hat{x} \) is obtained already and the length of TL is sizeable, there exists at least one solution that is relatively close to the boundary of \( \Lambda(x_0) \). Therefore, \( \text{length(TL)} \gg 1 \rightarrow \exists x \in \text{TL} \land \| x - x_B \| < \gamma \), where \( x_B \) is a boundary point and \( \gamma \) is the maximum allowance.

ii) During the search process if a current \( x_0 \) is relatively close to boundary of \( \Lambda(x_0) \) as stated in (i), together with a certain radius \( \rho \) that is relatively large enough to be able to reach some solutions outside \( \Lambda(x_0) \), the best solution of a current \( \Psi \) can be located outside \( \Lambda(x_0) \).
 iii) With proceeding a new search from a solution found outside $\Lambda(x_0)$ according to (ii), this restarts a new descent process to reach another local minimum of a new region nearby. By repeating the procedures with all $k$ different local minima being found within a finite search time $p \cdot m \cdot \text{Time}(x) < \sum w_i \cdot \text{Time}(x)$, and with the cost-value termination criterion being completely satisfied, one of the local minima is the global minimum.

3 Performance Evaluation

The ATS was coded in MATLAB$^\text{TM}$ for running on a Pentium 4, 1.6 GHz, 256 Mbytes RAM, 40 Gbytes HD. The search was conducted against three following functions to find the global minimum. Firstly, the unsymmetrical trigonometric sum function (TSF) is expressed by Eq.(3). The global minimum is on $x = -0.26$ making $f(x) = 4.56 \times 10^{-5}$ and is used as the termination criterion. Secondly, the Bohachevsky’s function (BF) [11], Eq.(4), and thirdly, the circle function (CF) [12], Eq.(5), are used. Both functions have the global minimum at $x = y = 0$ with $f(0,0) = 0$. We use $1 \times 10^{-5}$ to approximate zero and it is set as the termination criterion for the last two cases.

$$f(x) = (\sin(x) + 2.5 \sin(2x) + 1.5 \sin(4x) + 2 \sin(8x)) + x^2 + 4.4716 \quad (3)$$

$$f(x, y) = x^2 + 2y^2 - 0.3 \cos(3\pi x) - 0.4 \cos(4\pi y) + 0.7, \quad x, y \in [-1.0, 1.0] \quad (4)$$

$$f(x, y) = (x^2 + y^2)^{\frac{1}{2}} \left[ \sin^2 \left( 50 \left( x^2 + y^2 \right)^{\frac{1}{2}} \right) + 0.1 \right], \quad x, y \in [-0.5, 0.5] \quad (5)$$

We report only the performance of the ATS because the NTS is completely unable to locate the global minimum. Five parameters are considered to influence the search performance of the ATS. They are: i) the initial search radius ($R$), ii) the number of neighborhood members ($n$), iii) the number of repetitions of solution cycling before back-tracking ($n_{\text{re_max}}$), iv) the $k$th backward solution selected by the back-tracking mechanism ($k^{\text{th}}$ backward selection), and v) the percentage of search radius reduction compared to the radius before adaptation.

The first four tests begin with tuning four parameters ($R$, $n$, $n_{\text{re_max}}$ and $k^{\text{th}}$ backward selection), where the search radius is non-adaptive. Each parameter setting is carried out with the maximum of 1,000 trials. It starts with a random initial solution generated by MATLAB. It stops when either of the following termination criteria is met: i) the maximum search round of 10,000, or ii) the cost function $\leq \varepsilon$ (a very small number to approximate zero).

The setting from the first four tests that gives the best result is applied to conduct the fifth test, in which the search radius is adaptive. The adaptive radius scheme is set to have three steps of reduction as: i) if [cost function $< 10^{-1}$] then $R^{(\text{new})} = R^{(\text{old})}/DF$; ii) if [cost function $< 10^{-2}$] then $R^{(\text{new})} = R^{(\text{old})}/DF$; and iii) if [cost function $< 10^{-3}$]...
then \( R^{(\text{new})} = R^{(\text{old})}/DF \), where \( R^{(\text{old})} \) and \( R^{(\text{new})} \) are the search radius before and after adaptation, and \( DF \) is the factor of radius reduction and it is assigned by the following values: 10, 15, 20, 25, and 30% of the current radius. The parameters for the fifth test obtained from Tables 1-4 are given as follows: TFS–\( \{R = 2.5\%, n = 30, \text{n}_\text{re}\_\text{max} = 5, \text{and } k^{\text{th}} = -2\} \); BF–\( \{R = 10.0\%, n = 30, \text{n}_\text{re}\_\text{max} = 5, \text{and } k^{\text{th}} = -5\} \); and CF–\( \{R = 7.5\%, n = 40, \text{n}_\text{re}\_\text{max} = 5, \text{and } k^{\text{th}} = -5\} \). The results of this test are summarized in Table 5. It shows that 20-25% reduction of the search radius gives good performance in terms of speed and convergence.

4 Conclusions

The convergence analysis and the performance evaluation of the ATS method have been presented in this paper. The global convergence can be guaranteed. We present the ATS performance in terms of function minimization. The NTS is completely unable to locate the global solution while the ATS can. To apply the ATS successfully, one is recommended to evaluate its performance against a set of interested problems such that its parameters rendering fast search could be identified.

| R (%) | Average search round | Average search time (s) | No of successful trials |
|-------|----------------------|------------------------|------------------------|
|       | TSF      | BF    | CF   | TSF      | BF    | CF   | TSF      | BF    | CF   |
| 2.5   | 459.45   | 7923.30 | 9341.4 | 1.31     | 48.51  | 52.43 | 1000     | 203    | 214   |
| 5.0   | 936.41   | 7156.60 | 6610.2 | 2.81     | 41.18  | 38.87 | 1000     | 296    | 405   |
| 7.5   | 897.83   | 3876.60 | 3742.3 | 3.28     | 20.95  | 20.63 | 1000     | 757    | 885   |
| 10.0  | 924.17   | 1353.10 | 4878.8 | 3.92     | 6.06   | 25.40 | 1000     | 1000   | 816   |
| 12.5  | 992.33   | 2263.70 | 5955.1 | 4.78     | 10.60  | 33.03 | 1000     | 988    | 700   |
| 15.0  | 933.06   | 3071.40 | 6796.2 | 5.02     | 14.91  | 37.60 | 1000     | 954    | 548   |
| 20.0  | 1035.50  | 4832.40 | 8038.0 | 6.78     | 24.59  | 47.148| 1000     | 808    | 363   |

| n     | Average search round | Average search time (s) | No of successful trials |
|-------|----------------------|------------------------|------------------------|
|       | TSF      | BF    | CF   | TSF      | BF    | CF   | TSF      | BF    | CF   |
| 10    | 1473.5   | 4135.10 | 7134.4 | 2.14     | 10.33  | 21.82 | 998      | 893    | 526   |
| 20    | 707.15   | 2203.30 | 4802.4 | 1.44     | 7.72   | 20.23 | 1000     | 987    | 804   |
| 30    | 459.45   | 1353.10 | 3742.3 | 1.31     | 6.06   | 20.63 | 1000     | 1000   | 885   |
| 40    | 357.17   | 1089.10 | 3029.3 | 1.40     | 6.37   | 19.37 | 1000     | 1000   | 910   |
| 50    | 287.32   | 904.08  | 2928.5 | 1.33     | 6.49   | 25.06 | 1000     | 1000   | 873   |
| 60    | 230.81   | 802.53  | 2304.2 | 1.26     | 6.70   | 22.32 | 1000     | 1000   | 908   |
Table 3. Effects of n_re_max

| n_re_max | Average search round | Average search time (s) | No of successful trials |
|----------|----------------------|-------------------------|-------------------------|
|          | TSF      | BF     | CF     | TSF      | BF     | CF     | TSF      | BF     | CF     |
| 5        | 472.52  | 1310.90 | 3029.3 | 1.36   | 5.91   | 19.37 | 1000 | 1000 | 910 |
| 10       | 473.65  | 1322.10 | 3279.4 | 1.43   | 5.95   | 22.15 | 1000 | 1000 | 871 |
| 15       | 477.45  | 1353.10 | 3438.3 | 1.43   | 6.06   | 23.91 | 1000 | 1000 | 858 |
| 20       | 475.71  | 1518.80 | 3466.4 | 1.44   | 6.61   | 24.12 | 1000 | 998  | 851 |
| 25       | 485.43  | 1438.30 | 3360.0 | 1.42   | 6.51   | 23.37 | 1000 | 1000 | 862 |

Table 4. Effects of the kth backward selection

| kth      | Average search round | Average search time (s) | No of successful trials |
|----------|----------------------|-------------------------|-------------------------|
|          | TSF      | BF     | CF     | TSF      | BF     | CF     | TSF      | BF     | CF     |
| -1       | 461.60  | 1498.40 | 3455.1 | 1.49   | 6.33   | 24.09 | 1000 | 998  | 866 |
| -2       | 493.99  | 1488.40 | 3364.4 | 1.47   | 6.68   | 23.65 | 1000 | 998  | 871 |
| -3       | 477.19  | 1478.40 | 3253.7 | 1.57   | 6.90   | 22.21 | 1000 | 998  | 867 |
| -4       | 464.33  | 1586.30 | 3115.4 | 1.52   | 7.31   | 21.45 | 1000 | 998  | 884 |
| -5       | 470.31  | 1462.80 | 3029.3 | 1.53   | 6.21   | 19.37 | 1000 | 999  | 910 |

Table 5. Effects of reduced R

| Reduced R | Average search round | Average search time (s) | No of successful trials |
|-----------|----------------------|-------------------------|-------------------------|
|           | TSF      | BF     | CF     | TSF      | BF     | CF     | TSF      | BF     | CF     |
| 10%       | 11.07   | 24.36  | 1195.4 | 0.03   | 0.09   | 8.70  | 1000 | 1000 | 892 |
| 15%       | 13.36   | 26.18  | 1200.9 | 0.04   | 0.10   | 10.52 | 1000 | 1000 | 887 |
| 20%       | 17.85   | 30.16  | 600.94 | 0.05   | 0.13   | 4.20  | 1000 | 1000 | 942 |
| 25%       | 22.27   | 38.41  | 601.14 | 0.06   | 0.16   | 4.42  | 1000 | 1000 | 940 |
| 30%       | 34.64   | 64.24  | 802.92 | 0.11   | 0.31   | 7.15  | 1000 | 1000 | 914 |

References

1. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers and Operations Research 13 (1986) 533-549
2. Glover, F., Mulvey, J.M., Hoyland, K.: Solving dynamic stochastic control problems in finance using tabu search with variable scaling. Statistics and Operations Research Technical Report SOR-94-13, Princeton University, (1994)
3. Zhang, G., Habenicht, W., SpieB, W.E.L.: Improving the structure of deep frozen and chilled food chain with tabu search procedure. Journal of Food Engineering 6, 1 (2003) 67-79
4. Silva, E.L.D., Areiza, J.M.O., Oliveira, G.C.D., Binato, B.: Transmission network expansion planning under a tabu search approach. IEEE Trans Power Systems 16, 1 (2001) 62-68
5. Kulworawanichpong, T., Sujitjorn, S.: Optimal power flow using Tabu search. IEEE Power Engineering Review 22, 6 (2002) 37-40
6. Lin, B., Miller, D.C.: Application of tabu search to model identification. AIChE Annual Meeting Los Angeles. (2000)
7. Cordeau, J-F., Laporte, G.: A tabu search heuristic for the static multi-vehicle dial-a-ride problem. Transportation Research Part B: Methodological. 37, 6 (2003) 579-594
8. Puangdownreong, D., Areerak, K-N., Srikaew, A., Sujitjorn, S., Totarong, P.: System Identification via Adaptive Tabu Search. Proc. IEEE Int. Conf. on Industrial Technology (ICIT’02) 2 (2002) 915-920
9. Glover, F., Hanafi, S.: Finite Convergence of Tabu Search. Proc. MIC’2001-4th Metaheuristics Int. Conf, (2001) 333-336
10. Hanafi, S.: On the Convergence of Tabu search. J. Heuristics 7 (2000) 47-58
11. Bohachevsky, I.O., Johnson, M.E., Stein, M.L.: Generalized simulated annealing for function optimization. Technometrics 28, 3 (1986) 209-218
12. Miller, S.D., Marchetto, J., Airaghi, S., Koumoutsakos, P.: Optimization based on bacterial chemotaxis. IEEE Trans. Evol. Comput. 6 (2002) 16-29
Knowledge-Based Method to Recognize Objects in Geo-Images

Serguei Levachkine, Miguel Torres, Marco Moreno, and Rolando Quintero

Geoprocessing Laboratory (GEOLAB)
Centre for Computing Research (CIC) - National Polytechnic Institute (IPN)
{pallych, mtorres, marcomoreno, quintero}@cic.ipn.mx

Abstract. We present an approach to color image segmentation by applying it to recognition and vectorization of geo-images (satellite, cartographic) using knowledge-based learning and self-learning system. This approach exploits the user’s experience providing the knowledge domain in the form of the prescribed feature-attribute set. That is a simultaneous segmentation-recognition system when segmented geographical objects of interest ( alphanumeric, punctual, linear, and area) are labeled by the system in same, but are different for each type of objects, gray-level values. We exchange the source image by a number of simplified images (composites). Every composite is associated with certain image feature. Some of the composites that contain the objects of interest are used in the following object detection-recognition by means of association to the segmented objects corresponding “names” from the user-defined subject domain. The specification of features and object names associated with perspective composite representations is regarded as a type of knowledge domain, which allows automatic or interactive system’s learning. Additionally, we describe the fine-to-coarse scale method of the raster-to-vector conversion in which the “knowledge” of cartographic patterns into small-scale map aids in recognizing the corresponding patterns into large-scale map of the same territory. The results of gray-level and color image segmentation-recognition-vectorization are shown.

1 Introduction

Segmentation is fundamental to the field of image processing because it is used to provide the basic representation on which understanding algorithms operate. The ability to build up a representation from individual pixels of an image, which exploits relationships such as local proximity and highlights the structures of the underlying components, is important for the extraction of features during interpretation and recognition. In general, the nature of this representation is application dependent [1]. In the present work, we developed an application independent segmentation.

Up to the now a great variety of segmentation algorithms for gray-level images has been proposed. The majority of color segmentation approaches are based on monochrome segmentation approaches operating in different color spaces [2]. Gray-level segmentation methods can be directly applied to each component of a color
space; thus, the results can be combined in some way to obtain a final segmentation result. However, one of the problems is how to employ the color information as a whole for each pixel. When the color is projected onto three RGB color components, the color information is so scattered that the color image becomes simply multispectral image and the color information that humans can perceive is lost [3]. An alternative solution presented in this work is invariant image representation (composite images, or simply composites) that does not depend on the choice of particular color space. The processing of a color image is individual segmentation by each color component into image meaningful (or invariant with respect to a given, unnecessary color feature) regions, first and, then – image’s joint segmentation-recognition (or “objects of interest designing”). Moreover, the prescribed set of features is regarded as a type of knowledge domain. The composite image technique includes object-fitting compact hierarchical segmentation, binarization of segmented images, and synthesis of binary representations. The main goal of image synthesis consists of the object linking by its associated names. In the following sections, we build up composite image representations based on object-fitting compact hierarchical segmentation; see also [4][7].

2 Object-Fitting Compact Hierarchical Segmentation, Recognition and Vectorization

In our method, the image segments obtained as the result of the iterative procedure of successive increasing of the admitted gray-level and color thresholds in the segment merging form subsequently increasing compact hierarchical structure of the flat segment networks. Each segment of this structure can have the ancestor or a descendant. Thus obtained structure is called the adaptive dynamic data structure. The segment of image is a node of the spatial structure, whose attributes are primary numbers defined by the averages of color/gray-level segment’s features and by a set of pixels that represent the area and the shape of the segment. This allows organizing the object-oriented identification of semantically meaningful image’s regions. Our system has the interactive procedure of compulsory restructuration of the segment relationships as a tool of the semantic analysis of visual data. In other words, the system’s learning and self-learning with the prescribed set of associative identifiers are possible in the interactive regime.

Successive segment merging by some criteria leads to the segment structuring in a multi-level hierarchy that represents by the dynamic trees [7]. This hierarchy or the multi-level image partition is an efficient method of semantic identification of the image’s objects. The relationships between the dynamic tree nodes indicate the neighboring semantically meaningful regions. Because the image’s regions are identified by the corresponding tree nodes, the neighbor relation between them can be completely defined by a table of adjacency. A modification (elimination of some edges, i.e. segment relationships) of the dynamic tree allows modifying the resulting region and thus more exact object detection is reached. Each level of the tree of segments can be considered as alternative image interpretation in different semantics (see Fig. 3 and Fig. 5, Section 3). Adaptive dynamic tree structure regards the search for
meaningful objects as the combination of the object features that fit to the corresponding ranges and the following analysis of all admitted areas. This makes possible to use the automatic learning algorithms when the set of searched objects is given and it is necessary to define only the corresponding feature ranges (this is natural supposition for geo-images [4]). The learning process can be organized as follows. The user selects the appropriate level of segment hierarchy and points out the set of the suitable areas. These areas can be defined by combining (merging/splitting) the corresponding segments. Then the program computes the characteristics of the located segments and relationships between them and establishes the formal criterion of the search for the similar objects.

Object-fitting compact hierarchical segmentation is a sequence of embedded partitions without repetition of composed segments in different partitions. A partition is obtained by iterative segment splitting or merging. In the merging mode, any segment in each iteration merges into the nearest adjacent segment. The merging of segments into objects defines the image semantics. The image’s semantics in this context corresponds to the association of segment fields of different hierarchical levels being identified with identifying conceptions from the subject domain. For example, detection of a segment identifying a coastline or highway becomes semantically meaningful. Further, this set of segments is renamed as “coastline”, “highway”, etc. (Fig. 1). Segmented and recognized objects are subsequently vectored by applying a method described in [4] to be finally included into Geographic Information Systems (GIS).

\[ \frac{N}{N_0} = \left( \frac{V}{V_0} \right)^\alpha \]

In equation (1), \( N_0, V_0 \) denote number of segments and compressed volume of the source image respectively; \( \alpha \) is some real coefficient. We obtained that in the case of object-fitting compact hierarchical segmentation the exponent \( \alpha \) is approximately 2.9.
Knowledge-Based Method to Recognize Objects in Geo-Images

Note that for non-adaptive pyramidal segmentation [8] $\alpha$ is approximately 1.4. It is known that the volume of compressed data is closely related to the amount of information into data.

![Graphs](a) (b)

**Fig. 2.** Linear dependences: (a) Number of segments $N$ on iteration number $i$, (b) Compressed image volume $V$ on iteration number $i$

### 3 Composite Image Representation

We have found that in addition to natural decomposition (e.g. $R$, $G$ or $B$ – component splitting) of color images, artificial representations can also be useful for objects of interest detection-recognition [4][7]. Our approach provides composite representations of the source image by means of reduced number of color or tone components and segments. Composite image representation is a sequence of binary representations, which are packed into different bit planes. These binary images are the result of two-valued classification of source image by some feature (intensity, area, invariant moments, etc. [4]).

A bit component of composite image (Fig. 3) computes by means of global dynamic thresholding of the current segmented image. The threshold is equal to the average all over the image intensity, geometric or other feature, denoted by $I^0$. To threshold the image, $I^i$ is compared with its average over the pixels of each segment, denoted by $I^0$, as follows: $I^0 \geq (\leq) \xi I^0$, where $\xi$ is a tuning parameter.

![Composite Images](Fig. 3. Bit components of composite images obtained by means of dynamic adaptive thresholding of Lena’s source and segmented images)

To compose these images, we also used the geometric features from the feature set [4] in addition to intensity feature. The bit components are packed in the resulting representation, where the extrema of intensity indicate the pixels associated with unchanged binary feature. Essentially, the composite images form a “book” in which
the objects of interest can be found on appropriated page(s). Thus, a “page number” defines the method of thresholding and the tuning parameter $\xi$.

### 3.1 Color Composites

Compact hierarchical segmentation of a color image is performed by each independent color component ($R$, $G$, and $B$) considering these as semi-tone images. In this way, coinciding intensities of resulting $R$, $G$, and $B$ composite images indicate the segments of equal color with respect to using feature. This can be used for invariant color image description. As a rule, compact hierarchical image segmentation implies that color segments are enlarged simultaneously in accordance to regularities presented in Section 2, Fig. 2. The method requires significant operative memory space. To overcome this disadvantage, we used special data organization in the form of irregular dynamic trees (Section 2 and [4][7]) that provides optimal in memory space computing for the successive scanning of image scales. The user needs only to make adequate choice to carry out task features from prescribed feature set (knowledge domain).

### 3.2 Applications of Composites

Image two-valued classification (binarization) remains one of the most important tasks in modern recognition methods. In the frameworks of the composite image technique, we obtained a few solutions for this task [4][7]. By applying composites, we are able to extract cartographic data using $R$, $G$, and $B$ components of full-size color raster-scanned image (Fig. 4). Fig. 5 shows how our method insures object detection in the task of recognition of inclined digits embedded in graphics (note that this is old and very difficult problem that has been attracted much attention by image processing specialists [1][4][6]). This illustrates that each composite image contains machine-treatable bit-planes for target object detection and also purposeless bit-planes.

![Fig. 4. Cartographic pattern retrieval from a color map image 1082 x 1406 pixel (extreme left)](image-url)

![Fig. 5. Bit-planes suitable for digit recognition (left side) and other purposes (right side)](image-url)
4 Fine-to-Coarse Scale Method

In the following, we shall present the main ideas of the F2CS (Fine-to-Coarse Scale) method [10]. This method has been originated from unsolved problem of the vector description of raster objects. To date, it is hard to see the ways to obtain even a partial, but general, i.e. applied to any type of raster objects, solution of the problem. In certain sense, the F2CS method represents a promising alternative.

Let us suppose that we have a vector image $I_1$ (or already recognized raster image) in scale $1:s_1$ of a given territory $T$ and a raster image $I_2$ of $T$ to be vectored (recognized) in scale $1:s_2$, and $s_1 > s_2$ (e.g. $s_1 = 100,000$ and $s_2 = 50,000$). Our goal is to use the information from $I_1$ in vectorization of $I_2$. Note that $I_1$ can be considered as a “generalization” of $I_2$: $I_1 = G(I_2)$, i.e. if an object $O_2 \in I_2$, then there can exist $O_1 \in I_1$, such that $O_1 = G(O_2)$. We denote $\Omega$ - the set of all such objects $O_2$ from $I_2$ and $\Theta$ - the compliment of $\Omega$ in $I_2$: $I_2 = \Omega \cup \Theta$. We also put $\omega = G(\Omega)$ and note that $\omega \subseteq I_1$. Obviously, to vector objects from $\Omega$ and $\Theta$ we need two different strategies. The objects from $\Omega$ can be vectored, using the features (position, color or colors, shape, etc.) of the vector objects from $\omega$. After, the objects of $\Omega$ have been vectored we can vector the objects from $\Theta$ by one of the recognition modules [4] as a “new” cartographic material. To perform “vectorization-interpretation” or “segmentation-recognition” with F2CS we used a conceptual clustering [9] based on the set of prescribed object features [4] and the concept of the associated to the image function [10]. An example of application of the F2CS method is shown in Figure 6.

![Fig. 6. Recognition of a punctual cartographic pattern. a) Original image in scale 1:s, with the input information. b) Image in scale 1:s, (s < s) obtained after application of conceptual clustering. c) Recognition of the “coarse” pattern (function) “Palm”](image)

5 Conclusion

The problem of how and to what degree the semantic information should be employed in image segmentation has led us to the conception of composite image representation for mutual object detection-recognition at low level processing. We conjecture that modern segmentation systems must support object detection-interpretation-recognition, starting at low level, memorizing results at the intermediate level, and effectively communicating these results to the high level. The approach proceeding from this conjecture is called composite image technique. The idea is to prepare the source image as much as possible for subsequent high-level processing of image regions. In most of the existing color image segmentation approaches, defini-
tion of a region is based on similarity of color. This assumption often makes it difficult for any algorithms to separate the objects with highlights, shadows, shadings or texture, which cause inhomogeneity of colors of the object’s surface. To our knowledge, this method is one of the first attempts to design a segmentation-recognition computer system for complex color images of arbitrary type (Cf. [5]).

An alternative solution of the problem followed in this work is defining image regions by quantitative, qualitative, and nominal features (in addition to color feature), which on the whole render the user's knowledge domain. We believe that this is a kind of advanced simulation of the human’s visual perception. However, it is necessary to emphasize that for optimization of labor-intensive program training a strong formalization of composite image technique is now required. We are under way to solve this problem.

Automatic interpretation of color images presents certain difficulties for state-of-the art in image processing and artificial intelligence. We believe that only a system approach to the problem can be fruitful. In the context of the present work, this means first, decomposition of source image by multiple hierarchical components to achieve a stable, accurate representation in the presence of degraded images (raw data structuration stage). Second is the segmentation with mutual recognition of appropriate primitives (compression stage) and, if required, their vectorization to be directly included into application-oriented database, e.g. GIS (application-dependent stage). Finally, there is the development of a unified knowledge-based trainable and self-trainable system with optimal human-machine interaction for color image treatment.

References

1. Doermann, D.S.: An Introduction to Vectorization and Segmentation. Lecture Notes in Computer Science, Vol. 1389. Springer-Verlag, Berlin Heidelberg New York (1998) 1–8
2. Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.: Color Image Segmentation: Advances and Prospects. Pattern Recognition, Vol. 34, No. 12. (2001) 2259-2281
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. 3rd edn. PTR, NJ: Prentice Hall, New York (2002)
4. Levachkine, S.P., Velázquez, A., Alexandrov, V.V., Kharinov, M.V.: Semantic Analysis and Recognition of Raster-Scanned Color Cartographic Images. Lecture Notes in Computer Science, Vol. 2390. Springer-Verlag, Berlin Heidelberg New York (2002) 178-189
5. Kumar, K.S., Desai, U.B.: Joint Segmentation and Image Interpretation. Pattern Recognition, Vol. 32, No. 4. (1999) 577-589
6. Levachkine, S.P., Velázquez, A., Alexandrov, V.V.: Color Image Segmentation using False Colors and its Applications to Geo-Images Treatment: Alphanumeric Character Recognition. Proc. IEEE International Geosciences and Remote Sensing Symposium (IGARSS 2001) 9-13 July 2001, Sydney, Australia (2001) (IEEE Catalog Number (CD-ROM): 01CH37217C; Library of Congress Number: 01-087978; ISBN CD-ROM: 0-7803-7033-3)
7. Levachkine, S., Alexandrov, V.: Semantic-Mind Analysis and Object-Oriented Data Integration of Information Flows: A Primer. In: Levachkine, S., Serra, J., Egenhofer, M. (eds.). Proc. 2nd Int. Workshop on Semantic Processing of Spatial Data (GEOPRO 2003), November 4-5 2003, Mexico City, Mexico (2003) 11-21 (ISBN 970-36-0103-0)
8. Alexandrov, V.V., Gorsky, N.D.: Image Representation and Processing: A Recursive Approach. Mathematics and Its Applications, Vol. 261. Kluwer Academic Publishers, Dordrecht Boston London (1993)
9. Martínez-Trinidad, J.F., Guzmán-Arenas, A.: The Logical Combinatorial Approach to Pattern Recognition, an Overview through Selected Works. Pattern Recognition, Vol. 34, No. 1. (2001) 741-751
10. González-Gómez, E., Levachkine, S.: Fine-to-Coarse Scale Method of Color Cartographic Images Recognition (to appear)
Fast Design of 2-D Narrow Bandstop FIR Filters for Image Enhancement

Pavel Zahradník¹ and Miroslav Vlček²

¹ Department of Telecommunications Engineering
Czech Technical University Prague
Technická 2, CZ-166 27 Praha,
Czech Republic
Phone: +420-2-24352089,
Fax: +420-2-33339810

²Department of Applied Mathematics
Czech Technical University Prague
Konviktská 20, CZ-110 00 Praha,
Czech Republic
Phone: +420-2-24890720, Fax:+420-2-24890702
{zahradni,vlcek}@fel.cvut.cz

Abstract. Novel approach in the design of 2-D extremely narrow bandstop FIR filters is presented. The completely analytical design method is based on the 1-D optimal bandstop FIR filters. The 1-D FIR optimal bandstop filters are based on Zolotarev polynomial. Closed form formulas for the design of the filters are presented. One example demonstrates the design procedure. Application of the 2-D narrow bandstop FIR filter for image enhancement is presented.

Keywords: two-dimensional, FIR filter, bandstop filter, Zolotarev polynomial, equiripple approximation

1 Introduction

Two-dimensional narrow bandstop FIR filters play important role in the image and video enhancement/restoration tasks. They are frequently used in order to remove a disturbing frequency components from the spectrum of the image. The image filtering can be accomplished by both the nonlinear [1], [4], [7], [8], [9], [10] and linear [5], [12], [13], [14], [15] filters. In our paper we are concerned with completely analytical design of 2-D bandstop FIR filters with extremely narrow circularly symmetrical stop bands. The design of the 2-D narrow bandstop FIR filters is based on the Zolotarev polynomial [14]. We introduce the degree formula which relates the degree of the generating polynomial, the length of the filter, the notch frequency, the width of the stopbands and the attenuation in the passbands. The design procedure is analytical one and it does not require any FFT algorithm or any iterative technique.
2 Polynomial Equiripple Approximation

We assume the impulse response \( h(m) \) with odd length \( N = 2n + 1 \) and even symmetry. The transfer function of the filter is

\[
H(z) = z^{-n} \left[ h(0) + \sum_{m=1}^{n} 2h(m) T_m(w) \right] = z^{-n} \left[ h(0) + \sum_{m=1}^{n} 2h(m) T_m(\cos \omega T) \right]
\]

(1)

where \( T_m(w) \) is Chebyshev polynomial of the first kind and \( w = (z + z^{-1})/2 \). The 1-D equiripple narrow bandstop FIR filter is based on the Zolotarev polynomial \( Z_{p,q}(w) \) [14] of the degree \( n = p + q \) which approximates constant value in equiripple Chebyshev sense in the two disjoint intervals as shown in Fig. 1. The notation \( Z_{p,q}(w) \) emphasizes that \( p \) counts the number of zeros right from the maximum \( w_m \) and \( q \) corresponds to the number of zeros left from the maximum \( w_m \). Zolotarev derived the general solution of this approximation problem in terms of Jacobi’s elliptic Eta function \( H(u|\kappa)) \) [14]

\[
Z_{p,q}(w) = Z(u|\kappa)) = \frac{(-1)^p}{2} \left[ \left( \frac{H(u - \frac{p}{n} K(\kappa))}{H(u + \frac{p}{n} K(\kappa))} \right)^n + \left( \frac{H(u + \frac{p}{n} K(\kappa))}{H(u - \frac{p}{n} K(\kappa))} \right)^n \right].
\]

(2)

The position of the maximum value \( y_m = Z_{p,q}(w_m) \) is

\[
w_m = 1 - 2 \text{sn}^2 \left( \frac{p}{n} K(\kappa)|\kappa) \right) + 2 \frac{\text{sn} \left( \frac{p}{n} K(\kappa)|\kappa) \text{cn} \left( \frac{p}{n} K(\kappa)|\kappa) \text{dn} \left( \frac{p}{n} K(\kappa)|\kappa) \right]}{\text{dn} \left( \frac{p}{n} K(\kappa)|\kappa) \right) Z \left( \frac{p}{n} K(\kappa)|\kappa) \right). \]

(3)

Fig. 1. Zolotarev polynomial \( Z_{6,9}(w) \) with \( \kappa = 0.6966 \)
The maximum value $y_m$ useful for the normalization of the Zolotarev polynomial is

$$y_m = \cosh 2n \left( \sigma_m Z\left( \frac{p}{n} \textbf{K}(\kappa) | \kappa \right) - \Pi(\sigma_m, \frac{p}{n} \textbf{K}(\kappa) | \kappa) \right).$$

The degree of the Zolotarev polynomial $Z_{p,q}(w)$ expresses the degree equation

$$n \geq \frac{\ln(y_m + \sqrt{y_m^2 - 1})}{2\sigma_m Z\left( \frac{p}{n} \textbf{K}(\kappa) | \kappa \right) - 2\Pi(\sigma_m, \frac{p}{n} \textbf{K}(\kappa) | \kappa)}.$$  \hspace{1cm} (5)

The auxiliary parameter $\sigma_m$ is given by the formula

$$\sigma_m = F \left( \arcsin \left( \frac{1}{\kappa \text{sn} \left( \frac{p}{n} \textbf{K}(\kappa) | \kappa \right)} \sqrt{\frac{w_m - w_s}{w_m + 1}} \right) \mid \kappa \right)$$

where $F(\Phi|\kappa)$ is the incomplete elliptical integral of the first kind. The recursive algorithm for the evaluation of the coefficients $a(m)$ of the Zolotarev polynomial based on the expansion into Chebyshev polynomials of the first kind

$$Z_{p,q}(w) = \sum_{m=0}^{n} a(m) T_m(w)$$

was derived and presented in [14]. The impulse response coefficients $h(m)$ of the 1-D equiripple bandstop FIR filter are obtained by the normalization of the coefficients $a(m)$ as follows

$$h(n) = \frac{y_m - a(0)}{y_m + 1}, \quad h(n \pm m) = -\frac{a(m)}{2(y_m + 1)} \quad \text{for} \quad m = 1 \ldots n.$$ \hspace{1cm} (8)

3 Design Procedure

The goal of the design procedure is to obtain the 2-D impulse response $h(m, n)$ of the filter satisfying the specified notch frequency $\omega_{m1} T$, width of the stopbands $\Delta \omega_1 T$, the attenuation in the passbands $a_1 \ [\text{dB}]$ in the direction $\omega_1$ and the specified values $\omega_{m2} T$, $\Delta \omega_2 T$, $a_2 \ [\text{dB}]$ in the direction $\omega_2$. The design procedure is as follows:

1. For the specified values $\omega_{m1} T$, $\Delta \omega_1 T$ and $a_1 \ [\text{dB}]$ (Fig. 1) in the direction $\omega_1$ design the 1-D FIR narrow bandpass filter. The design procedure consists of the following steps:
   (a) Evaluate the Jacobi’s elliptic modulus $\kappa$

   $$\kappa = \sqrt{1 - \frac{1}{\tan^2(\varphi_s) \tan^2(\varphi_p)}}$$

   for the auxiliary parameters $\varphi_s$ and $\varphi_p$

   $$\varphi_s = \frac{\omega_m + \omega_m \Delta \omega / 2}{2} T, \quad \varphi_p = \frac{\pi - (\omega_m - \Delta \omega / 2)}{2} T.$$

   \hspace{1cm} (10)
(b) Calculate the rational values $p/n$ and $q/n$

$$
\frac{p}{n} = \left[ \frac{F(\varphi_s|\kappa)}{K(\kappa)} \right], \quad \frac{q}{n} = \left[ \frac{F(\varphi_p|\kappa)}{K(\kappa)} \right].
$$

(11)

(c) Determine the required maximum value $y_m$

$$
y_m = \frac{2}{1 - 10^{0.05 a[\text{dB}]} - 1}.
$$

(12)

d) Using the degree equation \[15\] calculate and round up the minimum degree $n$ required to satisfy the filter specification.

e) Calculate the integer values $p$ and $q$ defining the Zolotarev polynomial

$$
p = \left[ n \frac{F(\varphi_s|\kappa)}{K(\kappa)} \right], \quad q = \left[ n \frac{F(\varphi_p|\kappa)}{K(\kappa)} \right].
$$

(13)

The brackets $[ \ ]$ in (13) stand for the rounding operation.

(f) For the integer values $p$, $q$ and the elliptic modulus $\kappa$ evaluate the coefficients $a(m)$ \[7\] of the Zolotarev polynomial $Z_{p,q}(w)$ using recursive algorithm \[14\].

(g) From the coefficients $a(m)$ calculate the $M$ coefficients of the impulse response $h_1(m)$ of the 1-D equiripple bandpass FIR filter using \[8\].

2. Repeat the first step for the design of the 1-D FIR equiripple narrow bandpass filter in the direction $\omega_2$ specified by $\omega_{m_2}T$, $\Delta \omega_2T$ and $a_2$ [dB] resulting in the impulse response $h_2(m)$ of the length $N$ coefficients.

3. From the 1-D impulse responses

$$
h_1(m), \quad m = 0, \ldots, M - 1 \quad \text{and} \quad h_2(n), \quad n = n, \ldots, N - 1
$$

(14)

compose the 2-D impulse responses $h_1(m, n)$ and $h_2(m, n)$ by the zero padding. The non-zero coefficients are

$$
h_1\left(\frac{M-1}{2}, n\right) = h_1(m), \quad m = 0, \ldots, M - 1
$$

$$
h_2\left(m, \frac{N-1}{2}\right) = h_2(n), \quad n = 0, \ldots, N - 1.
$$

(15)

4. The 2-D impulse response $h_{BP}(m, n)$ of the dimension $M \times N$ of the narrow bandpass FIR filter is given by the 2-D linear discrete convolution

$$
h_{BP}(m, n) = h_1(m, n) * * h_2(m, n).
$$

(16)

5. The impulse response $h(m, n)$ of the final 2-D bandstop FIR filter is

$$
h(m, n) = -h_{BP}(m, n) \quad \text{for} \quad m \neq \frac{M-1}{2}, \quad n \neq \frac{N-1}{2}
$$

$$
h\left(\frac{M-1}{2}, \frac{N-1}{2}\right) = 1 - h_{BP}\left(\frac{M-1}{2}, \frac{N-1}{2}\right).
$$

(17)
4 Example

Design the 2-D bandstop FIR filter specified in the direction $\omega_1$ by the notch frequency $\omega_{m1}T = 0.32\pi$, width of the stopbands $\Delta\omega_1T = 0.1\pi$ for the attenuation in the passbands $a_1 = -1$ dB and in the direction $\omega_2$ by the values $\omega_{m2}T = 0.42\pi$, $\Delta\omega_2T = 0.1\pi$ for $a_2 = -1$ dB.

Using our proposed design procedure we obtain the two 1-D equiripple narrow band FIR filters with the impulse responses $h_1(m)$, $h_2(n)$ (step 1 and 2 in Sec. 3). The impulse responses $h_1(m)$, $h_2(n)$ of the length $M = N = 45$ coefficients are summarized in Table 1. The impulse responses $h_1(m)$, $h_2(n)$ are used for the design of the 2-D bandstop FIR filter (step 3, 4 and 5 in Sec. 3). The impulse response $h(m, n)$ of the 2-D narrow bandstop FIR filter contains $45 \times 45$ coefficients. The amplitude frequency response $20\log |H(e^{j\omega_1}, e^{j\omega_2})|$ of the 2-D narrow bandstop FIR filter is shown in Fig. 2. This filter was successfully applied for the enhancement of the rastered newspaper picture. In fact, the above mentioned filter specification results from the spectrum (Fig. 3) of the input image. Both the input and processed images are shown in Fig. 4.

Table 1. Coefficients of the Impulse Responses

| m, n | $h_1(m)$  | $h_2(n)$  |
|------|-----------|-----------|
| 0    | -0.0384055 | -0.039357 |
| 1    | -0.0112026 | -0.005884 |
| 2    | 0.0070961  | 0.017115  |
| 3    | 0.0225536  | 0.018027  |
| 4    | 0.0180424  | -0.009881 |
| 5    | -0.0060652 | -0.027843 |
| 6    | -0.0286170 | -0.005190 |
| 7    | -0.0261042 | 0.029543  |
| 8    | 0.0030887  | 0.023765  |
| 9    | 0.0334627  | -0.019610 |
| 10   | 0.0346624  | -0.038524 |
| 11   | 0.0018352  | -0.000823 |
| 12   | -0.0363314 | 0.042145  |
| 13   | -0.0427889 | 0.025541  |
| 14   | -0.0035272 | -0.030842 |
| 15   | 0.0367080  | -0.045288 |
| 16   | 0.0494999  | 0.006712  |
| 17   | 0.0157867  | 0.051591  |
| 18   | -0.0344320 | 0.022453  |
| 19   | -0.0539265 | -0.040650 |
| 20   | -0.0232385 | -0.046058 |
| 21   | 0.0297412  | 0.015396  |
| 22   | 0.113975   | 0.115323  |
Fig. 2. Amplitude frequency response

Fig. 3. Spectrum of the input image

Fig. 4. Input and filtered image
Acknowledgement

This work was supported by the grant No. CEZ:J04/98:212300014, Ministry of Education, Czech Republic.

References

1. Astola, J., Kuosmanen, P.: Fundamentals of Nonlinear Digital Filtering. CRC Press, 1997.
2. Abramowitz, M., Stegun, I.: Handbook of Mathematical Function. Dover Publication, New York Inc., 1972.
3. Achieser, N. I.: Über einige Funktionen, die in gegebenen Intervallen am wenigsten von Null abweichen. Bull. de la Soc. Phys. Math. de Kazan, Vol. 3, pp. 1 - 69, 1928.
4. Fischer, V., Drutarovsky, M., Lukac, R.: Implementation of 3-D Adaptive LUM Smoother in Reconfigurable Hardware. Springer Verlag LNCS 2438, pp.720-729.
5. Gonzales, R.C., Woods, R.E.: Digital Image processing. Wiley Interscience, NY, 2001.
6. Lawden, D. F.: Elliptic Functions and Applications. Springer-Verlag, New York Inc., 1989.
7. Lukac, R.: Binary LUM Smoothing. IEEE Signal Processing Letters, Vol. 9, No. 12, December 2002, pp. 400-403.
8. Lukac, R.: Adaptive Vector Median Filtering. Pattern Recognition Letters, Vol. 24, No. 12, August 2003, pp. 1889-1899.
9. Lukac, R.: Simplified Boolean LUM Smoothers. Proceedings of the 4th EURASIP-IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications VIPromCom-2002, Zadar, Croatia, June 16-19, 2002, pp. 159-162.
10. Lukac, R.: The Way How to Design and Implement an Adaptive Method Based on Center-Weighted Medians. Proceedings of the IEEE Scientific Workshop Signal Processing 2002, Poznan, Poland, October 11, 2002, pp.9-14.
11. Pitas, I., Venetsanopoulos A.N.: Nonlinear Digital Filters : Priciples and Applications. Kluwer Academic Publishers, 1990.
12. Pratt, W. K., Venetsanopoulos A.N.: Digital Image processing. Kluwer Academic Publishers, 1990.
13. Vlček, M., Jireš, L.: Fast Design Algorithms for FIR Notch Filters. Proc. of IEEE International Symposium on Circuits and Systems ISCAS’94, London, Vol. 2, pp. 297 - 300, 1994.
14. Vlček, M., Unbehauen, R.: Zolotarev Polynomials and Optimal FIR Filters. IEEE Transactions on Signal Processing, Vol. 47, No. 3, pp. 717-730, March 1999.
15. Vlček, M., Zahradnik, P., Unbehauen, R.: Analytic Design of FIR Filters. IEEE Transactions on Signal Processing, Vol. 48, pp. 2705-2709, September 2000.
Fast Design of Optimal Comb FIR Filters

Pavel Zahradník and Miroslav Vlček

1Department of Telecommunications Engineering,
Czech Technical University, Prague
Technická 2, CZ-166 27 Praha, Czech Republic
Phone: +420-2-24352089, Fax: +420-2-33339810
2Department of Applied Mathematics,
Czech Technical University, Prague
Konviktská 20, CZ-110 00 Praha, Czech Republic
Phone: +420-2-24890720, Fax:+420-2-24890702
{zahradni,vlcek}@fel.cvut.cz

Abstract. A novel fast analytical design method for highly selective
digital optimal equiripple comb FIR filters is presented. The equiripple
comb FIR filters are optimal in the Chebyshev sense. The number of
notch bands, the width of the notch bands and the attenuation in the
passbands can be independently specified. The degree formula and the
differential equation for the generating polynomial of the filter is presen-
ted. Based on the differential equation, a fast simple algebraic recursive
procedure for the evaluation of the impulse response of the filter is intro-
duced. Its arithmetic robustness outperforms by far the known analytical
design method. Highly selective equiripple comb FIR filters with thou-
sands of coefficients can be designed. One example demonstrates the
efficiency of the filter design.

Keywords: FIR, comb filter, narrow band, optimal filter, equiripple
approximation, analytical design, recursive algorithm

1 Introduction

In many applications of digital signal processing the removal of harmonic inter-
f erences is desired while leaving the broad-band signal unchanged, e.g. in the
processing of EEG/ECG signals, powerline communication etc. These tasks can
be accomplished by comb filters. In the design of comb filters, the width of the
notch bands is important. Comb IIR filters exhibit very narrow notchbands, but
their phase response is nonlinear. They produce substantial distortions of the
output signal which appear in the vicinity of the passbands due to the group
delay variation [7]. In numerous applications, especially in the processing of
pulse-like signals, the linear phase of FIR filters is essential. The narrowest pos-
sible notch bands are exhibited by comb FIR filters optimal in the Chebyshev
sense with equiripple behavior of the frequency response in the passbands. The
known analytical design technique of the equiripple comb FIR filter is based on
the optimal equiripple lowpass/highpass prototype FIR filter [2], [6]. The generating polynomial of the lowpass/highpass prototype FIR filter is expanded into the sum of Chebyshev polynomials to get the impulse response of the prototype filter. Finally, each delay of the lowpass/highpass prototype FIR filter is replaced by multiple delays to get the impulse response of the equiripple comb FIR filter [3], [4], [5]. However, the dynamic range of the coefficients of the generating polynomial of the prototype filter may be enormous, and consequently the conversion of the generating polynomial into the sum of Chebyshev polynomials requires very high precision arithmetics. This expansion represents the weak point of the known design procedure. In this paper, our novel simple algebraic recursive procedure based on the generating polynomial of the optimal equiripple comb FIR filter avoids the above mentioned weak point of the known approach. The simple algebraic recursive procedure removes the expansion of the generating polynomial into the sum of Chebyshev polynomials and thus it allows the design of highly selective FIR comb filters with hundreds and thousands of coefficients.

2 Generating Polynomial and Differential Equation

Due to the symmetry of the impulse response coefficients the transfer function of a linear phase FIR filter of the length \( N = 2M + 1 \) can be written in the form

\[
H(z) = \sum_{m=0}^{2M} h(m) z^{-m} = z^{-M} \left[ h(M) + 2 \sum_{m=1}^{M} h(M \pm m) \frac{1}{2} (z^m + z^{-m}) \right] = z^{-M} \sum_{m=0}^{M} a(m) T_m(w) = z^{-M} Q(w)
\]

where

\[
a(0) = h(M) \quad a(m) = 2h(M \pm m)
\]

and \( T_m(w) \) is Chebyshev polynomial of the first kind. The function \( Q(w) \) represents a polynomial in the variable \( w = \frac{1}{2}(z + z^{-1}) \) which on the unit circle reduces to a real valued zero phase transfer function \( Q[(\cos(\omega T))] \) of the corresponding FIR filter. The generating polynomial \( F(w) \) of the comb FIR filter is given by the compounded Chebyshev polynomial

\[
F(w) = T_n [\lambda T_r(w)] = \sum_{m=0}^{nr} B(m) w^m = \sum_{m=0}^{nr} A(m) T_m(w)
\]

For illustration, the generating polynomial \( F(w) = T_6 [1.15 T_5(w)] \) is shown in Fig. 1. The real parameter \( \lambda = 1/\kappa > 1 \) affects the amplitude of the ripples in the passbands of the FIR comb filter. The degree \( r \) of the inner Chebyshev polynomial determines \( r \) narrow bands. The positions of the narrow bands are identical with the positions of the extremal values of the inner polynomial \( T_r(w) \), which are \( w_{mi} = \cos(i \pi/r), \ i = 0, 1, ..., r \). In the frequency domain the narrow
bands positioned at $\omega_{mi} T = (i \pi)/r$, $i = 0, 1, \ldots, r$ are equally spaced inside the interval $[0, \pi]$. Because the marginal narrow bands at $\omega T = 0$ and $\omega T = \pi$ exhibit the half width we assume them as one narrow band. For illustration the comb filter shown in Fig. 3 based on the polynomial $T_6[1.15 T_5(w)]$ exhibits five narrow bands. The even degree $n$ of the outer Chebyshev polynomial $T_n(w)$ determines $n - 1$ local extremes of the same amplitude between the narrow bands. We have derived the differential equation for the generating polynomial in the form

$$U_{r-1}(w) (\kappa^2 - T_r^2(w)) \left[ (1 - w^2) \frac{d^2 F(w)}{d w^2} - w \frac{d F(w)}{d w} \right]$$

$$-r(1 - \kappa^2) T_r(w) \frac{d F(w)}{d w} + n^2 r^2 U_{r-1}(w) (1 - T_r^2(w)) F(w) = 0 \ . \hspace{1cm} (4)$$

The differential equation is indispensable for the derivation of the algebraic recursive algorithm for the evaluation of the coefficients of the impulse response $h(m)$ of the filter.

---

3 Optimality of the Equiripple Comb FIR Filter

The Chebyshev polynomial $T_n(w)/2^{n-1}$ represents an optimal equiripple approximation of the zero value on the interval $[-1, 1]$. The polynomial $T_n(w)$ exhibits $n - 1$ alternating extremes with the value $\pm 1$ inside the interval $[-1, 1]$, i.e. $n + 1$ extremal values including the interval edges $w = \pm 1$. The compounded polynomial $T_n[T_r(w)]$ exhibits $nr + 1$ alternating extremes with the value $\pm 1$ including the interval edges $w = \pm 1$ as $T_n[T_r(w)] = T_{nr}(w)$. Consequently the normalized generating polynomial $F(w)/(\lambda^n 2^{nr-1}) = T_n[\lambda T_r(w)]/(\lambda^n 2^{nr-1})$ represents the optimal equiripple approximation of the zero value on $r$ equally spaced disjoint intervals of the same width with respect to the equally spaced extremes of the same value between the intervals.
4 Degree Equation

The zero phase transfer function $Q(w)$ of the comb FIR filter is given by the normalization of the generating polynomial

$$Q(w) = 1 - \frac{1 + T_n[\lambda T_r(w)]}{C} = \sum_{m=0}^{nr} b(m) w^m = \sum_{m=0}^{nr} a(m) T_m(w) .$$

(5)

The normalizing constant $C$ follows from the condition $Q(w)|_{w=1} = 0$

$$C = 1 + \cosh[n \cosh(\lambda)] = 1 + T_n(\lambda) .$$

(6)

We emphasize the independence of (6) from the degree $r$ of the inner Chebyshev polynomial $T_r(w)$. The degree $n$ of the outer Chebyshev polynomial $T_r(w)$ is

$$n = \frac{\cosh(k)}{\cosh(\lambda)} = \frac{\ln(k + \sqrt{k^2 - 1})}{\ln(\lambda + \sqrt{\lambda^2 - 1})}$$

(7)

where parameters $\lambda$ and $k$ are

$$\lambda = \frac{1}{\cos\left(r \frac{\Delta \omega T}{2}\right)} , \quad k = \frac{1 + 10^{0.05a[dB]}}{1 - 10^{0.05a[dB]}} .$$

(8)

We call (7) the degree equation of the equiripple comb FIR filter. For illustration, the zero phase transfer function

$$Q(w) = 1 - \frac{1 + T_6[1.15 T_5(w)]}{1 + T_6(1.15)}$$

(9)

is shown in Fig. 2. The corresponding frequency response $20 \log |H(e^{j\omega})| \text{ [dB]}$ is shown in Fig. 3. Note that there are true zeros at the notch frequencies. In the filter design the real value $n$ (7) has to be up-rounded to the next even integer value. We emphasize that the up-rounding of $n$ preserves the specified number of notch bands and the width of the notch bands. The only affected parameter is the attenuation in passbands $a$ [dB] which is after the up-rounding equal or less than the specified value. The impulse response $h(m)$ of the filter consists of $2nr + 1$ coefficients, among which are $n + 1$ non-zero values.

5 Evaluation of the Impulse Response

Based on differential equation (4) we have developed a simple algebraic recursive algorithm for the algebraic evaluation of the impulse response $h(m)$ of the filter. This algorithm does not require high precision arithmetics beyond the precision of Matlab. It avoids a numerically fragile expansion of the generating function (5). The evaluation of the impulse response is summarized in Tab. 2. The design procedure consists of the following steps:
1. Specify the number of notch bands $r$, the width of the notch bands $\Delta \omega T$ and the maximal attenuation in the passbands $a$ [dB], as demonstrated in Fig. 3.
2. The degree of the inner Chebyshev polynomial is $r$.
3. Determine the auxiliary parameters $\lambda$ and $k$.
4. Evaluate $n$, round up $n$ to the next even integer value.
5. Evaluate the $2nr + 1$ coefficients of the impulse response $h(m)$ of the comb FIR filter using the algebraic recursive algorithm (Tab. 2).
6. Evaluate the actual attenuation in the passbands
\[
a_{act} [dB] = 20 \log \left( 1 - \frac{2}{1 + T_n(\lambda)} \right). \tag{10}
\]

6 Example

*Design an equiripple comb FIR filter with 20 notch bands. The specified width of the notch bands is $\Delta \omega T = \pi/50$ and the maximal attenuation in the passbands is $a = -1$ dB.*

| $m$ | $h(m)$  |
|-----|---------|
| 0   | 240     |
| 40  | 200     |
| 80  | 160     |
| 120 | 0.749920|

The degree of the inner Chebyshev polynomial is $r = 20$. The auxiliary parameters are $\lambda = 1.2361$ and $k = 17.3910$. From the degree equation (7) we obtain $n = 5.2623$ which is up-rounded to $n = 6$. The zero phase transfer function $Q(w)$ is
Table 2. Recursive algorithm for evaluation of the coefficients $h(m)$ of the impulse response

| Step | Equation |
|------|----------|
| **given** | $n$ (even integer), $r$ (integer), $\lambda > 1$ (real) |
| **initialization** | $\kappa = \frac{1}{\lambda}$  
$\alpha(n) = \lambda^n$  
$\alpha(n + 2) = \alpha(n + 4) = \alpha(n + 6) = 0$ |
| **body** | For $\mu = 1 \ldots \frac{n}{2}$  
$\alpha(n - 2\mu) = \left\{ \begin{array}{l} \alpha(n - 2(\mu - 1)) \left[ (1 - \kappa^2)(n - (2\mu - 1)) \right. \\
\times (n - (2\mu - 2)) + 3(\mu - 1)(n - (\mu - 1)) \\
- \alpha(n - 2(\mu - 2)) \left[ (1 - \kappa^2)(n - (2\mu - 4)) \right. \\
\times (n - (2\mu - 5)) + 3(\mu - 2)(n - (\mu - 2)) \\
+ \alpha(n - 2(\mu - 3))(\mu - 3)(n - (\mu - 3)) \end{array} \right\} / \mu(n - \mu)$ |
| **end loop on $\mu$** | $\alpha(0) = \frac{\alpha(0)}{2}$ |
| **coefficients $A(m)$ of the generating polynomial $F(w)$** | For $\mu = 0 \ldots \frac{n}{2}$  
$A(nr - 2\mu) = \alpha(n - 2\mu)$ |
| **end loop on $\mu$** | |
| **coefficients $a(m)$ of the zero phase transfer function $Q(w)$** | $C = 1 + T_n(\lambda)$  
$A(0) = 1 - \frac{1 + A(0)}{C}$ |
| **end loop on $\mu$** | $a(\mu) = -\frac{A(\mu)}{C}$ |
| **coefficients of the impulse response $h(m)$** | For $\mu = 1 \ldots nr$  
$h(nr) = a(0)$ |
| **end loop on $\mu$** | $h(nr \pm \mu) = \frac{a(\mu)}{2}$ |
\[ Q(w) = 1 - \frac{1 + T_6 \left[ 1.2361 T_{20}(w) \right]}{1 + T_6 \left( 1.2361 \right)}. \]  

(11)

For the values \( n, r \) and \( \lambda \) the impulse response \( h(m) \) is evaluated using the algebraic recursive procedure (Tab. 2). The impulse response \( h(m) \) with a length of 241 coefficients consists of seven non-zero coefficients. They are summarized in Tab. 1. The actual parameters of the comb FIR filter are as specified, except for the maximal attenuation in the passbands, which is \( a_{\text{act}} = -0.6080 \) dB due to the up-rounding of the degree \( n \). The amplitude frequency response \( 20\log|H(e^{j\omega})| \) [dB] of the equiripple comb FIR filter is shown in Fig. 4.

Acknowledgement

This work was supported by the grant No. CEZ:J04/98:212300014, Ministry of Education, Czech Republic.

References

1. Antoniou A.: Digital Filters, Analysis, Design, and Applications, McGraw-Hill Book Co., New York, 1993.
2. Herrman, O., Rabiner, L. R., Chan, D. S. K.: Practical Design Rules for Optimum Finite Impulse Response Lowpass Digital Filters, The Bell System Technical Journal, Vol. 52, No. 6, 1973, pp. 769-799.
3. Mitra, S. K.: Digital Signal Processing A Computer-Based Approach, McGraw-Hill, 1998.
4. Sophocles, J. Orfanidis : Introduction to Signal Processing, Prentice-Hall Inc., Upper Saddle River New Jersey, 1996.
5. Proakis, J. G., Manolakis, D. G.: Digital Signal Processing Principles, Algorithms and Applications, Prentice-Hall Inc. Upper Saddle River New Jersey, 1996.
6. Saramäki, T.: Finite Impulse Response Filter Design, in Handbook for Digital Signal Processing, eds. S. K. Mitra and J. F. Kaiser, New York: Wiley, 1993, ch. 4, pp. 195-198.
7. Vlček, M., Zahradník, P.: Digital Multiple Notch Filters Performance, Proceedings of the 15th European Conference on Circuit Theory and Design ECCTD’01. Helsinki, August 2001, pp. 49-52.
8. Zahradník, P., Vlček, M.: Digital FIR Double Notch Filters, Proceedings of the IASTED International Conference on Signal and Image Processing. Honolulu, August 2001, pp. 369-372.
9. Zahradník, P., Vlček, M.: Equiripple FIR Triple Narrow Band Filters, Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems APCCAS’02. Bali, October 2002. Paper No. 218, 4 pages.
10. Zahradník, P., Vlček, M.: Analytical Design of Optimal FIR Comb Filters, Proceedings of the IEEE 2003 International Conference on Communications ICC’03. Anchorage, Alaska, May 2003. Paper No. GC31-2, 4 pages.
11. Zahradník, P., Vlček, M.: Fast Analytical Design Algorithms for FIR Notch Filters, IEEE Transactions on Circuits and Systems I : Fundamental Theory and Applications, in print.
Artificial Intelligence Methods in Diagnostics of the Pathological Speech Signals

Andrzej Izworski, Ryszard Tadeusiewicz, and Wieslaw Wszolek

AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
{izwa, rtad}@agh.edu.pl, wwszolek@uci.agh.edu.pl

Abstract. In the work excerpts of research are presented, concerning the application of modified acoustic signal processing methods in the problem of “understanding” of selected pathologies of vocal tract. The concept of the research scheme is based on the technique of advanced acoustic signal analysis and it refers to the analysis of artificial neural networks functioning in the task of recognition of selected types of vocal tract pathologies. The method is based on utilization of an internal model of the considered signal’s generator and it is directed towards such a structure analysis of the examined sound. The described method allows to achieve more subtle differentiation for signal characterized by small diversification of measurable features, observed for the classes being recognized, what is the case in the problem of identification of selected pathologies considered here.

1 Introduction

In many problems of medical diagnosis, as well as planning and monitoring of therapy and rehabilitation of speech related organs, it is necessary to evaluate qualitative features of the acoustic signal of deformed speech. Tasks related to analysis and recognition of pathological acoustic signal of speech, characterizing selected pathological states, are exceptionally difficult. The difficulty results from the fact that forms of speech organ pathology, which are to be recognized (or classified) manifest themselves in various forms of speech signal deformation, often hard to predict and very inconvenient to be revealed in real, recorded speech signal of a given patient being examined. The correlation between phonetic and acoustic phenomena, observed in the temporal or spectral representation of speech signals, in general poorly correlates with morphological or pathophysiological features of the deformed speech generator. It happens that minor pathological elements (e.g. occlusion defect) strongly manifest in the speech signal, while very serious pathological changes (e.g. tumor) give only a weak and hardly readable picture of speech disturbances. Therefore it is very difficult to diagnose the condition and pathological changes of the voice tract using speech signal [1], in spite of existence of multiple examples of successful automated speech recognition in the semantic (recognition of the utterance contents for e.g. voice control of machines and devices) or personal aspect (verification and identification of persons by using their speech samples). Neither is
there a simple way to transfer the experience related to diagnosis of technological system, because the problems of pathological speech diagnosis are specific by the fact that for such tasks it is very difficult to find an appropriate rule for the preliminary signal analysis. What's more, it is also difficult and sometimes even impossible to indicate a proper recognition algorithm for the pathological speech signal [2]. It follows from the fact that during the identification of voice tract pathological states based on the analysis of generated deformed speech it is necessary to resort to highly specialized (atypical) methods, both for the signal parameterization as well as its categorization and classification. On the basis of the statement, that for the cases of analysis of speech pathology forms and sources discussed here the well-known methods of automated signal recognition cannot be applied, the authors propose in the present paper a new approach, based on the concept of automated understanding. Because of possible multiple meanings of that phrase it should be stressed that the meaning used in the presented work concerns the automated understanding of the nature and character of the pathological speech signal deformations. The exact meaning of the term understanding has no connections with the frequently discussed problem of semantic understanding recognition of the speech signal i.e. the contents of the pronounced sentences.

![Fig. 1. Spectra of the “A” vowel, correct and deformed (after chordectomy, hemilaryngectomy and fronto-lateral laryngectomy)](image)

Such a new approach is necessary because several previous attempts, made by the authors of the present work, directed towards the construction of a system recognizing types of speech organs pathologies, in spite of unquestionable successes, did not lead to final solutions. The reason seems to be the fact, that every attempt of a simple recognition of speech pathology must be based on the evaluation of some measure of difference between the specific utterance of a given patient and some standard of the correct speech. Alas such a simple recognition concept, superbly working in recognition of the utterance contents, or in verification of the speaker, does not meet the expectations in the attempts of recognition and classification of forms of speech.
pathology. The reason lies in the great changeability and diversity of speech. It concerns both regular and pathological speech [3]. Every person speaks in somewhat different way, various (with respect to the contents or speed) utterances of the same person reveal various phonetic and acoustic features of his/her speech signal, and even various registrations of the same utterance recorded from the same person but e.g. in various days, can be very different. Particularly troublesome is the considerable diversity of correct speech, as it is rather difficult to refer to (in the measurement of degree of speech pathology) a set of speech samples, in which the temporal, spectral and parametric features exhibit huge dispersion of values. It is almost a rule, that various samples of correct speech signal exhibit a greater variety of measurable acoustic parameters, that the measurable differences of the same parameter between these samples and the registered samples of speech, which is obviously pathological (see Fig. 1).

All this is the reason that one cannot confine to models of pathological speech signal recognition in a space based on the set of its features, but in every case one should try to understand, how did such a phonetic or acoustic phenomenon occur. It means, that the diagnostic system must contain an internal model of the signal generator, based on the knowledge about the speech signal and the ways of its generation - in regular and pathological conditions.

2 The Material of the Study

The studies of the speech articulation have been carried out for persons treated for the larynx cancer (men after various types of operations). Depending on the stage of the tumour, various types of partial larynx surgery have been applied. The final acoustic material has been collected from 95 persons divided into two groups:

- the reference group (the standard group), 25 persons with a correct pronunciation
- the group of patients (75 persons) treated by the following surgery methods
  - hemilaryngectomy (28 persons)
  - chordectomy (17 persons), enlargata (6 persons)
  - laryngectomy subtotalis (14 persons)
  - laryngectomy fronto-lateralis (5 persons)

Both the patients and the persons from the reference group pronounced the same text (three times), which consisted of: vowels (A,U,E,I), words containing vowels. The selection of phrases and sets of words pronounced by the examined persons has been based on morphological and functional analysis of the expected (for a given pathology) distinctions of speech organs, what resulted in collection of research material including sets of words selected with respect to their phonetic features in order to carry the maximum amount of information.

In order to receive undisturbed results, ensuring a precise and sometimes even very subtle evaluation of the quality and usefulness of specific sets of input parameters, it was necessary to collect signal samples of very high quality. After preliminary processing of the registered signal the result is a multispectrum, digitized in time, frequency and amplitude by the acoustic analyzer.
3 The Concept of Research

The research task of the present work is the evaluation of origins of speech signal deformations after larynx surgery treatment. One of the important problems encountered during the elaboration of the collected samples was the reduction of the very large information file, the source of which was the analyzed acoustic speech signal (e.g. in the form of dynamic spectra), to the space of features with reduced number of dimensions but information contents sufficient and useful from the diagnostic point of view. In the further signal processing stage the dynamic spectra has been transformed to several variants of feature vectors.

The above mentioned features have been selected during the long-time studies concerning the evaluation of the speech deformation level and the search for features combining the following three advantages:

- are insensitive to the content of the statement and personal features of the speaker’s voice
- exhibit great sensitivity for distinguishing between various forms of the same type of pathology and in classification of various stages of development for a given pathology
- are easy to determine from the registered speech signal samples and exhibit the required numerical stability (are insensitive to small errors in the signal measurement)

The authors have selected and studies several feature vectors, for which the respective spaces could be satisfactorily metricized, and which are presented below:

\[
< f_1, f_2, \ldots, f_{96} > = X_1
\]

where: \( f_i \)- the averaged level values in the \( i \)-th frequency band, with \( \Delta f = 125 \text{Hz} \)

\[
<F_1, F_2, F_3, M_0, M_1, M_3> = X_2
\]

where: \( F_1,F_2,F_3 \)- formants, \( M_0,M_1,M_3 \)- spectral moments

\[
< M_0, M_1, M_3, C_w, C_p, J, S > = X_3
\]

where:

\( C_w, C_p \)- the relative power coefficients,
\( J \)- Jitter (denotes the deviation of the larynx tone frequency in consecutive cycles from the average frequency of the larynx tone)
\( S \)- Shimmer, (denotes the deviation of the larynx tone amplitude in the consecutive cycles from the average amplitude of the larynx tone)

The concept described in the introduction has been presented in Fig. 2.

The model of signal generation represents all the knowledge about the pathological speech signal. The products of the model are spectra of the signal. The actual signal of pathological speech (obtained from a particular patient) after its transformation into the vector of features is compared with a transformed output signal of the model.
4 The Model of Speech Organs Simulation

The complex process of acoustic speech signal generation can be presented in the form of a theoretical model mapping functions performed by particular organs. It is essential for the simulation model to enable the determination of the signal spectrum, based on the geometrical parameters of the vocal tract specific for the articulation of particular speech sounds. The basis for presentation of the model has been taken from the works [7,8]. In the simulation model three principal modules have been distinguished:

- the source of the acoustic wave $G$, characterized by impedance $Z_g$,
- four-terminal network, characterized by transmittance $K(j\omega)$,
- load impedance $Z_{lo}$

which are presented in Fig. 3.

In the present work a model of larynx generator has been assumed, considered as a source of signals of frequencies $F_0$, $2F_0$, $3F_0$ etc., the schematic diagram of which is presented in Fig. 4.
The introduced notation is as follows: $F_{sou}$ reflects a simplified envelope of the spectral characteristic $|A_g(j\omega)|$.

$$F_{sou}(f) = \frac{1}{2} \left( \frac{f}{F_0} \right)^2$$

(4)

while the resistance $R_{agav}$ and the source’s acoustic mass $L_{agav}$ are taken for respective of these elements for average value of the glottis section $A_{gav}$.

5 Results

On the basis of general recommendation it has been decided to use in the study three layer networks of the feedforward organization type, built of elements with sigmoidal characteristics of the nonlinear transfer functions, connected according to the rule of full network connection between the input layer and the next layer of the network.

The criterion for the learning process termination has been connected with the changes (monitored during the learning process) of a purposely constructed coefficient:

$$\text{DELTA} = 1 – \text{pos}(Y) – \text{neg}(Y)$$

(5)

where:

$$\text{pos}(Y) = \max_{1 \leq i \leq n} (1 - z_i) y \quad \text{neg}(Y) = \max_{1 \leq i \leq n} (1 - z_i) y$$

This coefficient is a measure of domination of the recognition accepted by the network over the other competing recognitions. The description of application methodology for this original criterion can be found in the paper [4].

The product of such comparison and evaluation is a signal used for modification of internal model parameters, in order to minimize the difference between the vectors of features of the actual pathological speech signal and the signal generated by the model. The size and direction of the model modification is a measure of the speech signal deformation degree. In Fig. 5 the spectrum of the I vowel speech signal has been presented for the actual utterance.

\[ \text{Fig. 5. Spectrum of utterance of vowel I - correct pronunciation (left) and pathological pronunciation (right)} \]
The signal obtained from the model has been presented in Fig. 6.

The introduced concept of signal understanding consists of introduction of quantitative factors, describing the essence of the origins of signal deformation (e.g., various pathologies of the vocal tract). The speech signal recorded for a particular patient and the signal created by the generation model (in the form of the spectrum) are processed to the form of vectors of features and then compared (using the artificial neural networks) with respect to their similarity.

![Simulated spectrum of I vowel – correct (left) and pathological (right)](image)

The result of the evaluation is used for elaboration of such correction of the respective model parameters, which result in the greatest similarity of both signals. The magnitude of changes of the selected model parameters is a measure of the signal deformation, and the information specifying which of the model parameters induced the signal change ensuring the greatest similarity determines the level of “understanding” of the deformation origins.

6 Conclusions

Multiple psychological and neurophysiological experiments convince that people acquire information not by simple analysis and processing of sensory stimuli, but mainly by the way of specific cognitive resonance. For every essential perception category there must be a previous so called gnostic field, existing in the brain, usually elaborated by the learning process - although the inborn (genetically determined) origin of some of them cannot be also excluded. The perception always consists in confrontation of the present stream of sensory sensations received by the receptors with a stream of expected values of these impressions, predicted by action of internal models of perception categories localized exactly in the gnostic fields. Because of the multitude of learned perception categories, the wave of excitations coming from the receptors is actually confronted with a whole bunch of waves produced by various gnostic units, related to various perception categories. As a result of a specific interference the stream of data from the receptors is enhanced by the stream of information from a specific (for a given category) gnostic unit, what is sometimes called a cognitive resonance. Similar interference with other gnostic units leads to mutual wave suppression, what leads to revealing and enhancing of the proper stimulus recognition with accompanying efficient elimination of alternative interpretations (perception hypotheses). The perception concept sketched here is widely known, but the original contribution from the authors was its application as a standard for a construction of the system of automated understanding of selected
voice system pathologies, treated as objects which are being diagnosed by the analysis of the acoustic speech signal, deformed by these pathologies.

With such a solution, essentially different from solutions employed in the presently constructed diagnostic systems, the speech signal, received from a given patient, is recorded and analyzed (usually by use of neural networks [4]), and then it is confronted (most often in the plane of properly selected temporal and spectral multispectral features) with the reference signals, formed by the internal models mentioned above (the generators of pathological speech specific for known pathology forms). The process of adjustment of parameters of the registered signal, obtained from a given patient, and the signals obtained from the internal generators, leads first to the selection of this generator for which the strongest cognitive resonance is observed, and in the next stage of the perception modeling to the process of adjustment of internal generator parameters, executed for tuning it to the parameters of the observed signal, which leads to formulation of more exact (final) diagnostic hypothesis.

The described concept includes a series of elements unquestionably difficult in practical realization. For the traditional way of solving diagnostic problems the answer is frequently found more easily. However longtime experience of the authors in the problems related to analysis, evaluation and classification of pathological speech signals have proved, that for this task really a new approach is required. It is because for recognition of pathological speech the standard signals processing and classification methods, used in semantic speech recognition (comprehension of the utterance contents) or voice recognition (speaker identification), totally fail [3]. These methods include the spectral analysis (sometimes executed with application of the wavelet transformation technique, popular recently), discrimination analysis, linear prediction coefficients or cepstral coefficients. They cannot satisfactorily describe the pathological speech, because of its phonetic and acoustic structure, dissimilar with respect to the correct speech signal, and also because the goal of the recognition process is totally different for that case [4]. At the same time the amount of needs related to technical assistance in diagnostic, prognostic check up tasks, performed by the physicians dealing with speech pathology, constantly grows. Successful attempts of construction of picture recognition systems [5], [6] indicate that the proposed way may be effective.

In conclusion it can be stated, that in the field of automated diagnosis of pathological speech it is necessary to construct special methods of automated understanding of the nature of processes leading to speech deformation, which could replace the presently employed methods of typical acoustic signal analysis and recognition, and which would be fully adapted to the specificity of the considered problem. Such an attempt of construction of a new, fully original, special method of understanding is the concept described in the present work. It refers to solving of the problems considered here, making use of proper representation of the knowledge regarding the studied articulation process and the consequences of its deformation (in the form of adaptively adjusted models). The studies of the new method have just started, and it cannot be told whether this technique will be able to solve all the problems and to overcome all the difficulties. In general it is known, that in the tasks of acoustic signal (picture) analysis and recognition the unification of methods and standarization of algorithms has always encountered serious problems. The main
source of those difficulties is the fact, that in almost every task of signal analysis, different features and different parameters, closely related to the specificity of task being solved, have to be extracted, and they are used for finding answers to different questions. Similarly in the tasks of acoustic signals recognition the criteria and goals of their classification can be very different - even for the same signals. Because of that the proposed method will have to be considerably modified, in application to various specific tasks. The adaptation will affect both the methods of preliminary processing of acoustic signals, which obviously have to be oriented for the specific features of every identification task considered, and the techniques of internal modeling of the generation processes for various forms of speech pathologies. Also the techniques of appointing of the optimal model have to be specific, because as mentioned before, the tasks of pathological speech analysis are special by the fact that no shape of standard signal, to which a reference or relation could be made, can be found.

References

1. Tadeusiewicz R., Wszołek W., Wszołek T, Izworski A.: Methods of Artificial Intelligence for Signal Parameterisation Used in the Diagnosis of Technical and Biological Systems, 4th World Multiconference on Systemics, Cybernetics and Informatics, July 23-26, (2000) Orlando, FL, USA.
2. Tadeusiewicz R., Wszolek W., Izworski A., Wszolek T.; Methods of deformed speech analysis. Proceedings, International Workshop Models and Analysis of vocal Emissions for Biomedical Applications, Florence, Italy, 1-3 September (1999), 132-139.
3. Tadeusiewicz R., Izworski A., Wszołek W.: Pathological Speech Evaluation Using the Artificial Intelligence Methods, Proceedings of “World Congress on Medical Physics and Biomedical Engineering”, September 14-19, (1997), Nice, France.
4. Tadeusiewicz R., Wszołek W., Izworski A., Application of Neural Networks in Diagnosis of Pathological Speech, Proceedings of NC’98, “International ICSC/IFAC Symposium on Neural Computation”, Vienna, Austria, (1998), September 23-25.
5. Leś Z., Tadeusiewicz R.: Shape Understanding System - Generating Exemplars Of The Polygon Class, in Hamza M.H., Sarfraz E. (eds.): Computer Graphics and Imaging, IASTED/ACTA Press, Anaheim, Calgary, Zurich, (2000), 139-144.
6. Ogiela M. R., Tadeusiewicz R.: Automatic Understanding of Selected Diseases on The Base of Structural Analysis of Medical Images, Proceedings of ICASSP 2000, Salt Lake City, (2001).
7. Flanagan J.L.: Speech analysis, synthesis and perception. Springer-Verlag, Berlin-Heidelberg-New York, (1965).
8. Kacprowski J.: An acoustic model of the vocal tract for the diagnostic of cleft palate. Speech analysis end synthesis (ed. by W.Jassem), vol.5, Warsaw, (1981), 165-183.

This paper was supported by AGH grant 10.10.130.941
Intelligent Sub-patch Texture Synthesis Algorithm for Smart Camera

Jhing-Fa Wang, Han-Jen Hsu, and Hong-Ming Wang

Multimedia & Communication IC Design Lab
Department of Electrical Engineering, National Cheng Kung University
92A79, EE, No.1, Ta-Hsueh Road, Tainan, Taiwan
wangjf@csie.ncku.edu.tw
{hjhsu, homeming}@icwang.ee.ncku.edu.tw
http://icwang.ee.ncku.edu.tw

Abstract. We propose an intelligent sub-patch texture synthesis algorithm for smart camera. Redundant human removal is a novel function in the next generation smart camera. In order to fill the lost region, texture synthesis is an important research topic. We present an efficient constrained texture synthesis algorithm which can past a patch each time and used in human removal. The proposed intelligent algorithm is suitable to be integrated in smart camera and used in image reconstruction. The searching neighborhood pixels can be changed due to different patch size. The intelligent algorithm consists of three steps: line-expanding, texture synthesis and block-smoothing. The boundary artifact can be removed by the block-smoothing method. The comparison of previous algorithms is also provided in this paper. We present a first algorithm used in constrained texture synthesis by regular synthesizing order.

1 Introduction

Smart camera can remove redundant human when people capture the scene in real time. Constrained texture synthesis can generate similar texture by matching the boundary texture and source texture in the input image. Many texture synthesis algorithms have been presented in previous research. The classification of the algorithms consists of pixel-based, patch-based and feature matching. Pixel-based algorithm synthesize a pixel each time and need much computation time. Li-Yi Wei proposed a texture synthesis by fixed-neighborhood searching [2]. The algorithm can be accelerated by tree-structured vector quantization. About patch-based algorithm, Efros proposed a patch-based by image quilting [3]. The overlap part of patch boundary can be decided by minimal error cut calculation. Liang presented a real-time texture synthesis by patch-based sampling [4]. The efficiency of patch-based algorithm is better than pixel-based one. Yu-Hen Hu proposed a block-based method based on the method of Efros which can be used in constrained texture synthesis for image post processing [5]. Some algorithms generate new images by matching the feature in the example texture. Simoncelli generate textures by matching the joint statistics of the image pyramids [6].
In wireless image transmission, the image bitstream maybe lost by fading channels. We can use texture synthesis to reconstruct the damaged image. Rane and Bertalmío significantly presented a texture synthesis and image inpainting algorithms in wireless image reconstruction [7][8]. The searching shape is also block-based in the part of texture synthesis.

In order to decrease the redundant neighborhood pixels in small lost region, we present a scalable sub-patch algorithm. The intelligent algorithm can change the neighborhood pixels and patch size. In each time, we past one pixel or one line each time. In constrained texture synthesis used in structural texture, the algorithm can synthesize texture without boundary artifact.

2 Intelligent Sub-patch Texture Synthesis Algorithm

This section will introduce the proposed intelligent sub-patch texture synthesis algorithm. At first, the algorithm extends the segmentation boundary by several pixels horizontally. The reconstructed result is affected by different pixel numbers. At second, the searching neighborhood size can be changed intelligently in different patch size. Finally, the block-smoothing method is described in Section 2.3. In order to maintain the advantages of different texture synthesis algorithms, we use an inverse-U shape to capture the probability model of texture. The proposed algorithm is faster than pixel-based algorithm because of pasting a patch each time. The procedure of the algorithm used in hole-filling is shown in Fig. 1. For example, we want to synthesize the hole in the image as the original texture. The gray part of Fig. 1 is $T_{\text{hole}}$. The dark gray part is the current patch $P_{\text{cur}}$. We take a sample of input texture $T$ to find the best matching neighborhood around of $T_{\text{hole}}$. The neighborhood shape is inverse-U to

| parameter | Meaning                        |
|-----------|--------------------------------|
| $T_{in}$  | 2D input texture image         |
| $T$       | 2D input texture image without hole |
| $T_{\text{hole}}$ | Hole of the input texture image |
| $P_{\text{cur}}$ | Current patch |
| $P_{\text{can}}$ | Candidate patch |
| $N(P_{\text{cur}})$ | Neighborhood of current patch |
| $N(P_{\text{can}})$ | Neighborhood of candidate patch |
| $W_{\text{patch}}$ | Width of patch |
| $W_B$     | Width of block size |
measure the similarity. In each step, we past a best matching patch $P_{can}$ of $T$ into the synthesized texture $P_{cur}$ of $T_{hole}$. The similarity of $N(P_{cur})$ and $N(P_{can})$ is determined by the SSD equation (the sum of squared differences) in the neighborhood.

$$D(N_1, N_2) = \sum_p \{(R_i(p) - R_z(p))^2 + (G_i(p) - G_z(p))^2 + (B_i(p) - B_z(p))^2\}$$  \hspace{1cm} (1)

As above, R, G, B are pixels value in red, green and blue channels respectively. The searching neighborhood pixels of the proposed algorithm is scalable because $W_{patch}$ can be changed in different condition. Table 1 is the list of parameter.

2.1 Scalable Searching Neighborhood

The quality of the synthesized result is dependent on the size of neighborhood. When we use the algorithm in hole-filling, the scalable $W_{patch}$ can be changed by different hole size. After pasting a line at the first step, we synthesize the next patch of hole continually. An example of constrained texture synthesis in different conditions is shown in Fig. 1. The algorithm also can be used in image reconstruction which is shown is Fig. 2. When the lost region is a single line in a row, the inverse-U shape can be adjusted as the line. When the lost region is a single window over several columns, the shape can use the same shape from top to bottom. When $W_{patch}$ is $W_{L1}$, the best matching patch $P_{can1}$ is synthesized to $P_{cur}$. The synthesizing step is executed sequentially. In addition, when $W_{patch}$ is $W_{L2}$, the best matching patch $P_{can2}$ is synthesized to $P_{cur2}$. The total number of neighborhood pixels is $W_{patch} + 4$.

2.2 Synthesizing Order

Many constrained texture synthesis algorithm adopt spiral order to synthesize the lost block of texture. In our algorithm, we use regular scanning order from top to bottom. In Fig. 1, the hole is filled from the boundary texture in two directions: from top to center and bottom to center. In Fig. 2, the synthesizing order includes vertical direction and horizontal direction. In multiple- texture image, different synthesizing order can increase the flexibility of constrained texture synthesis.

2.3 Block-Smoothing Method

Fig. 3 shows the post-processing technique: block-smoothing method. The dotted line is the boundary of the human silhouette. The boundary blocks can be replaced by averaging the left ($W_{left}$) and right ($W_{right}$) blocks. In different texture background, the different size of $W_{B}$ can decide the quality of output image.

$$W_{avg}(x,y) = (W_{left}(x,y) + W_{right}(x,y))/2$$  \hspace{1cm} (2)
The neighborhood pixels is defined as follows:

\[ N(P_i) = \{ P_{i-1,j}, \ Pixel_{i-1,j-1}, \ Pixel_{i-1,j+1}, \ Pixel_{i,j-1}, \ Pixel_{i,j+1} \} \]

1 The neighborhood pixels is defined as follows:

\[ N(P_i) = \{ P_{i-1,j}, \ Pixel_{i-1,j-1}, \ Pixel_{i-1,j+1}, \ Pixel_{i,j-1}, \ Pixel_{i,j+1} \} \]

\( i \) means the row of the patch and pixels, \( j \) means the column of the patch and pixels.
3 Experimental Results

In order to verify the algorithm, we use MIT VisTex texture set [9] to test the proposed algorithm. The algorithm is implemented by MATLAB. The computation time is about several minutes in Pentium 4, 1.8G MHz. Example of human removal is shown in Fig. 4 and Fig. 5. In Fig. 4, the original image is shown in first row. The result after texture synthesis produces artifact (first image on second row). The artifact is reduced by the block-smoothing method whose block size is $3 \times 3$ (second image on second row). We use two directions to synthesize the lost line of input texture. The algorithm also can be used in larger hole size. Fig. 6 shows the experimental result of constrained texture synthesis used in hole filling. The size of the input texture is $256 \times 256$ and the hole size is $128 \times 128$. The comparison of the algorithms is shown in Table 2.

4 Conclusion

In this paper, we present an intelligent texture synthesis algorithm for smart camera. The neighborhood pixels can be changed by different patch size of the input texture. Most previous methods synthesize the hole from outside to center in spiral order. The novel regular scanning order is the first algorithm in constrained texture synthesis. Our method also need not much buffer to store the candidate patch. For this reason, the proposed algorithm is more suitable for VLSI implementation to be embedded in smart camera. The line-based searching pixels also can be reused in computing the similarity.
Fig. 4. Comparison of the proposed block-smoothing method

Fig. 5. Experimental result of human removal (the number of extended pixels is 6 and the block size is 7x7)

Fig. 6. Experimental result of image hole filling (The hole size is 25% of the whole image)
Intelligent Sub-patch Texture Synthesis Algorithm for Smart Camera

Table 2. Comparison of algorithms

| Method         | Searching pixels | Output pixels | Scalability |
|----------------|------------------|---------------|-------------|
| Wei [1]        | Pixel            | 12            | 1           | No          |
| Liang [3]      | Patch            | $4\times(\frac{1}{6}W_B^2) + 4\times(\frac{1}{6}W_B^2)$ | $W_B^2$     | No          |
| Hu [4]         | Patch$^2$        | $W_B^2 - n$   | n           | Yes         |
| Bertalmio [7]  | Pixel            | 64            | 1           | No          |
| Ours           | Sub-patch        | $W_{\text{patch}} + 4$ | $W_{\text{patch}}$ | Yes         |

References

1. G.R. Cross and A.K. Jain. “Markov random field texture models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(1):25–39, January 1983.
2. L. Y. Wei and M. Levoy, “Fast texture synthesis using tree-structured vector quantization”, Proceedings of SIGGRAPH, pp. 479-488, July, 2000.
3. A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer”, Proceedings of SIGGRAPH, pp. 341-346, 2001.
4. L. Liang, C. Liu, Y. Q. Xu, B. Guo, and H. Y. Shum, “Real-time texture synthesis by patch-based sampling”, ACM Transaction on graphics, vol. 20, pp. 127-150, 2001.
5. Y. H. Hu and R. A. Sambhare, “Constrained texture synthesis for image post processing”, IEEE International Conference on Multimedia and Expo, vol. 1, no. 6-9, pp. 113-116, July, 2003.
6. E. Simoncelli, and J. Portilla, “Texture characterization via joint statistics of wavelet coefficient magnitudes”, International Conference on image processing, vol. 1, pp. 62-66, October, 1998.
7. S. D. Rane, G. Sapiro, and M. Bertalmio, “Structure and texture filling-in of missing image blocks in wireless transmission and compression applications”, IEEE Transactions on image processing, vol. 12, no. 3, March, 2003.
8. M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Simultaneous structure and texture image inpainting”, IEEE Transactions on image processing, vol. 12, no. 8, August, 2003.
9. MIT Media Lab. Vision texture. http://www-white.media.mit.edu/vismod/imagery/Vision-Texture/vistex.html

$^2$ In Hu’s algorithm, $W_B$ is the patch block size and n is the pixels in target region.
Exploration of Image Features for Describing Visual Impressions of Black Fabrics

Chie Muraki Asano, Satoshi Hirakawa, and Akira Asano
Hiroshima University, Hiroshima 739-8521, Japan
chiem@mikeneko.jp

Abstract. Quantitative evaluation of human sensibility, or "Kansei," has recently attracted much attention. Evaluation of human visual impression is one of the important topics of this area. It is proposed in this paper that an objective evaluation method of human impressions based on the logistic discriminant analysis. This method applies the logistic discriminant analysis using various image features for sample images that have been already discriminated by sensory tests. An image feature yielding the logistic discrimination model of relatively high fitness to the human discrimination is regarded as that expressing human sensibility more appropriately than other features yielding low fitness. This method employs no subjective words to express human impressions but enables objective comparison of image features with respect to the ability of expressing human sensibility. This method was applied experimentally for exploring image features to describe human visual impressions of black fabrics. Furthermore, the comparison with the result of the discriminant analysis employing some contracted scores of the features by Principal Components Analysis (PCA) shows superiority of our proposed method.

1 Introduction

Evaluation systems of human sensibility, or Kansei, which occurs from human sense and cognition, have recently attracted much attention and interest. Especially, objective systems by the physical quantification values have been required. Various approaches to Kansei, based engineering design and their evaluation systems have been proposed in the field of Kansei engineering [1]∼[4]. These studies mainly apply multivariate statistical analysis to the extraction of relationships between results of sensory tests and physically observed features. Some of the conventional studies extract the relationships by employing the factor analysis with subjectively generated words to express human impressions, and some other studies analyze human impression using all of physical features that seem to have some. The former studies have a problem that the results are often interpreted with biases of the observers. The latter studies also have a problem that the results obtained by all the features are often difficult to be interpreted. Therefore, an effective method for feature selection is required and important in these cases.
We propose in this paper an objective evaluation method of human visual sensibility using feature selection technique based on the logistic discriminant analysis, instead of subjective descriptions based on the factor analysis. This method applies a sensory test to categorize a set of sample images. Then the logistic discriminant analysis using several image features is applied to distinguish an image pair extracted from one category and that from two different categories, and the parameter set of logistic model simulation expressing the human discrimination as appropriately as possible is derived. It is obvious that the logistic model is too simple to simulate the categorization based on the sensory test completely. However, we get from the fitness of logistic discrimination model that an image feature or a combination of image features yielding relatively high fitness are regarded to express the human categorization based on the sensory test more appropriately than other features yielding low fitness. Consequently, our method extracts image features relatively closely related to the human sensibility.

This paper shows the following feature sets which are developed and employed for the proposed method: texture anisotropy expressing orientation distribution of textural elements contained in an image, texture segment features expressing several geometrical characteristics of texture elements extracted by image segmentation using clustering of its intensity histogram, and periodic structures, which express frequency elements of textures with discrete fourier transform.

We apply our proposed method to the analysis of human sensibility in visual impression of black fabrics. The visual impression on black fabric is not influenced by color effect, but is affected by delicate differences of texture caused by fiber and/or textile structures. However, only a few quantitative evaluation methods of visual impressions have been proposed. We have recently investigated an evaluation method of human visual impression and the objective discrimination and evaluation by employing textural image features. The evaluation method of fabrics based on digital image processing has an advantage that it is reproducible because of the stability of captured images, contrarily to the instability of physical measurements of fabrics. The proposed method is applied experimentally to the exploration of the most effective ones from these textural features for describing human visual impressions of black fabrics in this paper. This study also presents fundamental knowledge for constructing an objective evaluation system of human visual impressions, which requires reproducibility.

2 Texture Anisotropy

Texture anisotropy is the direction distribution of texture elements. Uneven distribution of a texture suggests the existence of a directional structure in the texture. We evaluate texture anisotropy by a kind of histogram, called edge orientation histogram, indicating how frequently edges of each direction appear. Details of the feature are described in reference. Figure shows micrographs of some fabrics and their edge orientation histograms as an example.
3 Texture Segment Features

We introduce the concept of texture segment for characterizing texture images. A texture segment is a particle that is considered to compose a texture image. Statistics on geometries of texture segments characterize a texture. To extract the segments, the whole range of pixel values is divided into a small number of disjointed ranges. Each segment is extracted as a set of connected pixels assigned to the same range. An appropriate division for an effective extraction of the segments is achieved by an appropriate clustering of the intensity histogram. We apply $k$-means method for the clustering, and the total number of clusters is determined by the method described in [4].

Each pixel set assigned to the same cluster and connected by means of 4-neighborhood is defined as a segment. A unique label is assigned to each segment, and the following statistics on geometries of the segments are calculated to characterize a texture.

**Number of Pixels per Segment** - The mean of the number of pixels belonging to a segment for each cluster is calculated as a feature.

**Mean Length Ratio of the Distance Between the Foci to Major Axis** - The ellipse whose quadratic moment is the same as that of a segment is determined for each segment, and the ratio of the distance between the foci of the ellipse and its major axis length is derived as shown in Figure 2. The mean of the length ratio for each cluster is calculated as a feature.

**Mean Length Ratio Between Major and Minor Axes** - The ellipse as the above is determined, and the ratio between the major and minor axis lengths of the ellipse is derived as shown in Figure 2. The mean of the length ratio for each cluster is calculated as a feature.

**Number of Pixels per Segment** - The ratio of the number of pixels calculated for each segment against the ellipse determined as the above is derived as shown in Figure 2. The mean of the ratio for each cluster is calculated as a feature.
4 Periodic Structure Features (Periodicity Features)

Texture images such as fabrics have periodic structures. Thus it is important to extract characteristic periodic structures from the texture images. The images applied by two-dimensional discrete Fourier transformation (DFT), which is defined as follows, decomposes an image into some periodic elements.

\[
G(k, l) = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} f(x, y) \exp(-j2\pi x k/X) \exp(-j2\pi y l/Y)
\]  

(1)

Let \( X \) and \( Y \) be the height and width of the transformed image, and \( f(x, y) \) is the pixel value at a position \((x, y)\). \( G(k, l) \) is a complex number, and its strength is expressed to the absolute value called an amplitude spectrum, whose square is called power spectrum. The middle of Figure 3 shows a visualization of power spectrum whose pixel values. The center of this figure shows the spectrum with lowest frequency and the distance from that shows the height of a frequency. The following method is applied in order to extract characteristic frequency elements, and remainder pixels as shown in a right figure are employed as characteristic frequencies.

(1) The frequency elements of amplitude which are less than or equal to 10% of the maximum frequency are deleted, because most of these frequencies correspond to noise.

(2) \( n \times n \) size filter window is defined. If a target pixel of a power spectrum image shows the maximum and enough high value of amplitude in the window, the pixel is preserved. Otherwise the pixel is removed. This operation is applied to each pixels.

The range of \([0, 2\pi]\) is divided into 12 bins of equal interval on the basis of center of the image. The frequency element which have the longest distance between the center and its point is extracted at each bin.

Using the extracted frequency elements as shown in the right of Figure 3, following features are extracted from the image processed by \( 3 \times 3 \) filter window.
- The maximum and minimum and mean of amplitude of these frequencies.
- The maximum and minimum of the distance between the center of these frequencies.
- The area surrounded by these frequency elements.

Using four \( n \times n (= 3, 5, 7, 9) \) filters, the frequency images are applied to the above filtering operation. Then the transition of the number of extracted frequency elements is approximate to an exponential function \( \lambda e^{\alpha x} \). The fixed number \( \lambda \) and the coefficient \( \alpha \) is also calculated as features.

5 Discriminant Analysis

We explained the details of this analysis method in reference [4]. Some important definitions are shown in this section.

The set of distance for \( m \)-th image pair in the \( r \)-dimensional features’ space is defined by \( D_m = (d_{m1}, d_{m2}, \ldots, d_{mr}) \).

We define a binary random variable \( Y \) for the \( D_m \) as follows: \( Y = 1 \) if two images in \( m \)-th image pair are chosen from the same category and \( Y = 0 \) otherwise. We apply the logistic regression model to \( Y \) by explanatory variables \( D_m \).

\[
p(D_m) = \frac{1}{1 + \exp(-Z(D_m))}, \quad Z(D_m) = \beta_0 + \sum_{s=1}^{r} \beta_s d_{ms} \tag{2}
\]

where \( Z(D_m) \) is a discriminant function described with the constant term \( \beta_0 \) and the coefficients \( \beta_s \) for each term, and \( p(D_m) \) is the probability of \( Y = 1 \). We compare 0.5 and \( p(D_m) \). If \( p(D_m) \) is larger than 0.5, the \( m \)-th image pair is determined to be in different categories, otherwise in the same categories. The probability of \( Y = y_m \), where \( y_m \) takes 0 or 1, is defined as follows:

\[
Pr\{Y = y_m|\beta\} = p(D_m)^{y_m}(1 - p(D_m))^{1-y_m}. \tag{3}
\]

In fact, maximizing the log likelihood about this probability, the maximum log-likelihood, which is denoted by \( MLL \), and the estimated parameter set of \( \hat{\beta} \) are calculated.

6 Features Selection and Discriminant Function Evaluation Based on AIC

The discriminant function derived in a discriminant analysis often employs a number of features. The ability of the function is evaluated by the small discrimination error rate in general. However, employing all available features does not always maximize the discrimination ability. Thus we introduce a method of selecting feature combination that is fit to the categorization based on human sensibility. And in our problem of interest, the optimal fitness and likelihood of a feature combination are important, since the discriminant function is evaluated
not only on the discrimination ability for the sample data but also on the low error rate expected for unknown data.

We propose an exploration method of the optimal feature combination by employing AIC(Akaike’s Information Criterion)[7], which is one of information criteria for statistical model selection. Our feature selection method iterates the following procedures for each subset of all available features:

1. Selecting a feature combination from the subset, and measuring the distance between arbitrary two images with respect to the feature combination.
2. Deriving $Z(D_m)$ by estimating $\hat{\beta}$ using the maximization of the log-likelihood log like($\beta$). The maximum log-likelihood is denoted $MLL$.
3. Calculating the discrimination error rate for all the image pairs, denoted $W$, that for some of those classified into different categories by human sensibility, i.e. in the case $Y = 0$, denoted $W_D$, while that for other classified into one category, i.e. in the case $Y = 1$, denoted $W_S$.
4. Calculating $AIC = -2 \times MLL + 2 \times \alpha$, where $\alpha$ is the number of employed features in the feature combination.

The feature combinations are ranked both by the AIC values in ascending order and by the error rates in ascending order. Basically, a discriminant function with a feature combination yielding the smaller AIC value is rated higher. If the AIC values of two discriminant functions are similar, the function yielding the lower error rate is rated higher. A high-rated feature combination yields a discriminant function of high fitness to the human categorization, and is regarded to be relatively closely related to the human sensibility.

7 Experiments and Evaluation

Thirteen samples of black fabrics were used for this study. It is known that some impressions on black fabric are influenced by the illumination condition but other impressions are stable. In order to explain this observation, two images which differ on the illumination condition were captured from each sample fabric, that is, one image was captured under a simple natural light while the other was illuminated by an additional spotlight. Furthermore, four different subimages were extracted from each image. All of these subimages ($104 = 13 \times 2 \times 4$) were used as sample images for the discrimination and exploration method mentioned above. The discriminant experiments were carried out for the following two cases:

Case A: Two sample images from one fabric are regarded to be in one category regardless of the spotlight illumination.

Case B: Two sample images from one fabric under different illumination conditions are regarded to be in different categories.

Combinations of image features describing visual impression appropriately were explored in each case by the discriminant analysis. Besides the proposed method of discriminant analysis, the dimensions of the features are reduced by Principal Components Analysis (PCA), and the components data were employed
to the discriminant analysis. This analysis was applied in order to validate this research.

Table 1 shows the results of discrimination in Case A with the best combinations explored from all image features combinations by the proposed method. The features $X_1 \sim X_{30}$ in this table show the image features, where $i = 1, 2, \cdots, 10$ is assigned for each class in descending order of edge intensity, $i = 11, 12, \cdots, 22$ is assigned for each segment features extracted from major 3 clusters of texture segments, and $i = 23, \cdots, 30$ is assigned for each segment periodicity feature respectively. $W, W_D$ and $W_S$, shown besides AIC in the table, denote the error rates (%) derived in Sec. 6. Table 2 shows the results of discrimination in Case B.

Table 1. The result of discrimination with the best combinations of image features in Case A

| Combination of features | $W_D$ (%) | $W_S$ (%) | $W$ (%) | AIC |
|------------------------|-----------|-----------|---------|-----|
| $X_1, X_5, X_{11}$    | 20.83     | 8.51      | 14.67   | 564.09 |
| $X_1, X_5, X_{10}, X_{30}$ | 19.85 | 9.89      | 14.87   | 555.30 |
| $X_1, X_5, X_{10}, X_{16}, X_{19}$ | 18.95 | 11.26     | 15.10   | 546.55 |
| $X_1, X_5, X_{10}, X_{11}, X_{26}, X_{29}$ | 19.25 | 8.79      | 14.02   | 550.65 |
| $X_1, X_5, X_6, X_{10}, X_{11}, X_{16}, X_{30}$ | 19.47 | 8.51      | 13.99   | 550.33 |

Table 2. The result of discrimination with the best combinations of image features in Case B

| Combination of features | $W_D$ (%) | $W_S$ (%) | $W$ (%) | AIC |
|------------------------|-----------|-----------|---------|-----|
| $X_1, X_6, X_{16}$    | 17.75     | 8.97      | 13.36   | 223.96 |
| $X_1, X_6, X_{10}, X_{16}$ | 16.23 | 7.69      | 11.96   | 215.75 |
| $X_1, X_6, X_{10}, X_{16}, X_{25}$ | 15.76 | 8.33      | 12.05   | 211.49 |
| $X_1, X_6, X_{10}, X_{19}, X_{27}, X_{29}$ | 15.98 | 8.33      | 12.15   | 215.76 |
| $X_1, X_6, X_{10}, X_{14}, X_{16}, X_{24}, X_{25}$ | 15.46 | 8.33      | 11.89   | 213.10 |

Table 3. The results of the discrimination using principal components of all image features by PCA

| Number of PC | Case A |           | Case B |
|--------------|--------|-----------|--------|
|              | $W_D$ (%) | $W_S$ (%) | $W$ (%) | AIC | $W_D$ (%) | $W_S$ (%) | $W$ (%) | AIC |
| 1            | 26.10  | 15.11     | 20.61   | 670.45 | 22.21 | 14.74     | 18.48 | 262.61 |
| 2            | 21.69  | 10.71     | 16.20   | 599.63 | 19.90 | 10.90     | 15.40 | 245.89 |
| 3            | 21.75  | 10.99     | 16.37   | 601.53 | 19.94 | 10.90     | 15.42 | 247.78 |
| 4            | 21.73  | 9.62      | 15.68   | 600.22 | 19.81 | 10.90     | 15.35 | 249.73 |
| 5            | 21.83  | 9.89      | 15.86   | 601.57 | 19.58 | 9.62      | 14.60 | 248.55 |
| 6            | 20.91  | 7.97      | 14.44   | 586.07 | 25.08 | 11.54     | 18.31 | 270.75 |
| 7            | 39.20  | 9.34      | 24.27   | 728.20 | 25.62 | 10.26     | 17.94 | 272.45 |
Case B with the best combinations. This shows that the result of the combination of five features shows the lowest AIC value while the combinations of six or more features show lower discriminant error rates. Tables 3 shows the results of the discrimination using the principal components calculated from all image features by PCA in Case A (left side) and in Case B (right side) respectively. This result shows discriminant analysis by the proposed method is superior to that by a contracted dimension of the features by PCA.

8 Conclusions

In this paper, we have proposed a method exploring a logistic discrimination model highly fitted to the discrimination by human sensibility, and extracting the feature combinations employed in the explored discrimination function as an appropriate feature combination closely related to the human sensibility. The fitness of the logistic discriminant model has not been evaluated by rating the discrimination error only, but also by the information criterion for the estimation of the fitness and likelihood expected for unknown data in this study. The experiments have shown that not only the features in a specific intensity and those of periodic structure but also the features that are hardly recognized such as the features of edge in low intensity and those of shape in small segments improve the discrimination. In addition, the result of the discriminant analysis using contracted dimension of the features by PCA has shown that the proposed method brings improved results.

References

1. M. Kihara, I. Yoshizawa, M. Murakami, C. Muraki Asano, A. Asano and T. Fujimoto, The Determination of "pure blackness" of the Fabric in Women’s Formal Dresses, Proc. 6th Asian Textile Conference, paper no. 265 (2001).
2. S. Hirakawa, A. Asano and C. Muraki, Employment of clustering characteristics using information criterion on low-contrast image identification, IEICE Technical Meeting of Pattern Recognition and Media Understanding, Tokyo, Japan, PRMU2002-237, (2003)85–90.
3. H.J Park, E. Koyama, T. Furukawa, M. Takatera, Y. Shimizu and H. Kim, An Impression Evaluation Model for Apparel Product Retrieval Based on Image Analysis, 3, 1(2001)11–18.
4. S. Hirakawa, C. Muraki Asano, N. Hamahashi, and A. Asano, "Discriminative Exploration of Image Features for Describing Visual Impressions of Black Fabrics," Third International conference on Hybrid Intelligent Systems, Melbourne, Australia, 509-518 (2003. 12).
5. A. Soria-Fisch, M. Köppen and T. Sy, Usage of Fuzzy Aggregation for Perceptual Relevance Evaluation Authors, Proc. Int. Conf. on Information Processing and Management of Uncertainty, Annecy, France, (2002)953–960.
6. E.J. Wood, Applying Fourier and Associated Transforms to Pattern Characterization in Textiles, Textile Research Journal, 60(1990)212–220.
7. H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Contr., AC-19(1974)716–723.
Distributed Resource Allocation via Local Choices: General Model and a Basic Solution

Marian F. Ursu¹, Botond Virginas², and Chris Voudouris²

¹ Department of Computing, Goldsmiths College, University of London, London, SE14 6NW, UK
m.ursu@gold.ac.uk
http://homepages.gold.ac.uk/marian-ursu
² Intelligent Systems Lab, BT Exact, Adastral Park, Ipswich, IP5 3RE, UK
{botond.virginas, chris.voudouris}@bt.com

Abstract. This paper describes a solution to resource allocation, modelled as a distributed system. The solution was developed to complement an existing system built for the same purpose, but in a centralised approach (i.e., based on a central multi-criteria optimisation algorithm). Both are part of ARMS (Automated Resource Management System), an integrated system for the customer service operations of British Telecommunications plc. Resource allocation is modelled here as an iterative communication process, based on a 4-step communication protocol, between resources agents, on one hand, and requirements agents, on the other. Agents “own” disjoint sets of resources or requirements – i.e., they have complete decision power over their allocation. The global allocation emerges from this interaction/communication, where choices/decisions are made locally, within each agent. The paper describes the main aspects of the distributed system and illustrates, with a concrete implementation, the emergence of a good global solution from the distributed algorithm.

1 Introduction

Accurate resource management is critical to a company’s performance and profitability. It is crucial that a company manages its resources efficiently and effectively in order to remain competitive. The need for automated resource management is well recognised and has been the subject of considerable research and development [1]. The Intelligent Enterprise Technologies group within BTExact Technologies, building on previous experience in resource management, has developed ARMS [2], an Automated Resource Management System. The main application of ARMS is the continuous allocation of engineers to jobs in a large organization. At the core of ARMS are patented Operational Research and Artificial Intelligence techniques.

1 The work described in this paper was undertaken within the British Telecommunications plc Short-Term Research Fellowship Scheme, Summer 2003.
ARMS has three major functional components: *Forecasting and Job Generation* (FJG), *Dynamic Planner* (DP) and *Collaborator*. FJG takes as input historical job information and produces forecasts of job volumes. DP finds optimal allocations of engineers to jobs, using both forecasted and real jobs, so as to facilitate efficient and cost-effective scheduling of engineers by BT’s Work Manager System. DP is a single-domain resource-planning tool operating at local CTS (Customer Service Team) level. Collaborator has been built on top of DP to cater for *multi-domain allocation* (operating at RBA - Regional Business Area level) of resources/engineers that remained unused by DP, to requirements/jobs that were left unresolved by DP. Two main approaches were taken in the development of Collaborator: *global/centralised* and *distributed*. They are discussed below.

2 Global Versus Distributed Approaches to Resource Allocation

There are two main approaches that can be taken for resource allocation. Obviously, they represent extreme points of view, many combinations between them being possible.

**Global/Centralised Approach.** The organization’s resources are distributed centrally. The focus is on the company’s global interests and priorities. They need not be propagated and expressed at local level. They are expressed via a set of, usually conflicting, global criteria. The decision power at local level is, thus, not apparent. The computational model suitable for this approach is that of a *central multi-criteria optimisation* algorithm [3].

**Distributed Approach.** The focus is on local interests and priorities. They are (partially) achieved incrementally through interactions between local managers. The interactions are entirely driven by the managers’ own decision powers. The company’s overall interests and priorities are not apparent – there is no central manager to enforce them. Their accomplishment emerges (should emerge, rather) from the individual accomplishment of those of the local managers (note that some may be explicitly propagated and expressed at local level).

The version of Collaborator built in the *global/centralised approach* was reported in [4]. Its focus is on the balancing of the workforce across multiple domains (possibly across the entire organization). This requirement is expressed via a number of conflicting criteria, such as “maximise the number of jobs resolved”, “minimise the number of engineers used” and “minimise the overall travelling distance”. Collaborator uses a central Pareto multi-criteria optimisation algorithm [3] to infer a subset of optimal solutions – the Pareto set – all “equivalent” from the point of view of the stated criteria. Local domain preferences (e.g. preferred working time and travelling destination) are then employed, still centrally, to select the best-preferred solution from the Pareto set. The version of Collaborator built in the *distributed approach* – Distributed Collaborator (DC) – is outlined in the remainder of this paper.

Several authors have proposed that distributed allocation problems be solved via market mechanisms [5,6]. Market-based control is a paradigm that can be applied to
complex systems when other forms of control are not feasible. The solution presented here subscribes to this paradigm.

3 System Model in a Distributed Approach

3.1 Structural Model

Large organizations are normally partitioned in domains. Different criteria may be considered as the basis of such partitions. BT and Distributed Collaborator (DC) use spatially based partitions. Each domain has a set of resources and a set of requirements. Each domain has decision power over its requirements and resources. In the distributed approach, the computational model for such an organization consists of a set of autonomous and interacting agents. The global allocation of resources is accomplished through communication and exchanges between agents. This may involve negotiation, but it is not necessarily required. Our initial model is based on non-negotiable offers – i.e., they are either accepted or not (the nature of an offer needs not to be specified at this stage).

The core structural element of resource allocation in a distributed approach is that of an atomic allocation. Within the context of this paper, an atomic allocation is made only between two agents: namely between a set of atomic resources (A, B and C, in Fig. 1), belonging to one resources agent, and a set of atomic requirements (1 and 2, in Fig. 1), belonging to one requirements agent.

The resources must match the requirements to which they are allocated. Different application domains may have different meanings for the term “match”; different meanings may be possible even within the same domain. For example, for the allocation of engineers to jobs, “matching” may mean “can carry out to completion” or “can complete within the specified deadline”. Note that the definition of an atomic allocation considered here is not inherent to the distributed model. Yet, it simplifies the problem without considerably affecting its generality.

Each domain has two corresponding agents, one for its requirements and one for its resources. They enact the domain’s interest and priorities via the decisions/choices they make. Local interests and priorities may be expressed in two ways: (1) (declaratively) via a set of criteria that are to be minimised/maximised (e.g., minimise the engineers’ travelling distance); or (2) (procedurally) via set of rules of operation or...
strategies (e.g., attend to jobs of higher priority first). We consider both forms of expression. More so, extending our work on design regulations [7], we are working towards declarative representations for rules of operation.

In a purely distributed approach, all the intelligence is located/represented in the constituent agents – i.e. in the leaves of the system. The overall behaviour of the system emerges completely from the agents’ interaction. It is possible to relax the purely distributed approach, and to enforce certain interaction structures via central agencies (otherwise they, too, will have to be emergent). Examples of interaction structures include explicit asynchronous co-ordination (e.g., collaboration, competition and negotiation) and synchronism (e.g., request resources by a certain deadline).

3.2 Data Model

So far, the discussion was carried out mainly in terms of “requirements” and “resources”. They are too generic for the construction of a solution to a specific problem. In each particular case, they will have to be concretised and described as accurately as possible. This section provides a brief illustrative example of the data model employed in DC (refer to the ER diagram in Fig. 2). For the remainder of this paper, requirements are taken to be jobs and resources are taken to be engineers.

A domain may have a number of jobs and a number of engineers. A job has an inception, a deadline by which it has to be completed, a set of required skills (that an engineer will have to have in order to carry it out) and an expected duration (the period in which an engineer with average productivity for the required skills completes it). Optionally, a job may be regarded as being made of other jobs and may require (as prerequisite) the completion of some other jobs. An engineer has a set of “skill-productivity/expertise” pairs and a number of preferences (with regards to the locations where they wish to travel and the skills they wish to employ). Engineers are assigned to jobs. Each assignment is stamped with the date when it is supposed to start, the date when it is supposed to be finished, and the actual date of completion. The solution described below, uses a slightly simplified data model.
4 Basic Solution

This section outlines the solution devised for DC and illustrates it in a specific implementation. The solution is quite comprehensive – in that it can be employed with a variety of data models – and versatile – in that it constitutes a good basis for further extensions.

4.1 General Design

The distributed solution consists of a society of interacting agents, having the characteristics outlined by the structural model (Section 3.1). There are agents for jobs – JobsAs – and for engineers – EngsAs. The allocation of resources is modelled as an iterative communication process, where each iteration consists of a 4-step communication protocol. A central agent – Monitor – is used for synchronisation (required for the enforcement of the communication protocol and the iterative process). The communication protocol (basic iteration) is outlined below.

1. JobsAs: broadcast requests (not binding)
2. EngsAs: reply with offers (not binding)
3. JobsAs: send contract proposals (binding)
4. EngsAs: reply with agreed contracts (binding)

In the first step, all JobsAs broadcast requests, based on their job requirements. In the simplest case, each JobsA broadcasts requests for all its jobs to all the EngsAs. This is possible because requests are not (legally) binding. This is the solution initially adopted in DC. We are now experimenting with more elaborated strategies, represented by both collaboration and competition heuristics. In particular we carry out trials with preferential and bogus broadcasts. For the former, each JobsA has a (local) classification of all the EngsAs in levels of preference (criteria for preference could be “location/nearest”, “partner/historically-good”, etc.). Broadcasts are initially made just to “the most preferred” partners, then, after these stop responding (after a few iterations), broadcasts are extended to the next group of partners, and so on. For the latter, broadcasts are made with requests that are different from the real ones, in an attempt to attract better offers.

In the second step, EngsAs respond with offers. Each EngsA has its own way of formulating offers (local choices). Offers are not (legally) binding. Thus, this step, too, may be regarded as a way of efficient information gathering. At one extreme, offers can be made with all matching resources to all JobsAs. This leads to good information gathering, but is not efficient. This is the solution adopted in DC. The other extreme is not well defined. Various systems of “partners” (see above) and/or investigation techniques (e.g., bogus offers) may be employed. The relevance of these methods increases with the amount of competition that exists or is allowed in the system.

In step three, JobsAs, faced with various offers, must decide to whom to propose contracts (legally binding). If all EngsAs’ offers consist of “real” resources, then the mere application of the local interests and priorities as main criterion of decision is
Distributed Resource Allocation via Local Choices

appropriate (with the downside of knowing that contract proposals are not guaranteed to be accepted). If not, then competition strategies will have to be considered at this step too. DC subscribes to the former case: each JobsA aims to maximise the number of jobs attended to and to minimise the travelling costs it has to support. This could be achieved via a local multi-objective optimisation algorithm. Subsequently, this would represent a natural partition of a global optimisation problem into a set of smaller and independent optimisation problems. DC takes the “strategy route” – it implements strategies of compiling contract proposals, mimicking the behaviour of CST managers. Strategies are currently expressed as procedures and are hard-coded in agents. We are also experimenting with (declarative) notations for their representation.

In **step four**, EngsAs decide whom to contract (some of) their resources. An interesting angle to the problem is brought by the fact that atomic allocations (Fig. 1) employ groups of jobs and of engineers. DC simplifies this by having only one engineer and one job employed in an atomic allocation/assignment. Since contract proposals are binding, EngsAs can make their decisions just on the basis of their local interests and priorities. For DC, each EngsA attempts to maximise the number of jobs to attend, whilst maintaining the overall travelling distance between reasonable limits. Here, too, DC takes the “strategy route”.

The synchronisation of the agents’ behaviour, necessary for the realisation of the 4-step communication protocol, is achieved via the Monitor. This ensures that a step is initiated only after all the agents have indeed completed the previous step. However, this solution lends itself easily to other methods of **timing**. An asynchronous model, delegates the decision of when an agent should intervene in the overall environment to the agent itself – this too, then, becomes a matter of a local choice, based on local criteria and/or strategies. A combined solution can be accomplished via a system of deadlines. Agents are allowed to decide when to initiate a process, but only up to a specific deadline that is enforced by the Monitor. The Monitor or different “central” agents may be employed in other types of mediation. This would lead to a mixed centralised-distributed approach/solution and will be reported elsewhere.

The solution outlined here is an excellent basis for the implementation of different experimental prototypes. The following section briefly presents such a prototype.

### 4.2 Implementation and Results

For the solution outlined above, we have implemented a DC prototype in Java using JMS (Java Messaging Service) on a BEA Weblogic Server 6.1. For the communication process between agents we employ one JMS topic. Our experiments were carried out with various data sets, but they were all based on simple local strategies. We manually compiled a series of cases that mirrored the data sets used in real life. They had, approximately, the following characteristics: 30 active agents, between 5 and 10 jobs/engineers per agent; jobs of 1 day length (therefore, allocation could be done per day); local interests: maximise number of allocations and minimise travelling distance. The results obtained were very encouraging. The emergent global solutions were of very good quality. In each case, the solutions converged rapidly, in not more
than 6 iterations. The running time, on an average PC, did not exceed 5 minutes, including the time consumed by data output tasks, and this considering that the efficiency of the procedures that implement local strategies was not in focus. Because the procedures that implemented local strategies were quite simple, we managed to construct cases when better global solutions could be constructed manually. However, we are confident that more accurate representations (e.g. algorithms) and slightly more complex strategies would be sufficient for the great majority of real life cases.

An interesting experimental avenue opened by these case studies was the sensitivity of the results to data ordering (as input or as represented locally). Although we obtained different global solutions depending on data ordering, in most cases, they were equally good.

A problem we encountered was the lack of formal definitions for satisfactory, good and bad solutions, and also for what makes a solution better than another one. We carried out qualitative analyses manually, using suitable visualisation techniques. These were possible due to the nature of the local priorities (suitable for visualisation).

5 Conclusions and Future Work

The work reported in this paper is aligned with the recent growing interest in the AI community in the study of emergent global behaviours of distributed intelligent systems, where the intelligence is mainly located in the leaves. Even with simple prototypes, we achieved very encouraging results.

Our immediate objectives are: (1) to experiment with richer data models and data sets; (2) to refine local strategies and implement them more accurately; (3) to devise declarative representations for local strategies; and (3) to devise suitable analysis concepts (e.g., “good solution”) and build associated tools. The latter is particularly relevant when local criteria are not anymore readily suitable for visualisation.

We are also planning to: (1) explore the usefulness of central agencies (such as the Monitor, in DC) for handling exceptions and emergencies; (2) extend the solution to dynamic resource allocation (i.e., as resources and requirements arise); and (3) experiment with asynchronous communication models.

References

1. Simpson, J., Noble, A., Egan, B., Morton, D., Richards, T. and Burstin, M. Experience in applying OR techniques to the solution of practical resource management problems. In: BT Technology Journal, (1995), 13, 1, 16-28.
2. Voudouris, C., Owusu, G., Dorne, R., Ladde, C and Virginas, B., ARMS: An Automated Resource Management System for British Tellecomuniction plc, In: EURO/INFORMS Joint International Meeting EURO Excellence in Practice Award, Istanbul (2003).
3. Ehrgott, M. and Gandibleux, X. (eds): Multiple Criteria Optimization: State of the Art Annotated Bibliographic Survey, Kluwer's International Series in Operations Research and Management Science, Volume 52, Kluwer Academic Publishers, Boston, (2002).

4. Virginas B., Owusu G., Voudouris C. and Anim-Ansah G. A two stage optimisation platform for resource management in BT. In: Twenty third Annual International Conference of the British Computer Society's Specialist Group on AI, (2003), 109-121.

5. Clearwater, S. Market-Based Control: A Paradigm for Distributed Resource Allocation. (ed. Clearwater, S), World Scientific Publishing, Singapore (1996).

6. Wellman, M.P, A market oriented programming environment and its application to distributed multicommodity flow problems. Journal of Artificial Intelligence Research, (1993), 1, 1-23.

7. Ursu, M.F. and Zimmer, R.: Compliance in Critiquing Intelligent Design Assistants: Isomorphic Representation of General Norms and Exceptions. In (eds: Damiani, E., et al): The Sixth International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES02, IOS Press (2002), 98-104.
Behavior Profiling Based on Psychological Data and Emotional States

Rajiv Khosla, Chris Lai, and Tharanga Gooneseokera

Business Systems & Knowledge Modelling Laboratory, School of Business, La Trobe University, Victoria 3086, Australia
{R.Khosla,C.Lai,T.Gooneseokera}@latrobe.edu.au

Abstract. The provision of web cameras has opened possibilities of user behaviour profiling and web personalisation based on their emotional state. In this paper we report on a neural network model for emotional state profiling and its implications for behaviour profiling and personalisation.

1 Introduction

Most existing approaches of salesperson’s recruitment rely on interview process and psychometric techniques [1] for evaluating behaviour profiling and behaviour categorization of a sales candidate. These approaches have met limited success. We have developed Salesperson Recruitment and Benchmarking System (SRBS), which is based on integration of selling psychology based behaviour model and intelligent technologies like expert systems and self organizing maps. SRBS employs psychological data to determine a sales candidate’s selling behaviour category.

Emotional Intelligence has gained increasingly popularity in recent years, especially in its application to the workplace and personal development [2,3,4]. It has been associated with the ability to provide constructive feedback, team achievement and fits within the configuration of the total personality [4,5]. Within the context of personality, temperament and moods, emotions rise and fall in short-term cycles. In this paper we report on the preliminary work of correlating a candidate’s emotional state (eg; happiness, anger…etc.) with candidates selling behaviour category. The idea of correlation is to determine whether the emotional state confirms or negates the other psychological data related inputs provided by the candidate to SRBS.

The paper is organized as follows. Section two outlines the selling behaviour model. Section three outlines design and results of the SRBS and section four outlines design of neural network used for modelling the emotional state information of the candidate. Section five provides a discussion on the correlation between the psychological emotional data inputs and their implications for user profiling in general. Section six concludes the paper.

2 Selling Behavioural Model

The behavioural model developed by Buzzotte, Lefton and Sherberg [7] had been used for building predicting selling behaviour profiling. Another similar study done
by Anderson [8, p33] developed a social style matrix frame [9] for combining the assertiveness and responsiveness scales to establish social style. The behavioural model [10, p17] used here is shown in figure 1.

| Dominant-Warm | Dominant-Hostile |
|---------------|------------------|
| Sales are made when customers become convinced that they can satisfy a need by buying. The salesperson’s job is to demonstrate to the customer that their product would best satisfy the customer’s need. | The salesperson must impose their will on the customer by superior determination and strength. Selling is a struggle the salesperson must win. |

| Hostile | Warm |
|---------|------|
| Submissive-Hostile | Submissive-Warm |
| Customers buy only when they are ready to buy. Since persuasion does not work, salesperson’s job is to take their order when the customer is ready to give it. | People buy from salespersons they like. Once a prospect becomes a friend, it is only reasonable that he should also become a customer. |

**Fig. 1. Salesperson Behaviour Profile [10, p171]**

It has two dimensions namely, ‘Warm-Hostile and Submissive-Dominant’. A warm person is optimistic and willing to place confidence in others. Hostility is lack of regard for others, the attitude that other people matter less than oneself. A hostile person rarely trusts others. Submission is the disposition to let others take the lead in personal encounters. Dominance is the drive to take control in face-to-face situations. This model has been used based upon interactions with senior managers in the sales and human resources arena in the consumer and manufacturing industries in Australia [10].

### 3 SRBS Design and Results

For analysing the selling behaviour profile of a salesperson 17 areas have been identified for evaluation of a sales candidate behaviour profile as *selling as a profession, assertiveness, decisiveness, prospecting, product, customers, competition, success and failure, boss, peers, rules and regulations, expenses and reports, training, job satisfaction, view about people, relationship with non-selling departments, general attitudes* [11]. These areas have been identified after several discussions with sales managers and knowledge available in the literature [6, 12, 13]. Weights have been assigned to 17 areas on a scale of 1 to 10 using AHP (Analytical Hierarchy Process) technique [14]. The different behavioural categories have been determined in the form of a questionnaire. A sample set of four questions related to the area of competition is shown in figure 2.
1. In sales, the law of the jungle prevails. It’s either you or the competitor. You relish defeating your competitors, and fight them hard, using every available weapon.

| Behavioural Category | DH |
|----------------------|----|

2. You may not be aggressive otherwise, but when it comes to competition you are just the opposite. You spend a good deal of your time explaining to the customer why he should not buy from the competitor.

| Behavioural Category | SH |
|----------------------|----|

**Fig. 2.** Psychological Data Input: Sample Questions Related to the Competition Area

```plaintext
IF max (score DH, score SH, score SW, score DW) = score DW
   AND
   score DW / Total score < 0.65
THEN
   Pursue max (score DH, score SH, score SW)

IF
   Pursued category = DH
   AND
   score SH / score DH > 0.6
   score (SW + DW) / score (DH + DW) <= 0.9
   score (SH + SW) / score (DH + DW) >= 0.7
THEN
   Pursue max (score SH, score SW)
```

**Fig. 3.** A Sample Selling Behaviour Categorization Rule

Selling behaviour analysis was carried out on the feedback given by the salesperson on themselves and to determine the primary behavioural category of the salesperson. An answering pattern was determined based on accumulated answer score to all the questions in each behavioural category. A typical selling behaviour categorization heuristic used to determine the selling behavioural categorization is Figure 3.

**Fig. 4.** Candidate Result Screen with Area-wise Break-up

**Fig. 5.** The Comparison of Candidate Result with Benchmarking Profile
Area wise distribution of candidate’s behavior profile shows (figure 4) the overall distribution of four category scores. SRBS creates a benchmark profile for all candidates who are evaluated by it as shown in Figure 5. The dark line represents the candidate’s profile and lighter line represents the benchmark profile. This profile comparison bar chart provides feedback on the extent of similarity/difference between the two profiles.

4 Neural Network Based Facial Expression & Emotional State Modeling

The proposed system attempts to make use of the candidate’s emotional state to determine the correlation or commitment of the candidate has to the entered response. It is proposed that each of the four behaviour categories have associated with them a characteristic emotional profile. That is the emotional response or state of the candidate at the time of answering the questions can be also used as an indicator of which of the four behavioural profiles the candidate can be classified into. For example a candidate with overall behaviour category DH will have a positive emotional response when answering a question relating to the DW behavioural category such as DH question shown in figure 2 and a negative emotional response when answering a question relating to, say, the SH behavioural category also shown in figure 2. It is proposed that where a correlation does not exist between the emotional state and the response to a question that that particular area be further examined at the interview stage in order to accurately ascertain the behavioural category of the candidate.

Happy, angry and neutral emotional states (shown in Figure 6) were used to indicate positive, negative and neutral emotional responses. The work presented here is preliminary and plans to further expand the range of emotions analysed are in place. The number of emotional states was kept to a minimum in order to show a proof of concept. We have employed facial expression as a means of determining a candidate’s emotional state and its application to recruitment and benchmarking of salespersons.

Fig. 6. Happy and sad expressions images, middle, were subtracted from neutral images, left, to generate difference images, right. Classification was improved for non neutral expressions when difference images were used
4.1 Representation

Two common techniques being used for feature representation in face recognition and analysis are eigenvectors [14] and Gabor wavelets [16,17,18,19]. Daugman [16,17] presented a description of the 2D complex Gabor transform and generalised the Gabor function. Marcelja [18] and Daugman [16,17] presented that simple cells in the visual cortex can be modelled by Gabor functions. Lee [19] derived a family of 2D Gabor wavelets that satisfies the wavelet theory and the neurophysiological constraints of simple cells. Lyons, Akamatsu, Kamachi and Gyoba [20] used instances of Gabor filters at varying scales and orientations to represent facial expression images. We adopt the methodology used by Lyons, Akamatsu, Kamachi and Gyoba to represent images in this paper.

Image representation was by a Gabor vector calculated at points on a grid overlaid on the normalised image. See figure 7. The values of the gabor vector were the magnitude of the complex valued Gabor wavelet filter calculated at points on the grid within the image. The filters used were at three scales an octave apart and at six orientations within each scale. The bandwidth of the masks was approximately an octave. The response of the filter is given by

\[ R_k(\tilde{r}_0) = \int \psi(\tilde{k}, (r_0 - r)) I(r) d\tilde{r} \]

where

\[ \psi(\tilde{k}, \tilde{x}) = \frac{||\tilde{k}||^2}{\sigma^2} \exp \left( -\frac{||\tilde{k}\|^2 + ||\tilde{x}\|^2}{2\sigma^2} \right) \left[ \exp (i\tilde{k} \cdot \tilde{x}) - \exp \left( -\frac{\sigma^2}{2} \right) \right] \]

Fig. 7. G grid overlaid on the normalized image defining points where the gabor vector is calculated.
The wave vector $\vec{k}$ determines the scale and orientation of the Gabor filter. The substraction of the term $\exp\left(-\frac{\sigma^2}{2}\right)$ from the real part of the filter renders the filter to image illumination. The imaginary part of the filter is insensitive to illumination by default. The real and imaginary parts of the filter are often referred to as odd and even filters respectively.

4.2 Classification

Classification was done by a multilayer feed forward network using backpropagation training. The architecture was 1296 input nodes, 10 middle layer nodes and 3 output nodes. The training set for the network consisted of the images from six of the seven subjects with the seventh subject being the test set. That is training and testing was done on a “leave one out basis” for each of the subjects.

4.3 Results

Each of the images in our data set was tested with a network trained on the remaining images to generate the overall performance of the system. The results of training/test runs are given in table 1. A total of 63 networks were trained and tested on a “leave one out basis” to produce the results below. The naming convention used for the subject column is $s<\text{subject number}><\text{emotion class}>_\text{norm}_<\text{diff}>$ where subject number is between 1 and 7, emotion class is \{n,h,a\} which is equivalent to \{neutral, happy, angry\}, _norm indicates the image is normalised or derived from normalised images and _<diff>_ indicates whether the image is a difference image (see figure 6) or not.

5 Discussion

The psychologists point out that facial expressions can be effectively used for determining the change in emotional state of a person. In this paper we have reported on the first stage of our integration of emotional state information with psychological inputs provided by a sales candidate during an interactive session with a computerised SRBS. That is, the neural network model of “anger”, “happy” and “normal” emotional states represents the first stage of integration. At present we are in the process of integrating the neural network for modelling the emotional state of a sales candidate in real-time interactive session with SRBS. This work has implications for user profiling of internet and other ICT device (e.g., mobile phones) users for personalisation of information content. In other words, the emotional state of a user can be employed to determine the degree of interestingness in the information presented to the user.
6 Conclusion

In this paper we have reported on two methods, namely, psychological data based behaviour profiling and emotional state profiling using intelligent techniques in sales recruitment domain. We have discussed the implications for combining psychological and emotional data for selling behaviour profiling and user profiling and personalisation in general. At present we are integrating the two methods for selling behaviour profiling and user profiling in general.

Table 1. Results of 9 classification training runs. an asterix next to the result classification indicates a miss classification

| Subject | Test run 1 | Test run 2 | Test run 3 | Test run 4 | Test run 5 | Test run 6 | Test run 7 | Test run 8 | Test run 9 | Accuracy per Image | Accuracy per subject |
|---------|------------|------------|------------|------------|------------|------------|------------|------------|------------|----------------|---------------------|
| s10_norm | n          | n          | n          | n          | n          | n          | n          | n          | n          | 100.00        |                     |
| s10_norm | a          | a*         | a          | h          | h          | h          | h          | a*         | h          | 86.67        | 85.19              |
| s10_norm | a          | h*         | a          | a          | a          | a          | a          | a          | a          | 100.00        | 92.59              |
| s10_norm | n          | n          | n          | n          | n          | n          | a*         | n          | n          | 86.67        |                     |
| s10_norm | h          | h          | h          | h          | h*         | a          | h          | h          | h          | 100.00        | 100.00             |
| s10_norm | a          | a          | a          | a          | a          | a          | a          | a          | a          | 100.00        | 100.00             |
| s10_norm | n          | n          | n          | n          | n          | n          | n          | n          | n          | 100.00        |                     |
| s10_norm | a          | h          | a*         | h          | h          | h          | h          | h          | h          | 77.76        |                     |
| s10_norm | a          | a          | a          | a          | a          | a          | a          | a          | a          | 100.00        | 92.59              |
| s10_norm | n          | n          | n          | n          | a*         | n          | n          | n          | n          | 86.67        |                     |
| s10_norm | h          | h          | h          | h          | h          | h          | a*         | h          | h          | 86.67        |                     |
| s10_norm | a*         | h          | h          | a*         | a          | a          | a*         | a          | a          | 44.44        | 74.07              |
| s10_norm | n          | n          | n          | n          | n          | n          | n          | n          | n          | 100.00        |                     |
| s10_norm | h          | h          | h          | h          | h          | h          | h          | h          | h          | 100.00        |                     |
| s10_norm | a          | a          | a          | a          | a          | a          | a          | a          | a          | 100.00        | 100.00             |
| s10_norm | n          | n          | n          | n          | n          | n          | n          | n          | n          | 100.00        |                     |
| s10_norm | a*         | h          | a*         | a          | a          | a*         | a          | a          | a*         | 22.22         |                     |
| s10_norm | a          | a          | a          | a          | a          | a          | a          | a          | a          | 100.00        | 74.07              |

Number misclassification: 2 3 2 2 2 2 1 1
Percentage accuracy: 90.48 85.71 90.48 90.48 85.71 80.95 85.71 95.24 85.36 88.36

Overall Accuracy: 88.36

References

1. Murphy, K. A. and Shon, R. De :Progress in Psychometrics: Can Industrial and Organizational Psychology Catch Up?, Personnel Psychology vol. 53 (2000) 913-924
2. Goleman, D. Emotional Intelligence Bantam: New York (1995)
3. Goleman, D. Working with Emotional Intelligence Bantam: New York (1998)
4. Abraham, R. : The Role of Job Control as a Moderator Dissonance and Emotional Intelligence-Outcome Relationships, The Journal of Psychology Vol 134 March (2000) 169-184
5. Pellitteri, J. :The Relationship Between Emotional Intelligence and Ego Defence Mechanisms, The Journal of Psychology vol 136 March (2002) 182-194
6. Buzzotte, V.R., Lefton, R.E. and Sherberg, M.: Effective Selling Through Psychology: Psychological Associates New York (1981)
7. Anderson, R.E.: Professional Personal Selling, Prentice-Hall Inc: Englewood Cliffs New Jericy (1991)
8. Rich M. K. and Smith, D. C.: Determining Relationship Skills of Prospective Salespeople, J.of Bus.& Indus. Mkting vol. 15 (2000) 242-259
9. Khosla, R., Damiani, E.and Grosky,W. :Human-Centered E-Business, Kluwer Academic Publishers Massachusetts USA (2003)
10. Khosla, R. Goonesekera, T. and Mitsukura, T.: Knowledge Engineering of Intelligent Sales-Recruitment System Using Multi-Layered Agents Methodologies, presented at 14th International Symposium on Methodologies for Intelligent Systems (ISMIS), 28-31 October Maebashi Japan (2003)

11. Khosla, R. and Goonesekera, T.: An Online Multi-Agent e-Sales Recruitment Systems, presented at IEEE/WIC International Conference on Web Intelligence (WI) Halifax Canada (2003)

12. Szymanski, D.M.: Determinants of Selling Effectiveness: The Importance of Declarative Knowledge to the Personal Selling Concept, J. of Mkting vol. 52 (1988) 64-77

13. Weitz, B.A., Sujan, H. and Sujan, M.: Knowledge, Motivation and Adaptive Behavior: A Framework for Improving Selling Effectiveness, J. of Mkting vol. 50 (1986) 174-191

14. Saaty, T.L.: The Analytic Hierarchy Process, NY McGraw Hill (1980)

15. M.A. Turk, A.P. Pentland, “Face Recognition Using Eigenfaces”, Computer Vision and Pattern Recognition, 1991. Proceedings CVPR ’91., IEEE Computer Society Conference on, Pages:586 – 591, 3-6 June 1991.

16. J. Daugman, “Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile”, Vision Research I, vol.20, pp. 847-856, 1980.

17. J.G. Daugman, “Uncertainty Relation for Resolution in Space, Spatial, Frequency and Orientation Optimised by Two-Dimensional Visual Cortical Filters”, J. Optical Soc. Amer., vol. 2, no. 7, pp. 1,160-1,169, 1985.

18. S. Marcelja, “Mathematical Description of the responses of Simple Cortical Cells”, J. Optical Soc. Amer., vol. 70, pp. 1,297-1,300, 1980.

19. T. Lee, “Image Representation Using 2D Gabor Wavelets.”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(10):9590971, Oct. 1996.

20. S. Akamatsu, J. Gyoba, M. Kamachi, M. Lyons, “Coding facial expressions with Gabor wavelets”, Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, Pages:200 – 205, 4-16 April 1998.
Extension of Multiagent Data Mining for Distributed Databases

Ayahiko Niimi and Osamu Konishi
Department of Media Architecture, Future University-Hakodate, 116-2 Kamedanakano-cho, Hakodate 041–8655, Japan
{niimi, okonish}@fun.ac.jp

Abstract. We proposed a technique for using multiagent technology in data mining intended for two or more text databases. In this paper, we discuss data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining. First, we proposed data mining technique using multiagent technology. The proposed technique is applied to document databases, and discuss its results. Next, we extend the proposed technique with Stem algorithm, English morphological analysis, changed development language, adding the experiment data, and adding data mining algorithm.

1 Introduction

In KES2003, we proposed a technique for using multiagent technology in data mining intended for two or more text databases. [1] We applied our proposed approach to data mining from the document database, and discuss its problems. To apply proposed approach, we constructed only a minimum mounting which runs only UNIX local machine with process communications as agent communication and file system as black board model. It was confirmed to be able to switch the database and the data mining algorithm that used the constructed data mining system. We discussed data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining.

In this paper, we extend the proposed technique with Stem algorithm, English morphological analysis, changed development language, adding the experiment data, and adding data mining algorithm. The accuracy improvement of data mining intended for English document is expected by mounting of the Stem algorithm and an English morphological analysis. It becomes easy to correspond to the distributed environment by uniting the development language to Ruby. (Ruby is object-oriented script language.) It aims at the construction of the no-nonsense system by adding of data and adding mining algorithms.

The proposed extensions are applied to document databases, and discuss its results. We describe our proposed data mining that takes the multiagent technology; we describe extensions of proposed method. We construct multiagent system for data mining from document databases with these extensions, and apply to some analyzing and show the result.
Section 2 describes proposed data mining approach that uses multiagent techniques, and our proposal approach is applied to data mining from document databases. Chapter 3 describes the content added this time, and considers the mounting, the analysis, and the result. Section 4 describes conclusion and enhancing in a future.

2 Multiagent Data Mining with Databases

In KES2003, the multiagent technology is defined as a technology that processed information by cooperatively operating two or more independent programs (agent).

[1] Generally, multiagent technology is discuss with an autonomous control of an individual agent, but in this paper, we do not discuss it mainly.

A communication between agents between one to one, one to multi, multi to multi. In this paper, we use one to one communication by UNIX process communication, one to multi by Black board model.

2.1 Agent Definitions

The definition of agent which is used for data mining in this paper is defined as follows.

**Query Agent**: Query agent receives used the database and the data mining algorithm from a user, and generates other agents. Query agent is generated at each demand of a user.

**Mining Agent**: Mining agent generates DB-access agent, acquires data from DB-access agent, and applies data mining algorithm. Mining agent is generated of each applied mining algorithm.

**DB-Access Agent**: DB-access agent acquires data from the database, and sends it to mining agent. DB-access agent is generated of each database and of each mining agent.

**Result Agent**: Result agent observes a movement of mining agents, and obtains result from mining agents. When result agent obtains all results, result agent arrangement/integrates, and shows it to a user.

**Black Board(BB)**: Place where results from data mining agent is written.

2.2 Flow of System

A flow of proposed system is defined as follows. (Fig. 1 shows flowchart of proposed system.)

1. A user generates Query agent, with setting the used database and the used data mining algorithm as its parameter.
2. The place of black board(BB) is set with Query agent.
3. Query agent generates Mining agent, and the place of BB is transmitted.
4. Query agent generates Result agent, and the place of BB is transmitted.
5. DB-access agent is generated, and Mining agent is accessed to the database.
6. DB-access agent gets data from the database.
7. Mining agent receives data from DB-access agent, and applies the data mining algorithm.
8. Mining agent writes the result of data mining on BB.
9. Result agent checks BB, and if all results are written, arranges the results and presents to the user.
10. All agents are eliminated.

2.3 Feature of Proposed Method

The proposal method has the following features.

The result of data mining can be made for more meaning result by building in the thesaurus agent as Mining agent, and making it can access the thesaurus database.

Query agent generates two or more Mining agent, it becomes possible to execute data mining algorithms in parallel. Moreover, it becomes possible that constructing the system and the access to the database and the processing of data are divided by separating DB-access agent accessed the database with Mining agent that processes data.

It becomes possible that the processing of each data mining algorithm and its arrangement/integration are separately thought by setting the agent which arranges the result. Moreover, it becomes easy to build arrangement/integration according to user’s purpose into the system.

The system user comes to be able to construct the system corresponding to the purpose by recycling DB Agent and Mining Agent, and do tuning of Query agent and Result agent.

In this paper, the black board model with the file was handled with the interprocess communication on UNIX, but it can be easily enhanced to the communication on TCP/IP. Then, it is possible to enhance proposed approach to application to database that has been distributed on Internet. The problem of proposed
approach is not using interprocess communication on UNIX but using black board model. Writing in the black board becomes a problem when the number of databases and data mining algorithm used increase, then the entire operation is influenced from the operation of the slowest agent. Therefore, the access to database and the processing of the data mining algorithm can be run parallel, but processing stops when checking results in the blackboard. It is necessary to consider that the maximum time is set to the black board writing check, and the system can show the result after each agent process.

2.4 Construction of Experimental Environment

We constructed an experimental environment which has multiagents with data mining algorithms to verify our proposed approach.

The constructed experimental environment was following.

Agents were defined as each program.

The communication between programs(agents) used taking an initial option and a standard output and black board.

One file was used as Black board which has unique file name generated by process ID of query agent on UNIX system.

We used two document databases and two thesaurus databases.

These databases and thesaurus were the same as before we used. [2] Of each was divided into two for this experiment. Each document database had about 1200 papers which nonlinear problems, and each thesaurus database had about 270 words.

To confirm accessing database which was described different form, the document database was described by RDB form, and the thesaurus database was described by text file.

For data mining agents, we used frequency analysis, n-gram, association rule analysis, and using thesaurus. [3–7]

Using thesaurus was defined as getting thesaurus information from the thesaurus database.

It was confirmed to be able to switch the database and the data mining algorithm that used the constructed data mining system. A same result can be obtained from the constructed system with multiagents and the constructed system without multiagents. But if the proposed frame is used, the environment distributed on the network can be constructed. It was able to be confirmed to operate proposed approach in parallel on a local machine by the experiment.

3 Extension of Proposed Method

The technique that we proposed in KES2003 has been improved. The improved point is as follows.

- Implement of Stem algorithm
- Implement of English morphological analysis
- Re-implement system by programming language of Ruby
Data addition (1959–1975, 2002–2003)
Time series data analysis

The Stem algorithm is an algorithm that extracts an original word from the word of the conjugation form or plural form. In bibliography information registered in the document database, there is a word registered by the both singular form word and plural form word, the gerund, and the past participle are registered. In KES2003, to treat these in the same way, the filtering processing was done by manual. We resisted additional data for experimental database after KES2003. We need to do same filtering for these additional data, and then the Stem algorithm was mounted on our system.

Moreover, the English morphological analysis was built into the system. The morphological analysis program was used to take out the word of Japanese sentences last time. [8] However, we obtained the result that a specific part of speech (such as noun etc.) often became a keyword on the last experiment result. Then, to use part of speech information on the word for analysis, the morphological analysis program was applied for not only Japanese but also English.

In the last implement, the system was constructed with two or more programming languages. In this experiment, the main part of data analysis was written by the programming language "Ruby". The programming by object-oriented can be done by using Ruby, and the entire prospect of the system has improved. Moreover, it becomes easy to enhance the system to execute it on the computer that separates the

| \( t.f.idf \) | keywords                        |
|-------------|--------------------------------|
| 564.07      | chaos                          |
| 554.89      | chaos(ja)                      |
| 377.79      | neural network(ja)             |
| 359.74      | bifurcation                    |
| 318.01      | bifurcation(ja)                |
| 315.79      | simulation(ja)                 |
| 294.19      | neural network                 |
| 227.84      | circuit model(ja)              |
| 191.58      | synchronization                |
| 187.92      | bifurcation phenomenon(ja)     |
| 186.89      | numerical example(ja)          |
| 183.20      | circuit equation(ja)           |
| 180.09      | neural networks                |
| 178.41      | model(ja)                      |
| 177.17      | hysteresis                     |
| 165.24      | hysteresis(ja)                 |
| 143.39      | associative memory             |
| 140.16      | coupled oscillator             |
| 138.96      | nonlinear circuit(ja)          |
| 136.76      | fundamental equation(ja)       |
network because the environment to the network distributed programming is in order in Ruby. The result of the frequency analysis of the keyword and the result of the association analysis are shown below though there is no change in the result only because it just changes the mounting language. (See Table 1, 2)

Though it is personally unrelated to the proposal technique, at this experiment, we added new document data from 1959 to 1976 and from 2002 to 2003. We schedule to do continuously for the update of data in the future. We want to make the useful database test-head for document data mining.

We implement Mining agent which is applied data mining algorithm every year with specific keyword as an analysis of data. A time flow of the specific keyword can be examined by using to analyze this Mining agent. It can be assumed the clue that analyzes the research trend of the specific research field by analyzing the symbolized keyword from the document database. The result of the analysis specific keyword of "chaos" as an analysis example is shown below. (See Table 3. The document database is a nonlinear problem society’s.)

| sup, con f (%) | keywords |
|----------------|----------|
| (1.0, 83.4)   | neural_network, optimization, problems, combinatorial |
| (1.2, 83.5)   | neural_network, memory, associative |
| (1.6, 91.7)   | neural_system, network, dynamical |
| (1.0, 83.6)   | neural_networks, problems, combinatorial |
| (1.0, 83.6)   | networks, optimization, problems, combinatorial |
| (1.5, 84.4)   | chaos, system, linear, piecewise |
| (1.0, 78.9)   | neural, optimization, problems, combinatorial |
| (1.2, 96.1)   | neural_networks, optimization, problems |
| (1.2, 97.4)   | neural_networks, optimization, combinatorial |
| (1.2, 86.7)   | chaos, control, feedback, delayed |
| (1.2, 80.6)   | network, memory, associative |
| (1.4, 49.9)   | neural_network, associative |
| (1.2, 58.9)   | bifurcation, nonlinear, circuit |
| (1.5, 32.4)   | neural_system, network |
| (1.1, 58.5)   | neural_network, model |
| (1.3, 86.1)   | control, feedback, delayed |
| (1.0, 41.5)   | neural, problem, optimization |
| (1.2, 59.6)   | chaos, system, dynamical |
| (1.2, 62.4)   | neural_networks, problems |
| (1.9, 60.5)   | chaos, system, piecewise |

Table 2. Result of the Association Analysis (Top20)

4 Conclusion

In KES2003, we proposed a technique for using multiagent technology in data mining intended for two or more text databases. In this paper, we extend the proposed technique with Stem algorithm, English morphological analysis, changed development language, adding the experiment data, and adding data mining algorithm. We propose data mining technique using multiagent technology. The proposed technique was applied to document databases, and discussed results. We describe our proposed data mining that takes the multi agent technology; we describe extensions of proposed method. We construct multiagent system for data mining from
document databases with these extensions, and apply to some analyzing and show the result.

**Table 3. Frequency of Keyword each year (chaos)**

| Year | count |
|------|-------|
| 1992 | 107   |
| 1993 | 99    |
| 1994 | 160   |
| 1995 | 153   |
| 1996 | 161   |
| 1997 | 198   |
| 1998 | 130   |
| 1999 | 186   |
| 2000 | 161   |
| 2001 | 154   |

To apply proposed approach, we construct only a minimum mounting which runs only UNIX local machine with process communications as agent communication and file system as black board model. It was confirmed to be able to switch the database and the data mining algorithm that used the constructed data mining system. We will enhance to make the agent which processes data mining can executed on the computer that separates the network, and to consider the load-balancing of the data mining processing in the future. And, we will discuss other data mining algorithms and other type of databases. The system implement to use distributed Ruby(dRuby) is doing for distributed environment.

**References**

1. Niimi, A., Konishi, O.: Data Mining for Distributed Databases with Multiagents. KES’2003, Proceedings, PartII, Lecture Notes in Artificial Intelligence 2774, Springer:pp.1412–1418 (2003)
2. Niimi, A.: Research Trend Investigation from Bibliographic Database about Chaos using Data Mining Technique, Technical report of IEICE, AI20002-57, pp.59–64 (2003). (In Japanese)
3. Ichimura, Y., Hasegawa, T., Watanabe, I., Sato, M.: Text Mining: Case Studies, Journal of Japanese Society for Artificial Intelligence, Vol.16 No.2,pp.192–200 (2001). (In Japanese)
4. Nasukawa, T., Kawano, H., Arimura, H.: Base Technology for Text Mining, Journal of Japanese Society for Artificial Intelligence, Vol.16,No.2,pp.201–211 (2001). (In Japanese)
5. Nagao, M., Mori, S.: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, In Proceedings of the 15th International Conference on Computational Linguistics pp.611–615 (1994).
6. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, September 1994:32pages (1994).

7. Nagata, M., Taira, H.: Text Classification - Showcase of Learning Theories -, IPSJ Magazine, Vol.42 No.1 pp.32–37 (2001). (In Japanese)

8. Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara, M.: Morphological Analysis System ChaSen version 2.2.1 Manual (2000). [Online] Available: http://chasen.aist-nara.ac.jp/chasen/bib.html.en
Agent-Based Approach to Conference Information Management

Hee-Seop Han¹, Jae-Bong Kim², Sun-Gwan Han², and Hyeoncheol Kim¹

¹Korea University
Department Of Computer Science Education,
College of Education,
Anam-dong Sungbuk-ku, Seoul, 136-701, Korea
anemon@korea.com

²Gyeongin National University of Education
Department Of Computer Education
45, Gyodae Street Gyeyang-ku, Incheon, 407-753, Korea
han@gin.ac.kr

Abstract. Agent-based conference information management systems are proposed. Three different types of agent-based systems are introduced: single agent, multi agent and mobile agent. They manage the information intelligently and autonomously in Semantic Web using RDF and DAML.

1 Introduction

Researchers want to locate proper conferences to present their contributions and to share their ideas. Unfortunately it is difficult for them to search right conferences on right time since there is so much conference information in the web and the information is subject to change anytime and not well-organized. The other side, the conference organizers want to attract researchers as many as possible. It is not easy for them to search target researchers and let them know their conference information on right time. Without effective agents between the two sides, researchers might not meet deadlines and conference organizers can not make the conference more successful.

To solve the problems, intelligent software agents and semantic web can be used. Information of conference sites are represented in standardized Semantic Web. The agents are placed between a conference management system and a user computer. The user agent and the conference agent exchanges useful information each other. Through this process, the agents can provide researchers proper conference information and schedule, and thus conference organizers can make their conferences more successful.

In this article, we propose three types of agent-based systems. First model is the single agent-based system in HTML environment. Second Model is the multi agent-based system using DAML(DARPA Agent Markup Language) and RDF-based Semantic Web. Third model is the mobile agent-based system in wireless and ubiquitous computing environment.
2 Semantic Web

The Semantic Web is the representation of data, which is based on the Resource Description Framework (RDF), on the WWW. The RDF integrates a variety of applications using XML for syntax and URLs for naming. The Semantic Web is a mesh of information linked up in such a way as to be easily process by machines, on a global scale.

The Semantic Web is composed of several layers. The XML layer represents data and the RDF layer represents the meaning of data. The Ontology layer represents the formal common agreement about meaning of data finally the Logic layer enables intelligent reasoning with meaningful data. [10].

The RDF is a model of entities and relationships, providing a means for adding semantics to a document. The information of data is stored in the form of RDF statements, which are machine understandable. It is be also understandable to intelligent agents, search engines, browsers and human users. RDF is implementation independent and may be serialized in XML [7,9].

Ontology provides a very powerful way to describe objects and their relationships to other objects. The DAML language is being developed as an extension to XML and RDF. The latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontology and to markup information so that it is machine readable and understandable [5].

3 Single Agent-Based Conference Management System (SACS)

3.1 Overview

The SACS is based on existing HTML web pages of conference site. The system uses single agent to search and parse conference sites, and provides proper information to users. Figure 1 illustrates overview of the system.

![Fig. 1. Overall of Single Agent-based System](image-url)
The SACS converts the conference information into RDF. The RDF is then stored at the conference KB (Knowledge Base). User profiles and requests are stored in user DB (Data Base). Inference module recommends the user suitable information which is inferred from user’s requests. Interface module provides an I/O interface between user and the inference engine.

### 3.2 Design of Conference RDF

Figure 2 shows the relation graph of key terms and information in conference sites. This analysis is necessary for constructing the semantic web.

**Fig. 2. Basic keywords in conference site**

With the graphical analysis of meaning and relationship of key words in conference site, we build RDF. The following is an example of RDF from Figure 2.

```xml
<continent rdf:ID="Europe"/>

<rdf:Description rdf:ID="Europe">
  <rdf:type>
    <rdfs:Class rdf:about="#continent"/>
  </rdf:type>
</rdf:Description>

<rdf:Description rdf:ID="Spain">
  <is_part_of rdf:resource="#Europe"/>
</rdf:Description>
```

User agent sends a query in DAML to the conference agent. The conference agent returns the proper reply if any. Otherwise, the agent returns result by inference to the user. For example, consider a case that a user wants to participate in a conference held in France when the conference is not planned in France, but in Spain. The agent
recommends the conference in Spain instead of giving “NO” message. Based on the RDF, the system can provide better results to users.

4 Multi-agent-Based Conference Management Systems

4.1 Overall of Multi Agent-Based Systems

Multi agent-based system uses facilitator to connect different agents. Not like RETSINA[8], our system is an model built in Web environment. As illustrated in Figure 3, if conference sites are constructed Semantic Web using XML, RDF, ontology and etc, we can make our agents more intelligent. Conference agents in conference sites and user agents in users communicate intelligently and autonomously.

![Fig. 3. Overall of Multi Agent-based Systems](image)

The multi-agent-based conference management system consists of user agents, conference agents, and a facilitator agent. The facilitator agent is located between users and conference sites and plays a role of mediator that inter-communicates basic information. The conference information includes conference title, the topics and URL, and the user information includes user’s research field and user agent ID.

The user agent contains an inference engine and user profile. The user profile includes user’s basic data and user’s requests for conference. If user agent requests the facilitator any conference information, it returns the registered conference information. The user agent searches conference information using the conference site URLs.
In this phase, the user agent reports to the users with the best conference site using user’s requests and profiles.

The conference site is built on RDF defined in previous section. Conference agent manages the conference information and replies a user agent’s request. A communication between a user agent and conference agent uses DAML [5,11]. Conference agent requests facilitator agent the user’s information related their conference. With the user information acquired, conference agent can announce their conference information and encourage the participation of users. Also, if conference information is changed, agent broadcast all registered users.

### 4.2 Intercommunication Between Agents Using DAML

In this system, user agent collects the conference information as follows. User agent connects to facilitator and receives the conference agent URLs. Then it sends a query to the conference agents in parallel. The conference agent sends “SUCCESS” message and basic conference information to the user if there is any matched. Otherwise it returns “FAIL” message. If the facilitator finds the suitable URL of conference, the facilitator gives the conference information to user agent. Next the user agent communicates the conference agents directly. Therefore, this process can decrease server traffic and overloads.

For example, consider the case that a user queries fall for conference. If conference date field value is not fall but autumn, traditional system will respond “FAIL” message. If we define it with RDF and DAML, this system will respond “SUCCESS” message by inference. With the RDF, it can infer autumn as well as September or October from the fall. The example can be expressed as follows.

```xml
<daml:Class rdf:ID="September">
  <rdfs:subClassOf rdf:resource="#Fall"/>
</daml:Class>
<daml:Class rdf:ID="October">
  <rdfs:subClassOf rdf:resource="#Fall"/>
</daml:Class>
<daml:Class rdf:ID="Autumn">
  <rdfs:label>Fall season</rdfs:label>
  <daml:sameClassAs rdf:resource="#Fall"/>
</daml:Class>
```

### 5 Mobile Agent-Based Conference Management Systems

Mobile agent-based system needs neither servers nor a mediate agent. As illustrated in Figure 4, the mobile agent moves to conference site directly, collects the conference information, and provides proper information requested by users. Also, the conference manager collects user information in the mobile agent, and then manages and utilizes it. The suggested system architecture can be used directly in mobile device in wireless communication or ubiquitous computing environment. Security and privacy problems, however, should to be solved.
6 Conclusion

In this study, we suggest three different types of agent-based systems for conference management. Single agent-based system searches conference sites in HTML and then converts the information into RDF format. Multi agent-based system runs in semantic web that makes the agents more effective. Finally, mobile agent-based systems can be used in a wireless and ubiquitous computing environment.

We used RDF to express the information of conference sites in Semantic web, and applied DAML to communicate with each agent. The DAML can infer the query in communication language, and can supply users with better results.

With the agent-based systems, agents help conference organizers to search target users and to inform them right information on right time. Also agents help users to search conferences of their interests efficiently and be noticed with correct information even if it is subject to change. Thus, the users don’t need to waste their effort and time to manage and update their schedule for participation and presentation, and the conference organizers can encourage participations of researchers and papers of good quality.

References

1. Danny B. Lange, M. Oshima.: Programming and Deploying java Mobile Agents with Aglets, Addison-Wesley (2001)
2. David L. Martin, Adam J. Cheyer, Douglas B. Moran.: The Open Agent Architecture: A Framework for Building Distributed Software Systems (1999)
3. H.C. Wong and K. Sycara.: A Taxonomy of Middle-agents for the Internet. In ICMAS’2000 (2000)
4. Hyacinth S. Nwana, Divine T. Ndumu.: A Perspective on Software Agents Research (1999)
5. J. Hendler and D. L. McGuinness.: Darpa agent markup language. IEEE Intelligent Systems, 15(6) (2001) 72–73
6. M. Wooldridge and N. R. Jennings.: Intelligent agents: Theory and practice, The Knowledge Engineering Review, 10(2) (1995) 115-152
7. S. McIlraith, T. C. Son, and H. Zeng.: Semantic web service. IEEE Intelligent Systems, 16(2) (2001) 46–53
8. Sycara, K., Paolucci, M., van Velsen, M. and Giampapa, J.: The RETSINA MAS Infrastructure. To appear in the special joint issue of Autonomous Agents and MAS, Volume 7, Nos. 1 and 2, July (2003)
9. T. Berners-Lee.: What the semantic web can represent, http://www.w3.org/DesignIssues/RDFnot.html (2000)
10. T. Berners-Lee, J. Hendler, and O. Lassila.: The semantic web. Scientific American, 284(5) (2001) 34–43
11. T. Finin, Y. Labrou, and J. Mayfield.: KQML as an agent communication language. In J. Bradshaw, editor, Software Agents. MIT Press, Cambridge (1997)
Mining Frequency Pattern from Mobile Users

John Goh and David Taniar
Monash University, School of Business Systems, Clayton, Vic 3800, Australia
{Jen.Ye.Goh, David.Taniar}@infotech.monash.edu.au

Abstract. Group pattern was introduced to find groups of mobile users associated by means of physical distance and amount of time spent together. This paper addresses the inherent problem of group pattern, that mobile user are often not physically close together when they use mobile technology, by proposing frequency pattern. Frequency pattern use creative method to calculate frequency of communication between mobile users. By using frequency rather than physical distance, the closeness of two mobile users can better be represented. Performance of the proposed method indicates a suitable segment size and alpha value needs to be selected to get the best result.

1 Introduction

Data Mining is an area of research that focuses on finding out useful knowledge from a set of data sources [1]. The data sources can come from time series data, spatial based data, transaction based data and many more. The goal is find useful knowledge suitable for decision makers. Within data mining, a branch of mobile data mining can be extended based on the special conditions of mobile environment. In a mobile environment, there are nodes, which can be cars, mobile phones, PDAs. These nodes are often equipped with limited resources, such as limited battery life, memory capacity and processing capacity [2].

Group pattern [4] represents a group of users which stays within a maximum level of distance magnitude over a minimum amount of time. For example, a group of mobile nodes which is together over a short distance and over a specified time would be considered as a group. Therefore, a group pattern represents a group of mobile users which are related together in terms of distance and time. [4]

Group pattern has its unique characteristics independent from association rules [1], as association rules are transaction based while group pattern is based on dynamic time series. Group pattern is also distinct from clustering [3] because clustering is based on grouping entities that have similar characteristics and features, while group pattern is based on the movement data of mobile users which are dynamic as the time and distance changes frequently. Group pattern uses Euclidean distance [4] as a method to determine whether to include a set of mobile nodes into a group, forming a group pattern. Unfortunately, the purpose of mobile applications is to perform tasks without distance limitation.
2 Related Work

There are relatively few papers written specifically for methods in finding useful knowing in the mobile environment, also known as mobile mining. One of the related work that aims to finding useful knowing of groups of mobile users was presented in [4]. The input contains a database of raw data which consists of mobile user identification, and their two dimensional location data over a time series. The process of the data mining involves determining groups of users that stays together over two variables (time and distance).

The aim of this process is to provide a list of groups of mobile users, which during the time series has a magnitude of distance lesser than the distance threshold, and has a magnitude of time greater than the time threshold. In another words, if there are a few mobile users that are close to each other over a certain time, it is then considered as a group pattern [4].

The advantage of group pattern is that it is only suitable in finding mobile user knowledge in an environment that is restricted, that is, mobile users moves within a closely monitored region. This is because of the process involves finding out the x and y axis of the location of the mobile user data, which can be very expensive in real life.

The disadvantage of group pattern is that, there is an inherent problem of the entire process that is the method was used in mobile environment. A mobile environment is an environment that mobile nodes move freely over a large geographical distance. The purpose of mobile equipment usage is to break the barrier of distance, enabling users to interact with others without needing to get physically close to each other.

Group pattern is only useful when mobile users are actually close to each other, as a group is a group of mobile user that is close to each other over time. This, in real life can rarely happens, as if two people are physically close to each other, they don’t need to use mobile devices to communicate to each other. Furthermore, in real life, it may return a lot of noises (wrong result), especially mobile nodes are physically close to each other in public places, such as bus stops, airport terminals, while they are waiting for something to happen. The above example can easily qualified as a group pattern, as a group of mobile users are close to each other over a certain time. But these mobile users may not really have much interaction with each other when they are close physically. Rather they might be busy sending communications using their mobile equipment to people in another country.

3 Frequency Pattern: Proposed Method

In the area of data mining, there is mobile mining which focuses on the goal of finding out useful knowing based on raw data of mobile users. In the area of mobile mining, the proposed method is called frequency pattern meets the goal of mobile mining of finding useful knowledge from raw data of mobile users by means of flexible method for calculating a frequency of communication between two mobile nodes and determining their logical proximity and returns useful knowledge of sets of mobile users that are close to each other.
Frequency pattern is discovered by calculating the relative frequency between two mobile nodes. First, the pre-specified criteria, such as the number of segments, the \textit{alpha} values and the size of the time series to be considered are provided. Then, a relative frequency is calculated. In order to determine whether the relative frequency is strong enough to be generated as knowledge, a threshold has to be set. If the relative frequency is stronger than the threshold, then it will be accepted. The final outcome is a knowledge, which contains mobile nodes and their strong relative relationships among each others.

Let the time series in consideration be represented as \{t_1, t_2, \ldots, t_{10}\} where each item represents an unit of timeslot. Let the segments be represented in the manner of segment 1 \{t_1 \text{ to } t_3\}, \alpha=0.2, which represents segment number 1 ranges from timeslot \(t_1\) to timeslot \(t_3\) inclusive, with an \textit{alpha} value of 0.2. The addition of all \textit{alpha} values in all segments must equate to 1. In another words, for the final relative frequency value, the \textit{alpha} value for each segment represents the amount of emphasis is given for a particular timeslots in a particular segment. All the conditions mentioned are called the \textit{pre-specified criteria}. The \textit{pre-specified criteria} is important because it enables the frequency calculation to be tailored to different data mining situations by providing different emphasis on different segments of the time series.

The calculation of relative frequency between two mobile nodes is non-directional. This means that a 0.4 value of frequency from mobile node 1 to mobile node 2, and from mobile node 2 to mobile node 1 are the same. The relative frequency between two mobile nodes is calculated in such a way that it takes on different emphasis on different segment in the time series based on the \textit{alpha} value. If relative frequency between mobile node \(A\) and mobile node \(B\) is equal to 0.4, this means that within the value of 0.4, 20\% of the value is supported from the average frequency counted in segment 1, 30\% of the value is supported from the average frequency counted in segment 2, and 50\% of the value is supported from the average frequency counted in segment 3, since \(\alpha_1=0.2, \alpha_2=0.3, \alpha_3=0.5\).

The concept of relative frequency is best explained by means of an example. In this example, the time series ranges from \(t_1\) to \(t_{10}\), which represents the raw mobile data collected between mobile node \(A\) and mobile node \(B\). The sample of communication collected between two mobile nodes in a mobile environment consists of a sequential order of timeslots, from \(t_1\) to \(t_n\). Each timeslot represents a binary value of either 0 or 1. By having the pre-specified criteria of 3 segments, where segment 1 ranges from \(t_1\) to \(t_3\) with \textit{alpha} value 0.2, segment 2 ranges from \(t_4\) to \(t_6\) with \textit{alpha} value 0.3 and segment 3 ranges from \(t_7\) to \(t_{10}\) with \textit{alpha} value 0.5. The relative frequency between mobile node \(A\) and mobile node \(B\) can be presented as an equal:

\[
\text{Relative Frequency} = \text{Average(Segment 1)} \times \alpha_1 + \ldots + \text{Average(Segment n)} \times \alpha_n
\]

For the sample data above, a result of relative frequency of 0.93 is obtained. In reality, a frequency of 0.93 would be considered highly logically proximate. The goal of data mining, which is to find relevant and accurate knowledge from raw data, when applied into the mobile environment, would require the time series to be extended to a reasonable amount of length.
After all the relative frequencies between mobile nodes are counted, a sample output would be represented a tabular form represented by Table 2 below. Note that there is no relative frequency within the node itself, and relative frequency is non-directional, thus relative frequency between node A and node B is equal to relative frequency between node B and node A.

| Relative Frequency | Node A | Node B | Node C | Node D |
|-------------------|--------|--------|--------|--------|
| Node A            | -      | 0.9    | 0.2    | 0.5    |
| Node B            | 0.9    | -      | 0.5    | 0.8    |
| Node C            | 0.2    | 0.5    | -      | 0.7    |
| Node D            | 0.5    | 0.8    | 0.7    | -      |

Finally, in order to show the knowledge from the raw data, a frequency threshold (β) has to be set. The frequency threshold represents the degree of relative frequency that a decision maker is willing to accept in order to recognise two mobile nodes to be logically close to each other. Therefore, within all the relative frequency calculated those which are lesser than β will be treated as not logically close. Therefore, the output is a table or graph where relative frequency ≥ β. From the sample output above, with β = 0.5, there are 7 recognised pair of mobile node.

Table 2. Sample Output

| Relative Frequency | Node A | Node B | Node C | Node D |
|-------------------|--------|--------|--------|--------|
| Node A            | -      | 0.9    | -      | 0.5    |
| Node B            | 0.9    | -      | 0.5    | 0.8    |
| Node C            | 0.5    | -      | -      | 0.7    |
| Node D            | 0.5    | 0.8    | 0.7    | -      |

Graphically, the knowledge is represented in Figure 1. In Figure 1, there are two set of output generated. The left hand side graph represented the output with a beta value of 0.5 while the right hand side graph represents the output with a beta value of 0.6. The purpose of representing the output in two different graphs is to show the result of increasing the beta value.

Here, the knowledge found is that mobile node A is significantly (0.9) related to mobile node B. There is also a strong relationship between mobile node B and mobile node D (0.8). Finally, there is also a good relationship between mobile node D and mobile node C (0.7). We can observe that there is no significant relationship between: node A and node C, node A and node D, node B and node C. In different circumstances, the inverse of the output can be used when there is significant relationship between most mobile nodes while the decision maker is interested in knowing finding mobile nodes that are not logically close thus saving resources.
Fig. 1. Output with Beta = 0.5 (Left), Output with Beta = 0.6 (Right)

The algorithm for frequency pattern generation is presented in Figure 2. The frequency generator computes the relative frequency between two mobile nodes with the pre-specified criteria into consideration. The display knowledge part checks and displays only mobile nodes having relative frequency greater than threshold.

```plaintext
Function Frequency_Generator (Node A, Node B) {
    Data Structure Representing Pre-Specified Criteria
    Define Configuration As Array Of {
        Array of Segment As Integer
        Array of Segment Size As Integer
        Array of Segment Alpha Value As Float
    }
    Define Freq As Float, I As Integer
    Set Freq To 0 // Each Relative Frequency = Average of Frequency Count * Alpha Value
    For I = 1 To I = Number of Segments
        Freq = Freq + [Average(Segment I) * Alpha(Segment I)]
}
Function Display_Knowledge(Table of Nodes) {
    For I = 1 to I = (No of Mobile Nodes)
        For J = 1 to J = (No of Mobile Nodes)
            If Frequency > Threshold Then Display Mobile Node I-J
}
```

Fig. 2. Frequency Generator Algorithm

4 Performance Evaluation

Performance evaluation is performed by measuring how accurate the proposed solution can determine three different sets of data which contains different kind of frequency distribution. The three sets of data consists of Set A, Set B and Set C. Set A consists of a distribution of frequency skewed to the left. Set B consists of a near
normal distribution of frequency. Set C consists of a skewed right distribution of frequency. The sample data sets are first generated, and the formula entered to calculate the relative frequency based on the pre-specified criteria which consist of segments and alpha values.

Figure 5 below represents the impact of increasing emphasis on recent communication to the relative frequency. The figure shows that as there is increasing emphasis on recent communication, Set C, the set of data which has more recent communication tends to have higher relative frequency, which will be regarded as having more significant logical relationship.

Figure 3 represents the performance result of increasing emphasis on recent and historical data. The higher the window, the longer the decimal places required. From the performance data, we can observe clearly that Set C responded strongly to recent emphasis and Set A responded strongly to emphasis on historical data.

![Fig. 3. Performance Data for Recent & Historical Emphasis](image)

Figure 4 below represents the impact of decreasing standard deviation of emphasis on relative frequency. The decreasing standard deviation also means that more emphasis will be placed on the middle region of the pre-specified criteria. As Set B contains a normally distribution frequency distribution, it stays within the average region. Set A, however only achieved a slight increase while Set C decreases significantly.

![Fig. 4. Performance Data for Decreasing Standard Deviation](image)
5 Conclusion and Future Work

The proposed frequency pattern [4] is more relevant in mining knowledge out of mobile users in real life. This is because the frequency pattern focuses on the logical communication as a mean to determine whether two mobile nodes are in fact close to each other logically rather than using physical distance which may be more appropriate in a non-mobile environment. The frequency pattern uses the concept of segments and \( \alpha \) values in order to allow decision makers to place different level of emphasis during different parts of the time series, such as during peak hour, or during hours that should be neglected such as out of office hour.

The performance of the proposed frequency pattern [4] method significantly reacts to the data set. It can further be tailored into different environment by dynamically changing the size of the segment, the \( \alpha \) value and the number of segments. The future work of this proposed paper is to looking at the relative relationship between mobile nodes in one further step. It may be possible to find out the relationship between two mobile nodes through an intermediate node.

References

1. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proc. 20th Int. Conf. Very Large Data Bases, 1994.
2. E.-P. Lim, et al. In Search Of Knowledge About Mobile Users. 2003.
3. R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. In 20th International Conference on Very Large Data Bases, September 12--15, 1994, Santiago, Chile proceedings, 1994.
4. Y. Wang, E.-P. Lim, and S.-Y. Hwang. On Mining Group Patterns of Mobile Users. Lecture Notes in Computer Science, vol. 2736, pp. 287-296, 2003.
Semi-supervised Learning from Unbalanced Labeled Data – An Improvement

Te Ming Huang and Vojislav Kecman
School of Engineering, The University of Auckland, Auckland, New Zealand
v.kecman@auckland.ac.nz, huangjh@win.co.nz

Abstract. We present a possibly great improvement while performing semi-supervised learning tasks from training data sets when only a small fraction of the data pairs is labeled. In particular, we propose a novel decision strategy based on normalized model outputs. The paper compares performances of two popular semi-supervised approaches (Consistency Method and Harmonic Gaussian Model) on the unbalanced and balanced labeled data by using normalization of the models’ outputs and without it. Experiments on text categorization problems suggest significant improvements in classification performances for models that use normalized outputs as a basis for final decision.

1 Introduction

Today, there are many learning from data paradigms, the most popular and the most used ones being classification and regression models [2]. They belong to the so-called supervised learning algorithms in which a learning machine attempts to learn the input-output relationship (dependency or function) $f(x)$ by using a training data set $X = \{[x(i), y(i)] \in \mathbb{R}^m \times \mathbb{R}, i = 1,...,n\}$ consisting of $n$ pairs $(x_1, y_1), (x_2, y_2),..., (x_n, y_n)$, where the inputs $x_i$ are $m$-dimensional vectors $x \in \mathbb{R}^m$ and the labels (or system responses) $y \in \mathbb{R}$ are continuous values for regression tasks and discrete (e.g., Boolean) for classification problems. Another large group of standard learning algorithms are the ones dubbed as unsupervised ones when there are only raw data $x_i \in \mathbb{R}^m$ without the corresponding labels $y_i$ (i.e., there is a ‘no-teacher’ in a shape of labels). The most popular, representative, algorithms belonging to this group are various clustering and (principal or independent) component analysis routines.

Recently, however, we are facing more and more instances in which the learning problems are characterized by the presence of (usually) a small percentage of labeled data only. In this novel setting, the task is to predict the labels (or the belonging to some class) of the unlabeled data points. This learning task belongs to the so-called semi-supervised or transductive inference problems. The cause for an appearance of the unlabeled data points is usually expensive, difficult and slow process of obtaining labeled data. Thus, labeling brings the costs and often it is not feasible. The typical areas where this happens is the speech processing (due to the slow transcription), text categorization (due to huge number of documents, slow reading by humans and their
general lack of a capacity for a concentrated reading activity), web categorization, and, finally, a bioinformatics area where it is usually both expensive and slow to label huge number of data produced.

Recently several approaches to the semi-supervised learning were proposed. Here, we present, compare and improve the two transductive approaches, namely, the harmonic Gaussian model introduced in [6] and consistency method for semi-supervised learning proposed in [5].

However, none of the methods successfully analyzes the possible problems connected with the so-called unbalanced labeled data, meaning with the situations when the number of labeled data differs very much between the classes. We propose the normalization of the classifier outputs before a final decision about the labeling is done.

Paper is organized as follows: In section 2 we present the basic forms of the two methods. Section 3 introduces the normalization step which improves the performance of both the consistency method and the harmonic Gaussian model faced with unbalanced labeling significantly. It also compares the effects of normalization with the results of both methods obtained and presented in [5]. Section 4 concludes the presentations here and proposes possible avenues for the further research in this novel area of semi-supervised learning.

2 Consistency Method Algorithm and Harmonic Gaussian Model

There exist a great variety of methods and approaches in semi-supervised learning. The powerful software SemiL for solving semi-supervised (transductive) problems, used within this study, is capable of using 12 different models for a semi-supervised learning (as suggested in [4]). Namely, it can solve the following variously shaped semi-supervised learning algorithms: both the hard label approach with the maximization of smoothness and the soft label approach with the maximization of smoothness, for all three types of models (i.e., Basic Model, Norm Constrained Model and Bound Constrained Model) and by using either Standard or Normalized Laplacian. Presenting all the variety of results would require much bigger space than it is allowed within the constrained space allotted here. That’s why the presentation here will be focused on two basic models only, and on an introduction of a normalization step as the first possible significant stage in improving results to date.

Below we present Global consistency model from [5] which is a soft label approach with the maximization of smoothness that uses a normalized Laplacian without a norm constraint, as well as the Harmonic Gaussian method presented in [6] which is a hard label approach with the maximization of smoothness that uses a standard Laplacian also without a norm constraint.

2.1 Global consistency Model

The presentation here follows the basic model proposed in [5] tightly.

Given a point set $X$ as defined in the Introduction the first $l$ points $x_i$ are labeled, and the remaining points $x_u$ ($l+1 \leq u \leq n$) are unlabeled. The goal is to predict the label of the unlabeled points.
Let $F$ denote the set of $n \times c$ matrices with nonnegative entries. A matrix $F = [F_1^T, \ldots, F_n^T]^T \in F$ corresponds to a classification on the dataset $X$ by labeling each point $x_i$ as a label $y_i = \arg \max_{j \leq c} F_{ij}$. We can understand $F$ as a vectorial function $F: X \rightarrow \mathbb{R}^c$ which assigns a vector $F_i$ to each point $x_i$. Define an $n \times c$ matrix $Y \in F$ with $Y_{ij} = 1$ if $x_i$ is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. Clearly, $Y$ is consistent with the initial labels according the decision rule. The algorithm is as follows:

1. Form the affinity matrix $W$ defined by $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2. Construct the matrix $S = D^{-1/2}WD^{-1/2}$ in which $D$ is a diagonal matrix with its $(i, i)$-element equal to the sum of the $i$-th row of $W$.
3. Iterate $F(t+1) = SF(t) + (1 - \alpha)Y$ until convergence, where $\alpha$ is a parameter in $(0, 1)$.
4. Let $F^*$ denotes the limit of the sequence $\{F(t)\}$. Label each point $x_i$ as a label $y_i = \arg \max_{j \leq c} F_{ij}^*$. 

First, one calculates a pairwise relationship $W$ on the dataset $X$ with the diagonal elements being zero. In doing this, one can think of a graph $G = (V, E)$ defined on $X$, where the vertex set $V$ is just $X$ and the edges $E$ are weighted by $W$. In the second step, the weight matrix $W$ of $G$ is normalized symmetrically, which is necessary for the convergence of the following iteration. The first two steps are exactly the same as in spectral clustering [3]. Here, we did not solve the problem in an iterative way as shown above. Instead, we solve the corresponding equivalent system of linear equations $(I - \alpha S) F^* = Y$ for $F^*$ by using conjugate gradient method which is highly recommended approach for dealing with huge data set. Also, instead of using the complete graph we calculated the $W$ matrix by using only 10 nearest neighbors. This step decreases the accuracy only slightly, but it increases the calculation speed significantly. Note that self-reinforcement is avoided since the diagonal elements of the affinity matrix are set to zero in the first step ($W_{ij} = 0$). The model labels each unlabeled point and assigns it to the class for which the corresponding $F^*$ value is the biggest, as given in step 4 above.

### 2.2 Harmonic Gaussian Model

The presentation here also follows the basic model proposed in [6] tightly. The algorithm is as follows:

1. Form the affinity matrix $W$ defined by $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$.
2. Construct the diagonal matrix $D$ with its $(i, i)$-element equal to the sum of the $i$-th row of $W$. Note that we can use $W$ and $D$ as given in section 2.1 above too.
3. Form the following two matrices $W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$, $D = \begin{bmatrix} D_{ll} & 0 \\ 0 & D_{uu} \end{bmatrix}$ as well as the vector $f = [f_l \ f_u]^T$, where $l$ stands for the labeled data points and $u$ for the unlabeled ones.
4. Solve for $f_u$ as follows $f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$ which is the solution for the unlabeled data points.
More detailed description of the two basic models, namely, the global consistency model and the harmonic Gaussian model can be found in [5] and [6] respectively.

3 Performance of the Two Models and Possible Improvement

The extensive simulations on various data sets (as presented in [5]) have indicated that both models behave similarly and according to the expectations that with an increase in the number of labeled data points $l$, the overall models’ accuracies improve too. There was just a slightly more superior performance of the consistency model from [5] in respect to the harmonic Gaussian model, when faced with a small number of unbalanced labeled data. At the same time, the later model performed much better for extremely small number of the labeled data as long as they are balanced (meaning there is a same number of the labeled points for all the classes. Here, an extremely small number meant 1 labeled data per each class only, in the text categorization problem from [5]).

Such a behavior needed a correct explanation and it asked for further investigations during which several phenomena have been observed. While working with balanced labeled data (meaning with the same number of labeled data per class) harmonic Gaussian method performed better than the consistency model. On the contrary, for a small number of unbalanced labeled data, the harmonic Gaussian model performed worse than the consistency one. This indicates a sensitivity of the former while working with the unbalanced labeled data.

At the same time a simulation shows that in the harmonic Gaussian method the mean value of the class with less labeled points is lower than for the classes with more labeled data. Recall that the final decision is made based on the maximum of the $F^*$ values and obviously the elements of the class with less labeled data could be assigned to different class just due to the fact that the (mean) values of other classes are higher.

This led us to the introduction of a normalization step for the elements of the column vectors $F^*$, bringing them to the vectors with a mean $= 0$, and with a standard deviation $= 1$. Only now, after the normalization is performed, the algorithm searches for the maximal value along the rows of a matrix $F^*$ and labels the unlabeled $i$-th data to the class $j$ if $F^*_{ij} > F^*_{ik}$, $k = 1, c, k \neq j$.

The introduction of the normalization step improves the behavior of the algorithm significantly as it is shown in Fig. 1, where we compare performances of the two models without normalization as given in [5] to the performances of both models incorporating a normalization part.

Same as in [5], in the experiment here, we investigated the task of text classification using the 20-newsgroups dataset. The chosen topic was rec which contains autos, motorcycles, baseball, and hockey from the version 20-news-18828. The articles were processed by the Rainbow software package with the following options: (1) passing all words through the Porter stemmer before counting them; (2) tossing out any token which is on the stop list of the SMART system; (3) skipping any headers; (4) ignoring words that occur in 5 or fewer documents. No further preprocessing was done. Removing the empty documents, we obtained 3970 document vectors in a 8014-dimensional space. Finally the documents were normalized into TFIDF representa-
tion. The cosine distance between points was used here too. The mentioned procedure is the same as in [5] just in order to ensure the same experiment’s setting for same data set.

We played with various widths of the Gaussian RBF and the results with a few \( \sigma \)-s are shown in Fig. 1. The results in [5] use \( \sigma = 0.15 \) for both harmonic Gaussian method and consistency method. The test errors shown are averaged over 100 trials. Samples were chosen so that they contain at least one labeled point for each class. Thus, the setting of the experiment is identical to the one in [5].

![Graph showing error rates of text classification](image)

**Fig. 1.** The error rates of text classification with 3970 document vectors in an 8014-dimensional space for recreation data sets from version 20-news-18828. At least one labeled data for each class must be labeled. The smallest number of labeled data here is therefore 4. The normalized model outputs outperform the algorithms without normalization.

Several interesting phenomena can be observed in Fig. 1. First, the normalization improves the performances of both methods very significantly. This can be observed easily by comparing the error rates between the models with and without normalization. The error rates of the consistency method for four labeled points drop from 46% to 22%. When 50 points are labeled, the error rates drop from around 22% to about 13% and similar improvements can be found on the harmonic Gaussian method.

The only exception is in the case of the later method when there are only four labeled points available. In this situation, the error rate of the harmonic Gaussian is already much lower than the consistency method’s one, even without normalization and the improvement by normalization is not as significant as in other cases. This is a consequence of having balanced labeled data points from each class (1 in each class).
Hence, the mean values of $F^*$ along each column are closer to each other and there is no need for normalization.

In contrast, when the number of labeled points in each class is different (i.e., unbalanced which is the case whenever there is more than 4 labeled data for four classes and random labeling is used) the performance gain from normalization is more significant. The negative effect of unbalanced data can be observed from following the increase in error rate when working with ten data of labeled points and if normalization is not applied within the harmonic Gaussian method. Without normalization, the harmonic Gaussian method needs approximately forty unbalanced labeled points to match its very performance when having four balanced labeled points only. In contrast, the performance of the normalized model with ten unbalanced labeled data outperforms the result for the four balanced points. With a normalization step, the harmonic Gaussian method seems to be slightly better than the consistency method. This is not the case while working without the normalization. The best model for the text categorization data in our experiments is a harmonic Gaussian model with width equal to 0.3 which achieves an accuracy of 90% with only 50 labeled points out of 3970 of the total data points. For both methods with normalization of $F^*$, models with smaller width parameter perform slightly better than with the larger widths. Finally, for a 3970 data, the learning run based on a conjugate gradient algorithm takes only about 25 seconds of a CPU time on a 2MHz laptop machine for 100 random tests runs.

4 Conclusions

The extensive simulations have shown that an introduction of a normalization step improves the behavior of both transductive inference models (namely, consistency method and harmonic Gaussian one) very significantly. In both methods, the normalization of $F^*$ improves the performance up to fifty percents. However, the results are inconclusive, because many areas still need to be explored and more investigations are needed before final conclusions. For example, in this study we only investigate two basic models out of the twelve possible models mentioned earlier. Also, there are several parameters associated with these algorithms which can alter the overall performance of the model, e.g., the parameter for constraining the norm of $F^*$ (as suggested in [4]) can also have some impact on the performance of the models. This means that there may still be some space for improving the performance of the semi-supervised learning algorithms even further. In addition, the effects of a normalization step for other data set should also be further explored. The work presented here, can be treated as an initial step in this area only. It demonstrated that the way how the decisions are made from the output of these models can have a significant impact on the final classification performance. Our future work will go along the path of finding better decision strategies.

Acknowledgements

Our particular thank goes to Dr. Chan-Kyoo Park for all his support and communications on various guises of graph-based semi-supervised learning algorithms. We also thank Dr. Den-
gyong Zhou for introducing the first author to this area during his short stay at Max Planck Institute in Tübingen.

References

1. Huang, T. M., Kecman, V.: SemiL, Software for solving semi-supervised learning problems, [downloadable from: http://www.support-vector.ws/html/semil.html or from http://www.engineers.auckland.ac.nz/~vkec001], Auckland, (2004)
2. Kecman, V.: Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Systems, The MIT Press, Cambridge, MA, (2001)
3. Ng, A. Y., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm, Advances in Neural Information Processing Systems 14, (Eds.) Dietterich, T. G., Ghahramani, Z., MIT Press, Cambridge, Mass. (2002)
4. Park, C., Personal Communication, Tübingen, (2004)
5. Zhou, D., Bousquet, O., Lal, T. N., Weston, J., Schölkopf, B.: Learning with Local and Global Consistency, Advances in Neural Information Processing Systems 16, (Eds.) Thrun, S., L. Saul and B. Schölkopf, MIT Press, Cambridge, Mass. (2004) 321-328
6. Zhu, X.-J., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, (2003)
Handling Emergent Resource Use Oscillations

Mark Klein¹, Richard Metzler², and Yaneer Bar-Yam²

¹Massachusetts Institute of Technology
m_klein@mit.edu
²New England Complex Systems Institute
{richard, yaneer}@necsi.org

Abstract. Business and engineering systems are increasingly being created as collections of many autonomous (human or software) agents cooperating as peers. Peer-to-peer coordination introduces, however, unique and potentially serious challenges. When there is no one ‘in charge’, dysfunctions can emerge as the collective effect of locally reasonable decisions. In this paper, we consider the dysfunction wherein inefficient resource use oscillations occur due to delayed status information, and describe novel approaches, based on the selective use of misinformation, for dealing with this problem.

1 The Challenge

Business and engineering systems are increasingly being created as collections of many autonomous (human or software) agents cooperating as peers. The reasons for this are simple: the challenges we now face are simply too large, both in scale and complexity, to be handled by hierarchical control schemes. In many cases, moreover, political or other concerns exclude the possibility of centralized control even when it is technically feasible.

In such systems we face, however, the potential of highly dysfunctional dynamics emerging as the result of many locally reasonable agent decisions [1]. Such “emergent dysfunctions” can take many forms, ranging from inefficient resource allocation [2] [3] to chaotic inventory and price fluctuations [4] [5] [6] [7] to non-convergent and sub-optimal collective decision processes [8]. The properties of these dysfunctions often appear paradoxical, and their solutions often require new kinds of thinking.

In this paper we focus on one type of emergent dysfunction: resource use oscillation in request-based resource sharing. Imagine that we have a collection of consumer agents faced with a range of competing providers for a given resource (e.g. a piece of information, a sensor or effector, a communication link, a storage capability, or a web service). Typically, the utility of a resource is inversely related to how many consumers are using it. Each agent strives to select the least-utilized resource. This can be viewed as one element of workflow automation, as described in [9]. Such resource allocation is frequently carried out on a first-come first-served basis. This is a peer-to-peer mechanism - there is no one ‘in charge’ - which is widely used in settings that include markets, internet routing, and so on. It is simple to implement, makes minimal bandwidth requirements, avoids centralized bottlenecks and - in the absence of delays in resource status information – allows consumers to quickly converge to a near optimal distribution across resources.
Consumers, however, will often have a delayed picture of how busy each resource is. Agents could imaginably poll every resource before every request. This would cause, however, a N-fold increase in number of required messages for N servers, and does not eliminate the delays caused by the travel time for status messages. In a realistic open system context [10], moreover, consumers probably cannot fully rely on resource providers to accurately characterize the utility of their own offerings (in a way that is comparable, moreover, across providers). Resource providers may be self-interested and thus reluctant to release utilization information for fear of compromising their competitive advantage. In that case, agents will need to estimate resource utilization using other criteria such as their own previous experience, consulting reputation services, or watching what other consumers are doing. Such estimates will often lag behind the actual resource utility.

When status information is delayed, we find that resource use oscillations emerge, potentially reducing the utility achieved by the consumer agents far below optimum [11]. What happens is the following. Imagine that we have two resources, R1 and R2. We can expect that at some point one of the resources, say R1, will be utilized less than the other. Consumers at that point will of course tend to select R1. The problem is that, since their image of resource utilization is delayed, they will continue to select R1 even after it is no longer the less utilized resource, leading to an “overshoot” in R1’s utilization. When the agents finally realize that R2 is now the better choice, they will tend to select R2 with the same delay-induced overshoot. The net result is that the utilization of R1 and R2 will oscillate around the optimal equilibrium value. The amplitude of the oscillations, moreover, increases with the delay, to the extent that all consumers may at times select one resource when the other is idle:

![Figure 1](image.png)

**Fig. 1.** The utilization of two equivalent resources with and without status info delays

Such oscillations have two undesirable effects. One is that they can increase how long consumers have to wait for resources (i.e. reduce system throughput), because some resources may lay idle even when there are consumers not being served. They can also increase the *variability* in how long consumers have to wait for a resource, which...
may be significant in domains where consistency, and thus predictability, is valued. This problem is influenced, in seemingly paradoxical ways, by changing the number of resources and consumers. Reducing the server load, by increasing the number of servers, actually worsens the decline in throughput, and causes throughput losses to occur at lower status delay values. The throughput reduction can be substantial, reaching as high as 40%. Throughput losses increase and come at shorter status delays, moreover, as we increase the number of resources. The traditional ‘fix-all’ of increasing system capacity thus actually makes this emergent dysfunction worse. Despite their apparently counter-intuitive nature, these results can be explained simply. When the utilization of a resource is low, even small amplitude oscillations can cause it to go idle. And since all consumers shift to what they believe is the least-utilized resource, many resources can potentially go idle as a result of delay-induced oscillations.

Another paradox is that the more aggressive agents are at requesting only the least-utilized resource, the worse the problem gets. This strategy, moreover, is the individually rational one despite the throughput losses that can result. The incentives follow a prisoner’s dilemma game [12]. While everyone would be better off if all consumers occasionally selected what they believe to be the more heavily utilized resource (i.e. if everyone ‘cooperated’) the temptation is for agents to ‘defect’ (i.e. only request the least-utilized resource) to take advantage of the cooperators and/or avoid being taken advantage of themselves. Self-interested agents will thus find themselves driven to behaviors that cause resource use oscillations.

Resource use oscillations have been studied, primarily for the two resource case, in the literatures on “minority games” and distributed systems. The minority games literature [13] [14] has investigated how to design agents, typical using evolutionary techniques, so that their local decisions do not interact to produce oscillations.

While this approach does work under some conditions, it is unrealistic in an open systems context where agents are developed independently, so their resource request strategies are not subject to centralized control. The distributed computing work took the approach of creating an ecology of agents that each look at resource status information with some characteristic additional delay. Those agents whose total status delay matches the period of the resource use oscillation will, in theory, do a superior job of estimating current utilization and will come to dominate the population [15]. This approach has several disadvantages. First of all, it is a closed systems approach, in that it assumes that agents adhere to a centrally defined decision function. It also assumes that the delay in status information (and thus the oscillatory period) changes slowly or not at all. If the status delay changes more quickly than the agent population can evolve, the population will tend to be dominated by agents with inappropriate additional delays. It has been shown, in addition, that such systems are prone to sporadic bursts of strategy instability that can affect the period of resource use oscillations even in the absence of changes in actual status delays [16]. Finally, this work was only evaluated for the two resource case, so it’s value for larger numbers of resources is unknown. Our challenge, therefore, is to find an approach that moderates or eliminates oscillatory resource utilization dynamics without needing to control the design or operation of the consumer agents.
2 Efficiency Through Misinformation

As we have seen, emergent dysfunctions often have counter-intuitive properties. The solutions for emergent dysfunctions can, similarly, grow out of behavior that seems locally sub-optimal. This is the case with the techniques we have investigated. Our approach is predicated on resources (selectively) misinforming consumers about how busy the resource is. Paradoxically this can lead, as we show below, to superior resource allocation performance, including greater throughput and reduced variability.

The Scenario: All the approaches were evaluated in a scenario with multiple (from 20 – 50) consumers and multiple (2 – 5) resources. Consumers submit requests to the resource that they judge is the least heavily utilized. Resources differ in how quickly they can complete requests. When a request is received by a resource, it is placed on a queue and, once it reaches the front of the queue, the resource is allocated to the consumer for a length of time inversely proportional to the speed of the resource. When that period is over, a notification message is sent to the consumer. Messages take a fixed amount of time to travel from sender to receiver. Consumers wait a randomly selected amount of time, after a notification is received, before submitting a new request. The value of a resource to a consumer (though not of course the time it takes to access the resource) is independent of the resource’s utilization. The case where utilization does affect resource value is considered in [17]. The two metrics of interest to consumers in this scenario include (1) the aggregate throughput of the system, in terms of requests processed per time unit, and (2) the variability in request processing times. In our simulations, messages took 20 time units to propagate, the time gap between receiving a completion notification and a sending a subsequent request was normally distributed with an average of 40 and a standard deviation of 10, one server took 80 time units to service a request, and the other took 160 time units. Each simulation run was 10,000 time units long.

The dynamics of this system can be described analytically [18]. Queue lengths follow a triangle function where the frequency is determined only by the delay in status information, and the amplitude, for a given scenario, is determined only by the ratio of the delay time to the time it takes a resource to process a request. Message travel time has the same impact as status delays, because both increase the lag between a change in resource utilization and the response by consumers. When oscillations become so strong that the resources go idle periodically, the throughput of the system is inversely proportional to the status delay.

Status Misinformation: Let us assume that the resources have control over the status information that the consumers are given when they decide which resource to request. Let us further assume that consumers have a probability $p$ of being given information that leads them to select the ‘wrong’ (more heavily utilized) resource. The notion that agents can have somewhat ‘corrupted’ status information was broached in [15], but that work did not investigate how status misinformation can be beneficial by dampening delay-induced oscillations. Oscillations are damped because misinformation causes requests are spread to some extent to both resources, irregardless of which one is actually less utilized. It can be shown analytically [18] that for small levels of $p$, the variability in resource utilization is reduced linearly with $p$. As $p$ approaches 1, however, consumers get less and less ‘real’ information, and are increasingly likely to
choose resources without regards to their actual utilization, so resource utilization performs a ‘random walk’ [19], increasing the variability in request processing times and raising the possibility that the queue for one of the resources will empty out, thereby reducing throughput. So we are faced with a tradeoff. Small levels of \( p \) reduce the oscillatory amplitude, but larger ones increase it again due to the impact of random fluctuations. These insights are confirmed by simulations. When \( p \) is zero, we find that the variability in how long an consumer must wait for a resource increases, as we would expect, with the status information delay, due to periodic oscillations. When the delays get large enough to cause queue emptying, throughput drops. For intermediate values of \( p \), throughput is returned to near-optimal levels even with large delays, but variability is high. As \( p \) approaches 1, throughput drops off again (due to queue emptying caused by random walk fluctuations) and variability becomes higher yet. Throughput is maximized when \( p \) is about 0.7. Remarkably, performance is improved by imposing substantial misinformation.

Stochastic Request Rejection: The approach just discussed relies on the ability to control the information that consumers use to decide which resources to request. This is an unrealistic assumption, however, for many domains. In an open system, we do not have the control of consumer design that would be necessary to assure this. This approach also assumes that messages with resource status information are sent to consumers, either periodically (with a frequency at least as high as that of the delay-induced oscillations) or when they are about to make a resource request. This can substantially increase the message traffic required by the resource sharing protocol. This motivated us to explore an alternative approach for alleviating delay-induced resource use oscillations. The idea is simple: some fixed fraction of resource requests are rejected, at random, by resources. When a consumer receives a rejection message, it is (reasonably) assumed to send its request to some other server instead. The net effect is the same as with the previous approach in that, for some constant fraction of requests, consumers are misled about which resource is the least utilized. In the scenario we studied, throughput was maximized when 1/2 of all requests were rejected.

The stochastic request rejection approach can, however, reduce throughput if resource demands are low enough that the resource queues empty out due to request rejection. The reject messages also increase message traffic. The average number of rejections for a request, for \( p = 0.5 \), is 1, so an average of 2 requests will be needed to access a resource, increasing total required message traffic from 2 (one request and one notification) to 4 (two requests, one reject, and one notification).

Both of these disadvantages can be substantially ameliorated by adopting a load-dependent rejection scheme, inspired by the ‘random early drop’ scheme proposed for routers [20]. Instead of using a fixed request rejection frequency, resources reject requests with a frequency proportional to how full their queue is. The number of rejection messages generated is less (because high rejection rates are only incurred at the utilization peaks) and very few rejections occur when the resources are under-utilized, making it unlikely that throughput will be reduced because a request was rejected when a resource was available. Load-dependent rejection also offers the bonus of somewhat higher throughput than fixed-rate rejection; because the rejection rate (and thus the degree of damping) increases with the amplitude, the oscillations have a rounded shape that results in a smaller peak amplitude.
The average rate of rejection needs to be tuned to the current average load. There is a tradeoff involved. If the rejection regime is too aggressive, we incur excessive reject message traffic, and the possibility of causing queue emptying by rejecting requests when a resource is lightly utilized. If the rejection regime is not aggressive enough, however, there will be insufficient damping which can also lead to queue emptying and throughput loss.

The impact of the schemes we have discussed can be summarized and contrasted as follows. Misinformation-based techniques substantially increase throughput and reduce the variability in the time it takes to get a consumer request satisfied, for a wide range of delays, relative to the base case where these techniques were not used. Load-based rejection is the best technique in terms of throughput and variability, with the additional advantage of not assuming we can control the status information received by consumer agents, but incurs increased message traffic. These effects were statistically significant (p < .01).

One final refinement involves the realization that there is no point in incurring the increased message traffic caused by request rejection if there are no resource use oscillations, or if the oscillations are caused by variations in aggregate consumer demand rather than by status delays. This challenge, fortunately, is easy to address. Stochastic request rejection should only be activated if (1) there are significant periodic oscillations in resource utilization (determined by looking for above-threshold values in the power spectrum derived by a fast Fourier transform), and (2) the resource utilization across servers is negatively correlated (positive correlation would imply that aggregate demand is varying). We have implemented this approach and found that it successfully avoids being triggered by aggregate demand variations while remaining effective in responding to delay-induced oscillations.

The load-dependent stochastic rejection approach has also been shown, in our simulations, to effectively reduce the impact of status delay-induced oscillations when there are more than 2 resources.

3 Contributions and Future Work

We have presented a novel and promising approach for mitigating the deleterious effects of delay-induced resource-use oscillations on request-based resource sharing, by exploiting the paradoxical power of selectively misinforming consumers. The approach is designed to be appropriate for the important context of distributed systems with peer-to-peer coordination, where we can not rely on being able to control the design or operation of the resource consumers. Our future efforts will include empirical and analytic work. We will extend our analytic treatment to cover more than two resources. We also are developing an analytic way to determine the correct rejection regime for different contexts; we have done this empirically to date. We also plan to use our models to predict the degree of resource oscillation, as well as the potential benefits of selective misinformation, for real-world resources such as competing web sites.
Acknowledgements

This work was supported by the NSF Computational and Social Systems program as well as the DARPA Control of Agent-Based Systems program.

References

1. Jensen, D. and V. Lesser. Social pathologies of adaptive agents. In the proceedings of the Safe Learning Agents Workshop in the 2002 AAAI Spring Symposium. 2002: AAAI Press. Pgs 13 - 19.
2. Chia, M.H., D.E. Neiman, and V.R. Lesser. Poaching and distraction in asynchronous agent activities. In the proceedings of the Proceedings of the Third International Conference on Multi-Agent Systems. 1998. Paris, France. Pgs 88-95.
3. Hardin, G., The Tragedy of the Commons. Science, 1968. 162: p. 1243 - 1248.
4. Youssefmir, M. and B. Huberman. Resource contention in multi-agent systems. In the proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95). 1995. San Francisco, CA, USA: AAAI Press. Pgs 398-403.
5. Sterman, J.D., Learning in and about complex systems. 1994, Cambridge, Mass.: Alfred P. Sloan School of Management, Massachusetts Institute of Technology. 51.
6. Kephart, J.O., J.E. Hanson, and A.R. Greenwald, Dynamic pricing by software agents. Computer Networks: the International Journal of Distributed Informatique, 2000. 32(6): p. 731-52.
7. Ranjan, P., et al. Decision Making in Logistics: A Chaos Theory Based Analysis. In the proceedings of the AAAI Spring Symposium on Diagnosis, Prognosis and Decision Making. 2002. Pgs
8. Klein, M., et al., The Dynamics of Collaborative Design: Insights From Complex Systems and Negotiation Research. Concurrent Engineering Research & Applications, 2003. 11(3): p. 201-210.
9. Pham, H. and Y. Ye. Knowledgeable Objects as Data Agents for Business Automation. In the proceedings of the 2002 International Conference on Artificial Intelligence. 2002: CSREA Press. Pgs 1341-1347.
10. Hewitt, C. and P.D. Jong. Open Systems. Working Report Massachusetts Institute of Technology. 1982.
11. Hogg, T., Controlling chaos in distributed computational systems. SMC’98 Conference Proceedings, 1998(98CH36218): p. 632-7.
12. Osborne, M.J. and A. Rubinstein, A course in game theory. 1994, Cambridge, Mass.: MIT Press. xv, 352.
13. Challet, D. and Y.-C. Zhang, Emergence of Cooperation and Organization in an Evolutionary Game. arXiv:adap-org/9708006, 1997. 2(3).
14. Zhang, Y.-C., Modeling Market Mechanism with Evolutionary Games. arXiv:cond-mat/9803308, 1998. 1(25).
15. Hogg, T. and B. Huberman, Controlling chaos in distributed systems. IEEE Transactions on Systems, Man & Cybernetics, 1991. 21(6): p. 1325-32.
16. Youssefmir, M. and B.A. Huberman, Clustered volatility in multiagent dynamics. Journal of Economic Behavior & Organization, 1997. 32(1): p. 101-118.
17. Klein, M. and Y. Bar-Yam. Handling Resource Use Oscillation in Multi-Agent Markets. In the proceedings of the AAMAS Workshop on Agent-Mediated Electronic Commerce V. 2003. Melbourne Australia. Pgs
18. Metzler, R., M. Klein, and Y. Bar-Yam. *Efficiency Through Disinformation*. 2004. New England Complex Systems Institute. http://www.arxiv.org/abs/cond-mat/0312266
19. Bar-Yam, Y., *Dynamics of complex systems*. 1997, Reading, Mass.: Addison-Wesley. xvi, 848.
20. Braden, B., et al. *Recommendations on Queue Management and Congestion Avoidance in the Internet*. Working Report: 2309. Network Working Group. 1998.
A Practical Timetabling Algorithm for College Lecture-Timetable Scheduling

Kyoung-Soon Hwang, Keon Myung Lee, and Joongnam Jeon

School of Electric and Computer Engineering, Chungbuk National University, and Advanced Information Technology Research Center (AITrc)**, Korea

kmlee@cbnu.ac.kr

Abstract. A college timetable is a temporal arrangement of a set of classes and classrooms which all given constraints are satisfied. Timetabling has long been known to belong to the class of problems called NP-hard. This paper introduces a practical timetabling algorithm capable of taking care of both strong and weak constraints effectively, used in an automated timetabling system for a 2-year college. It presents what kind of the hard and soft constraints are imposed on the timetabling at the college and how to deal with them. It explains the data structures used to check the constraints. It also present the strategies for scheduling courses and for allocating classrooms to courses in the timetables. It shows some experiment results obtained in the application of the proposed algorithm to the college timetabling.

1 Introduction

Even though most college administrative works have been computerized, the lecture-timetable scheduling is still mostly done manually due to its inherent difficulties. The manual lecture-timetable scheduling demands considerable time and efforts. The lecture-timetable scheduling is a constraint satisfaction problem in which we find a solution that satisfies the given set of constraints.

The college lecture-timetabling problem asks us to find some time slots and classrooms which satisfy the constraints imposed on offered courses, instructors, classrooms, and so on. Therefore, the variables to be instantiated are times slots and classrooms of offered courses. Since the problem is a combinatorial optimization problem belonging to NP-hard class, the computation time for timetabling tends to grow exponentially as the number of variables gets larger.

There have been a number of approaches made in the past decades to the problem of constructing timetables for colleges and schools. Timetabling problems may be solved by different methods inherited from operations research such as graph coloring[2] and mathematical programming[6], from local search procedures such as tabu search and simulated annealing[8], from genetic algorithms[3], or from backtracking-based constraint satisfaction manipulation[4][5][7].

** This work has been supported by Korea Science and Engineering Foundation through AITrc.
We have developed an automated timetabling system for a 2-year college which provides lectures for both daytime students and night time students who work in daytime. In the developed system, we formulated the timetabling problem as a constraint satisfaction problem and proposed a practical timetabling algorithm capable of taking care of both strong and weak constraints and finding variables’ instantiation, which is based on the backtracking search[9].

2 Problem Formulation

The considered college is a two-year school who has the following characteristics on her course administration. The college offers courses for both daytime students and night time students who have jobs during daytime. The classes for daytime students are scheduled in the weekday’s daytime and Saturday morning. The courses for night time students are scheduled in the weekday nights and Saturday afternoon. The type of lectures are either theory lecture, seminars, or practicals. The class size of theory lectures is either 40 or 80. The seminar and practical classes have the size 40. A time slot is a 60 minutes interval. For theory and seminar classes, 1 slot time corresponds to 1 credit, and 1 credit for practicals takes 2 slots time.

Once an instructor decides to offer a course for a specific year-session of a department, an offered course $X_i$ takes place in the timetabling problem, which is expressed as a tuple of attributes (course, credits, department, instructor, year, section, class-group, course-type, lecture-type, $\text{TimeSlots}$, $\text{Rooms}$). Except $\text{TimeSlots}$ and $\text{Rooms}$, all attributes of $X_i$ are determined at the time the course is decided to be offered. Both $\text{TimeSlots}$ and $\text{Rooms}$ are list fields to contain assigned time slots and classrooms for the course $X_i$. To indicate an attribute $\text{attr}$ of an offered course $X_i$, we use the notation $X_i.\text{attr}$.

The time slots are assigned from 9 AM to 1AM for weekdays and from 9AM to 6PM Saturday. The 1st to 10th time slots of weekday and the 1st to 5th time slots of Saturday are for daytime student courses, the 11th to 6th time slots of weekday and the 6th to 10th time slots are for night time student courses. The daytime time slots are labeled as $T_i$ ($i = 1 \ldots 55$). The night time slots are labeled as $N_i$ ($i = 1 \ldots 35$).

There are various constraints to be satisfied at the time to instantiate variables about time slots and classrooms. The constraints can be categorized into strong and weak constraints as follows:

**Strong Constraints**

$C_1$: A classroom is not assigned to more than one lecture at the same time.

$C_2$: An instructor cannot teach more than one class at the same time.

$C_3$: Courses for the same year-session students of a department cannot take place at the same time.

$C_4$: The classroom for a course should have enough capacity to take students registered in the course.

$C_5$: The classroom should be well equipped with required facilities for the classes.
Weak Constraints

$C_6$: The lectures are not assigned to time slots which are in the instructor's forbidden time zones.

$C_7$: Instructors' daily lecture hours should be restricted to be within the allowed maximum hours.

$C_8$: As far as possible, classes are scheduled in the instructor's preferred time zones.

$C_9$: A lunch/dinner break must be scheduled.

$C_{10}$: If possible, the lectures should not be scheduled on too late night time slots.

$C_{11}$: The theory courses are scheduled on Monday and Tuesday, and the practical courses are scheduled on Wednesday, Thursday, and Friday.

$C_{12}$: For daytime students, the cultural subjects courses are scheduled in the morning time slots (1st to 4th time slots on weekdays), and major courses are scheduled in the afternoon time slots (5th to 8th time slots).

$C_{13}$: For nighttime students, the cultural-subjects courses are scheduled on the 11th to 12th slots, and the major courses are scheduled on the 13th to 16th time slots on weekdays.

$C_{14}$: If possible, the lecture hours for a course should be scheduled consecutively.

$C_{15}$: As far as possible, classes should be scheduled in their corresponding department's exclusive-use classrooms.

$C_{16}$: The classrooms should be allocated in a manner to minimize the distances between adjacent classes's classrooms.

It is desirable for timetables to satisfy all strong and weak constraints. However, it is usually not easy to meet all these constraints. The strong constraints must be satisfied all the times, but weak constraints can be somewhat sacrificed to find feasible timetables. Among the weak constraints, constraints from $C_6$ to $C_{14}$ are imposed on the allocation of time slots. Constraints $C_{15}$ and $C_{16}$ are imposed on the allocation of classrooms. The constraints are arranged in the order of importance in the scheduling. For example, if it is impossible to find schedules satisfying both $C_6$ and $C_7$ simultaneously, it is preferred to choose a schedule that satisfies $C_6$ but $C_7$ rather than a schedule satisfying $C_7$ but $C_6$. In the course of timetabiling, if needed, the weak constraints could be sacrificed in the reverse order of the sequence above-listed.

3 The Proposed Timabiling Method

This section presents the timetabling algorithm used in the implemented system. The algorithm uses backtracking-based search to find proper variable instantiations satisfying the imposed constraints. It also discusses in detail the variable ordering method, the time slots scheduling method, the classrooms scheduling method, and the application strategy to a timetabling problem of large size.
3.1 The Timetabling Algorithm

The timetabling algorithm takes as an input variables \( X = \{X_1, X_2, \ldots, X_n\} \) where \( X_i \) corresponds to an offered course which has the following attributes: (course, credits, department, instructor, year, section, class-group, course-type, lecture-type). Each variable \( X_i \) has two list fields to be instantiated: \( X_i.TimeSlots \) for the time slots allocated to the course of \( X_i \), and \( X_i.Rooms \) for the classrooms allocated to \( X_i \). The domains \( TS \) of time slots must be given as an input: \( TS = \{TS_1, TS_2, \ldots, TS_n\} \) where \( TS_i \) is the domain of time slots for \( X_i \); if \( X_i.class-group = \text{daytime} \), \( TS_i = \{T_1, T_2, \ldots, T_{55}\} \); otherwise, \( TS_i = \{N_1, N_2, \ldots, N_{35}\} \). The domains \( CR \) of classroom domains should be also provided: \( CR = \{CR_1, CR_2, \ldots, CR_n\} \) where \( CR_i = ER_i \cup SSR_i \). \( ER_i \) is the set of exclusive-use classrooms for the department of \( X_i \), and \( SSR_i \) is the set of shared classrooms of the school to which \( X_i \)'s department belongs. The constraints imposed among variables are also provided as the input to the algorithm.

The algorithm outputs the instantiated time slots \( X_i.TimeSlots \) and classrooms \( X_i.Rooms \) for each offered course \( X_i \) which satisfy the imposed constraints. Both \( X_i.TimeSlots \) and \( X_i.Rooms \) are lists of which size is the lecture hours of the corresponding course.

Procedure. Timetabling-by-Constraint-Satisfaction

Input: variables \( X = \{X_1, X_2, \ldots, X_n\} \) and the domains of \( X_i.TimeSlots \) and \( X_i.Rooms \); constraints on the variables \( C = \{C_1, C_2, \ldots, C_{16}\} \)

Output: instantiations to \( X_i.TimeSlots \) and \( X_i.Rooms \) for \( X_i \in X \).

Step 1. (Step forward) If the current variable \( X_{\text{cur}} \) is the last variable, exit with the instantiated values. Otherwise, select the next variable in the ordered sequence of variables and call it \( X_{\text{cur}} \). Initialize \( D'_{\text{cur}}.TimeSlots = TS_{\text{cur}} \) and \( D'_{\text{cur}}.Rooms = CR_{\text{cur}} \).

Step 2. (Schedule Time Slots) Select available time slots from \( D'_{\text{cur}}.TimeSlots \) for \( X_{\text{cur}}.TimeSlots \) that are consistent with all previously instantiated variables. Do this as follows:

(a) If \( D'_{\text{cur}}.TimeSlots = \phi \), extend \( D'_{\text{cur}}.TimeSlots \) by sacrificing a weak constraint according to the ordered sequence of weak constraints imposed on time slots one by one. If there are no more weak constraints to be sacrificed, go to Step 4.

(b) Select a value \( t \) from \( D'_{\text{cur}}.TimeSlots \) and remove it from \( D'_{\text{cur}}.TimeSlots \). If there is already some time slot scheduled for \( X_{\text{cur}} \), choose a time slot adjacent to the already scheduled time slot as far as possible.

(c) For each constraint defined on \( X_{\text{cur}} \) and \( X_i, 1 \leq i < \text{cur} \), test if the previously instantiated variables’ values are consistent with \( X_{\text{cur}}.TimeSlots \) appended with \( t \). If not, go to Step 2a.

(d) Append \( t \) to the end of \( X_{\text{cur}}.TimeSlots \).

(e) If it has scheduled as many time slots as the lecture hours corresponding to \( X_{\text{cur}}.credits \), go to Step 3. Otherwise, go to Step 2a.
Step 3. (Schedule Classrooms) Select available classrooms from $D'_{cur}.Rooms$ for course $X_{cur}.Rooms$ that are consistent with all previously instantiated variables. Do this as follows:

(a) Select a classroom $r$ from the exclusive-use classrooms $ER_i$ of $D'_{cur}.Rooms$, for a time slot $t$ of $X_{cur}$, which does not yet have an allocated classroom. If there is already some classroom scheduled for $X_{cur}$, choose the same classroom or near classroom to the already scheduled classroom, as far as possible. If so, append $r$ to the end of $X_{cur}.Rooms$ and remove $r$ from $D'_{cur}.Rooms$, and then go to Step 3d.

(b) Select a classroom $r$ from the school shared classroom $SSR_i$ of $D'_{cur}.Rooms$. If there is such a classroom and its usage is still within its department quota, then append $r$ to the end of $X_{cur}.Rooms$ and remove $r$ from $D'_{cur}.Rooms$, and then go to Step 3d.

(c) If there are no available classrooms for $t$ of $X_{cur}$ and $X_{cur}.lecture-type = practical$, go to Step 2a. Otherwise, append no-classroom to the end of $X_{cur}.Rooms$.

(d) If there remains some time slot of $X_{cur}$ for which classrooms are not yet considered, go to Step 3a.

(e) Go to Step 1.

Step 4. (Step backward) If $X_{cur}$ is the first variable, exit and declare that the problem’s constraints are inconsistent. Otherwise, select the previous variable and call it $X_{cur}$ and go to Step 2a.

3.2 Variable Ordering

The instantiation order of variables largely affects on the computation time and solution quality. The instantiation order can be either dynamically determined on the fly or predetermined before the application of the algorithm. Considering the constraints on time slots of courses, the proposed method determines the instantiation order earlier on. Cultural-subjects courses are supposed to be scheduled in some specific time zones to allow as many students to take courses as possible. Thus the variables are ordered in the following sequence:

- cultural subjects courses with preferred time zone
- cultural subjects courses without preferred time zone
- major courses with preferred time zone
- major courses without preferred time zone

If multiple courses are on the same rank, the practical courses requiring specific classrooms come first and the courses with large class size come next and then others come. If more one than a course are still on the same rank, the tie breaks are done at random.

3.3 Constraint Checking

To make easy to check the consistency of a time slot variable value with respect to already instantiated variables, the developed system maintains some data structures as follows: $CLRM(classroom-ID i, A_{CLRM}(i))$, $INS(instructor-ID i, A_{INS}(i))$, $DEPT-YEAR(department-ID d, year y, section s, A_{DYS}(d, y, s))$. 
**CLRM**(classroom-ID \(i\), \(A_{\text{CLRM}}(i)\)) is a data structure for a classroom \(i\) where \(A_{\text{CLRM}}(i)\) is an array of size \(N\) and \(N\) is the total time slots in a timetable. An element of the array corresponds to a time slot and its value is 1 if the classroom is scheduled to a course during the time slot period, 0 otherwise. \(\text{INS}(\text{instructor-ID } i, A_{\text{INS}}(i))\) is a data structure for an instructor \(i\) and \(A_{\text{INS}}(i)\) is an array of size of \(N\) for instructor \(i\). An element corresponding to a time slot has the following values: 1 if it is scheduled for a lecture of the instructor, 2 if it is in the preferred time zone of the instructor, 3 if it is in the forbidden time zone of the instructor, 0 otherwise. \(\text{DEPT-YEAR}(\text{department-ID } d, \text{year } y, \text{section } s, A_{\text{DY}}(d, y, s))\) is a data structure to prevent courses, offered to students of the same year-section and department, from being scheduled in parallel. The element value of \(A_{\text{DY}}(d, y, s)\) is 1 if a lecture is scheduled during the corresponding time slot period for a course of the corresponding year-section and department.

The constraints can be checked against these data structures. Constraint \(C_1\) can be handled with \(\text{CLRM}\), \(C_2\) with \(\text{INS}\), and \(C_3\) with \(\text{DEPT-YEAR}\). \(C_4\) and \(C_5\) can be checked by referring the attributes of classrooms. \(C_6\), \(C_7\) and \(C_8\) are taken cared by referring \(\text{INS}\). \(C_9\) can be taken into account by controlling allowed time slots. \(C_{10}\) can be implemented by considering available time slots in the increasing order of time. Constraints from \(C_{11}\) to \(C_{13}\) can be imposed by controlling the allowed time zones to courses according to their course type and course class. \(C_{13}\), \(C_{14}\) and \(C_{16}\) can be imposed by ordering available classrooms and referring the attributes of classrooms.

### 3.4 Classroom Scheduling

The classrooms are classified into exclusive-use classrooms for a department, shared classrooms for a school, shared classrooms for the entire college. Each department has some exclusive-use classrooms. It is desirable for a department to use its own exclusive-use classrooms to serve its lectures. If it is impossible, the department may use some shared classrooms of the school to which the department belongs.

To prevent a department from using a lot of shared classrooms, we impose the constraints to allow a department to use shared classrooms within the specified quota. If a school contains \(m\) departments and has \(n\) shared classrooms for its departments, the usage quota of the shared classrooms for a department is \((n/m)\times(\text{the number of total time slots})\). The proposed algorithm assigns classrooms to courses if it is possible. Otherwise, it labels the classroom fields with no-classroom. As the postprocessing, the implemented system fills the classroom fields for scheduled time slots. After the algorithm constructs timetables for departments, the implemented system fills the classroom fields with value no-classroom. To assign classrooms to the scheduled time slots without allocated classrooms, it collects all unoccupied classrooms for the corresponding time slots across the college, and then finds appropriate classrooms considering class size and distances.
3.5 Strategy for College Timetabling

It is possible theoretically to perform timetabling for all departments at a time. Due to inherent properties of combinatorial optimization like timetabling, the computation time for college-wise timetabling grows in exponential order. To avoid explosive computation time, the proposed method sorts departments in a sequence in some way, and finds timetables for departments according to the sorted sequence. After that, it does postprocessing to take care of unfilled classroom fields. This approach has been employed in practice in college timetabling. If there are some courses for which timetabling is not yet completed, administrator may involve in modifying the generated timetables so as to put those courses into them.

4 Implementation and Experiments

The system was developed using Powerbuilder 7.0 and PC Oracle 8.0 on Windows 2000 NT server with Pentium III-500, 256M RAM. The system provides the graphical user interface with which users input information about courses, classrooms, instructors, and update manually the produced timetables.

Table 1 shows the experiment data and results. The implemented system has been applied to timetabling of 5 departments: English Translation (ETN), Management Information System (MIS), Financial Information Systems (FIS), Fashion Industrial Design (FID), and Internet Information Systems (IIS). In the table, total indicates the total number of offered courses, lectures is the lecture hours, theory and practical are the numbers of theory courses and practical courses, respectively. Pt is the number of lecture hours with preferred time zones, NPt is the number of lecture hours that are scheduled in non-preferred time zones, and P is \((Pt - NPt)/Pt \times 100\). Ct is the number of courses with prohibited time zone, NCt is the number of courses for which scheduled time slot is in prohibited time zone, and \(C = (Ct - NCt)/Ct\). Time is the computation time elapsed to get timetables. The computation time and timetable quality were considered as

| dept | year | instructors | Courses | total | lectures | theory | practical | Pt | NPt | P (%) | Ct | NCT (%) | C | Time (min) |
|------|------|-------------|---------|-------|---------|--------|-----------|----|-----|-------|----|---------|---|-----------|
| ETN  | 1    | 11          | 13      | 53    | 8(1)    | 6      | 20        | 0  | 100 |       | 100|         |    | 13:26     |
|      | 2    | 11          | 12      | 53    | 7(1)    | 6      | 40        | 0  | 100 |       | 100|         |    |           |
| MIS  | 1    | 9           | 10      | 49    | 9(1)    | 3      | 20        | 0  | 100 | 4      | 0  | 100     |    | 9:13      |
|      | 2    | 7           | 8       | 41    | 8(1)    | 0      | 0         |    |     | 4      | 0  | 100     |    |           |
| FIS  | 1    | 9           | 12      | 47    | 10(1)   | 2      |           |    |     |       |    |         |    | 10:48     |
|      | 2    | 6           | 8       | 37    | 7(1)    | 1      |           |    |     |       |    |         |    |           |
| FID  | 1    | 7           | 8       | 56    | 8       | 6      | 120       | 60 | 50  | 40    | 0  | 100     |    | 21:80     |
|      | 2    | 10          | 10      | 39    | 10(6)   | 8(5)   | 20        | 0  | 100 |       |    |         |    |           |
| IIS  | 1    | 9           | 9       | 58    | 9       | 7      | 60        | 0  | 100 |       |    |         |    | 7:09      |
Fig. 1. A timetable generated by the system

acceptable by the administrative officer of the college. The postprocessing for classroom allocation for time slots with no classroom mark has taken less than 1 minute in the experiment. Figure 1 shows a screen shot of a timetable generated by the developed system.

5 Conclusions

This paper presented the characteristics of a college timetabling problem in terms of constraints. It discussed in detail a practical timetabling algorithm capable of taking care of both strong and weak constraints in the course of variable instantiations. It also presented the data structure used to check the constraints. The algorithm always instantiates the time slots without exceptions but allows some classrooms not to be scheduled for a while. The unscheduled classrooms are treated as a postprocessing by collecting all available resources. This strategy is useful because it is usually difficult to find available time slots and classrooms at the same time under the situation resources are limited. To save the computation time, the system takes the approach to ordering departments in some sequence, trying to find timetables sequentially according to the sequence, and then taking care of unscheduled classrooms. In the experiments, the implemented system found acceptable timetables even though sometimes administrator has to involve somewhat to update final timetables.

References

[1] J. Ferber, Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison Wesley Longman. (1999) 509p.
[2] E. K. Burke, D. G. Elliman, R. Weare. A University Timetabling System based on
Graph Colouring and Constraint Manipulation. in J. of Research on Computing
in Education. (1996).
[3] J. P. Caldeira, A. C. Rosa. School Timetabling using Genetic Search. In 2nd Int.
Conf. on Practice and Theory of Automated Timetabling. (1997).
[4] H.-J. Goltz, G. Kuchler, D. Matzke. Constraint-Based Timetabling for Universities.
In INAP’98. (1998) 75-80.
[5] H. Rudova, L. Matyska. Constraint-based Timetabling with Student Schedules. In
PATAT. (2000).
[6] N. L. Lawrie. An integer linear programming model of a school timetabling problem.
Computer Journal 12. (1969) 307-316.
[7] M. Yokoo, E. H. Durfee, T. Ishida, K. Kuwabara. Distributed Constraint Satis-
faction for formalizing distributed problem solving. In 12th IEEE Int. Conf. on
Distributed Computing Systems. (1992) 614-621.
[8] M. A. Saleh Elmohamed, P. Coddington, G. Fax. A comparison of annealing
techniques for academic course scheduling. In Practice and Theory of Automated
Timetabling II (E. Burke and M. Carter(eds.)). Springer-Verlag LNCS 1408. (1998)
92-112.
[9] R. Dechter. Backtracking algorithms for constraint satisfaction problems — a sur-
vey. Constraints, International Journal. (1998).
Java Bytecode-to-.NET MSIL Translator for Construction of Platform Independent Information Systems

YangSun Lee* and Seungwon Na**

*Dept. of Computer Engineering, Seokyeong University, 16-1 Jungneung-Dong, Sungbuk-Ku, Seoul 136-704, KOREA
yslee@skuniv.ac.kr

**Platform R&D Center, SK Telecom CO. LTD., 84, Taepyungro 1-Ga, Chung-Gu, Seoul 100-768, KOREA
nasw@dgu.ac.kr

Abstract. This paper presents the java bytecode-to-.NET MSIL translator, to construct a platform independent information system, which enables the execution of the java program in .NET environment without JVM(Java Virtual Machine), translating bytecodes produced by compiling java programs into MSIL codes. Java, one of the most widely used programming languages recently, is the language invented by James Gosling at Sun Microsystems, which is the next generation language independent of operating systems and hardware platforms. Java source code is compiled into bytecode as intermediate code independent of each platform by compiler, and also executed by JVM. .NET language such as C# and .NET platform in Microsoft Corp. has been developed to meet the needs of programmers, and cope with Java and JVM platform of Sun Microsystems. After compiling, a program written in .NET language is converted to MSIL code, and also executed by .NET platform but not in JVM platform. For this reason, we designed and implemented the java bytecode-to-.NET MSIL translator system for programs written in java language to be executed in the .NET platform without JVM.

1 Introduction

Information systems such as transaction procession systems(TPS), management information systems(MIS), and decision support systems(DSS) and executive information systems(EIS) are developed mainly by java language or .NET language such as C, C++ and VB. Java is dependent on java platform, JVM and .NET languages are dependent on .NET platform. Consequently, Information systems constructed by java or .NET language also are dependent on platform, and can not be executed on different platforms. So, we present a method to execute java programs on the .NET platform without JVM[17,18].

This work was supported by grant No.(R01-2002-000-00041-0) from the Basic Research Program of the Korea Science & Engineering Foundation.
Java is a language that generates execution files through compiling. The generated files are in the bytecode form, which is interpreted and executed by JVM (Java Virtual Machine), the Java interpreter. Therefore, Java programming language is both compiler and interpreter language, but more universal than normal compilers and runs faster and more efficiently than usual interpreters [2, 5, 8].

On the other, C# language provided by Microsoft .NET platform is offered as a platform-independent language like Java of Sun Microsystems, the development of components by C# is becoming widespread. C# is a language based on the characteristics of high coding convenience and easy maintenance of Microsoft Visual Basic and flexibility and intensity of C++ which enables much faster and easier development of COM++ and web services [1, 3, 4, 14].

On this environment of .NET platform, we can create any code that, once converted to MSIL, can generate an execution file optimized for the target platform regardless of the language used. Therefore, by converting the Java bytecode to .NET MSIL, we can run the Java programs on the Windows .NET platform even without JVM.

In summary, this paper presents a method to translate java bytecode to .NET MSIL code using a mapping table and the macro translation method, to construct a platform independent information system. Indeed, we designed and implemented a translator using this method, and translate a java bytecode program to MSIL code to run it on the .NET platform without JVM.

2 Bytecode and MSIL Code

2.1 Bytecode

Bytecode can be considered as a machine language for JVM (Java Virtual Machine). It is acquired in the stream format for each method within a class when JVM loads a class file. At this time, the stream is in the binary stream format of a 8-bit byte. Furthermore, a bytecode basically has a stack-oriented structure, originally intended for being used through an interpreter. In other words, it is either interpreted by JVM, or compiled when the class is loaded [2, 5, 8, 15, 16].

JVM (Java Virtual Machine) saves and executes the source code by java programming language using class file format. Since the class file is in binary format, it is very difficult to analyze and modify. On the other hand, Oolong code, another type of java intermediate language, is much easier to read and write compared to class file format. Oolong code is based on Jasmin language of John Meyer, and designed to work in the level of bytecode [5, 13].

2.2 MSIL Code

MSIL (Microsoft Intermediate Language) is an intermediate language of .NET language such as C#, comprising a set of stack-based instructions designed to be easily generated from source codes through a compiler or other tools. The instructions are classified into arithmetic/logical operations, flow control, DMA, exception handling,
and method calling. In addition, the virtual method calling, field access, array access, object allocation and initialization, which affect the programming structure, are also types of instructions directly supported by MSIL. The MSIL instruction set searches data types in stacks, and is directly interpreted.

Furthermore, MSIL is independent of hardware and platform, and was originally designed for JIT. Unlike Java, it was designed to be language independent of the first and targeted at generic programming. Consequently, the language adapts very well to the change of program functions and structures [1,3,6,7,9,11,13].

3 Java Bytecode-to-.NET MSIL Translator

For the translation system from Java bytecode to .NET MSIL code to construct a platform independent information system, we designed and implemented the code translator by using the mapping table and the macro conversion method. We must convert the matching between codes through the mapping table so that the instruction mapping part and the function format structure conversion part, which are the main components of the translator, would become functionally equivalent to the relationship between bytecode and MSIL code. Furthermore, in the actual implementation, we will refer to the table corresponding to each of the codes by using the macro conversion method.

3.1 System Configuration

Figure 1 shows the system configuration of the translator from Java bytecode, namely Oolong code, to .NET MSIL code. For the process of extracting the Oolong code from class file, with the class file as an input, the Oolong disassembler, Gnoloo, was used to output the Oolong code in the text assembly format. Next, with the code converter, we generated an MSIL code from the Oolong code used as the input file. Finally, we created an *.exe file from the MSIL code through ilasm.exe provided by the .NET platform.

3.2 Java Bytecode-to-.NET MSIL Translator

With the extracted Oolong code as an input to the translator, we used the instruction mapping table and the function format structure conversion table to generate MSIL code, which is the resulting code of the translator. As shown in Figure 2, we obtained an MSIL code from the Oolong code through the extraction process, and the result is generated from the two processes of mapping and conversion.

The development environment for the translator is divided into the JDK and Oolong part to create and disassemble the java class file to be input to the translator, and the .NET platform SDK part to convert the translated MSIL file into a *.exe file.

3.2.1 Data Type Mapping Table

This is the table for the basic data type mapping used by the translator for a conversion (Table 1). There are some basic data types for Oolong code, which are included in the 17 data types used by MSIL.
3.2.2 Instruction Mapping Table
Table 2 lists only the basic instructions. The instructions are classified based on the MSIL, to which the corresponding Oolong instructions are mapped.

4 Experiment Results and Analysis

The followings are the example of java array program. program 1 below depicts the extraction of Oolong code from java program source. The translator used the extracted Oolong code as its input, and translated the .NET MSIL code corresponding to each Java Oolong code through the mapping table and macro conversion within the translator, which is exactly the case for program 2.
Table 1. Data Type Mapping Table

| **Oolong** | **MSIL** | **Description** |
|------------|----------|-----------------|
| iadd, fadd, ladd, dadd | add | addition |
| imul, fmul, lmul, dmul | mul | multiplication |
| ... | ... | ... |
| newarray | newarr | creating array |
| string | string | unicode string |
| ... | ... | ... |
| byte | int8 | signed 8-bit integer |
| short | int16 | unsigned 16-bit integer |
| int | int32 | unsigned 32-bit integer |
| long | int64 | unsigned 64-bit integer |
| ... | ... | ... |

Table 2. Instruction Mapping Table

| **Oolong** | **MSIL** | **Description** |
|------------|----------|-----------------|
| iload.0 | ldarg.0 | load the argument number 0 on the stack. |
| ... | ... | ... |
| istore_0 | stloc.0 | store a value in local variable number 0 on the stack |
| ... | ... | ... |
| iadd | add | addition |
| imul | mul | multiplication |
| ifnonnull | btrue | branch if <value> is nonzero. |
| putfield | stfld | store into a field of an object. |
| getfield | ldfld | load a field of an object. |
| ... | ... | ... |

Public class ArrayTest {  
   public static void main()  
   {  
      int m[]={1,2,3};  
      for (int i=0; i<3; ++i)  
         system.out.println(m[i]);  
   }  
}
Program 1. Java Program, Extracted Oolang Code and Corresponding MSIL Code

Figure 3 shows the result of the execution after translating the java Oolang program to .NET MSIL program and converting it into executable file.

Fig. 3. Execution Result

It shows the same results from the execution of ArrayTest.j extracted by the Oolang disassembler from the class file generated by the Java compiler and the execution of MSIL file ArrayTest.exe generated by the translator with ArrayTest.j file as input.
5 Conclusion

In this paper, to construct a platform independent information system, and in order to enable the execution of java programs on the .NET platform environment, we compared the bytecode, intermediate language for JVM, with MSIL, intermediate language for the .NET platform, and then we designed and implemented the bytecode-to-MSIL translator to convert the bytecode into the MSIL code. In this process, we composed an assembly instruction mapping table as well as a function format structure conversion table. We are currently researching on the optimization method to generate the better codes to speed up and to be able to convert the entire code format for large programs as well.

References

[1] Andrew Troelsen, C# and the .NET Platform, APress, 2001.
[2] Bill Venners, Inside the JAVA Virtual Machine, 2nd ed., McGraw-Hill, 2000.
[3] Don Box & Chris Sells, Essential .NET Volume 1 The Common Language Runtime, Addison-Wesley, 2002.
[4] Eric Gunnerson, A Programmer’s Introduction to C#, APress, 2001.
[5] Hoshua Engel, Programming for the Java Virtual Machine, AddisonWesley, 1999.
[6] Jeff Prosise, Programming Microsoft .NET, Microsoft Press, 2002.
[7] John Gough, Compiling for the .NET Common Language Runtime(CLR), Prentice Hall, 2002.
[8] Ken Arnold & James Gosling, The Java™ Programming Language, 3rd ed., Addison-Wesley, 2000.
[9] Microsoft, C# Language Specification, Nov. 2000.
[10] Microsoft Corporation, Common Language Infrastructure(CLI), Dec. 2001.
[11] Microsoft, MSIL Instruction Set Specification, Nov. 2000.
[12] Microsoft, The IL Assembly Language Programmer’s Reference, Oct. 2000.
[13] Serge Lindin, Inside Microsoft .NET IL Assembler, Microsoft Press, 2002.
[14] Simmon Robinson, Professional C#, Wrox, 2002.
[15] Tim Lindholm & Frank Yellin, The Java™ Virtual Machine Specification, 2nd ed., Addison-Wesley, 1999.
[16] Troy Downing & Jon Meyer, Java Virtual Machine, O’REILLY, Mar. 1997.
[17] Ralph M. Stair, Principles of Information Systems: A Managerial Approach, Boyd & Fraser Publishing, 1992.
[18] James A. O’Brien, Management Information Systems: A Managerial End User Perspective, IRWIN, 1990.
A Scale and Viewing Point Invariant Pose Estimation

M. Y. Nam and P. K. Rhee

Dept. of Computer Science & Engineering Inha University
Biometrics Engineering Research Center
Yong-Hyun Dong, Incheon, Korea
rera@im.inha.ac.kr, pkrhee@inha.ac.kr

Abstract. This paper addresses the framework and implementation of a viewing point invariant face detection system using appearance-based learning method and hybrid learning approach. Multiple view-based object detection is attractive since it could accumulate the object models by autonomous learning process. An object, however, can be approximated even though it is represented with many scales and views. Face detection determines the location and size of each human face (if any) in an input image. A closely related problem is face recognition. And pose estimation is a hard problem in face recognition. Detecting varying pose human face in video frames is an important task in many computer vision applications such as human-computer interface. The scale invariant face detection system and pose estimation proposed in this paper employs hybrid learning approach and statistical method. We employ FuzzyART and RBF network and mahalanobis distance for optimal pose estimation and an optimal face detecting architecture. We achieve very encouraging experimental results.

1 Introduction

We carry a conceptual idea of real world objects in our thought as socrates explained to homogeneous. The difficulties in visual detection/recognition are caused by the variations in viewpoint, viewing distance, illumination, etc. Recognizing 3-D objects under various view-points and lighting conditions is the final goal of computer vision. Robust object detection needs high invariance with regard to those variations. Much research has been done to solve this problem [1]. The face detection systems can be divided into three major categories [2]: model-based method, image invariant method, and appearance-based learning method. In the model-based method, an object model is defined, and this model is matched to the image. The second one, image invariant method is based on a matching set of image pattern relationships. The last one learns the features of objects of labeled categorized object examples. Appearance-based approach has been employed successfully in the computer vision areas [3].

Face detection determines the location and size of each human face (if any) in an input image. A closely related problem is face recognition. And pose estimation is a hard problem in one of face recognition. Detecting varying pose human face in video frames is an important task in many computer vision applications such as human-computer interface, and multimedia retrieval. In this paper, FuzzyART (Fuzzy
Adaptive Resonance Theory) and RBF are employed to simulate the capacity of the high level cell in the attentive process face recognition system (Human visual system).

The pose invariant face detection system proposed in this paper employs combined supervised and unsupervised learning. We employ FuzzyART and RBFN for optimal pose estimation for an optimal face detecting architecture. We achieve very encouraging results. The outline of this paper is as follows. In section 2, we present the architecture of the proposed face detection system. The pose classification process is discussed in section 3. We give experimental results in section 4. Finally, we give concluding remarks.

2 The Proposed System Architecture

We propose a generic scheme of object detector for multiple viewpoints and scales. The specific task we are discussing in this paper is a generic face detector which is invariant viewing angles and scales. The system consists of the multi-scale module and the object vector representation module, object detector. Initially, seed appearance models of an object is manually gathered and classified for training the detector module. The detector with prior classes is trained by the initial seed data set.

![Fig. 1. Face Detection System Architecture](image)

System’s structure shows as Fig. 1, pose estimation have combined network RBF network and FuzzyART.

2.1 Scale-Invariance Face Detection by Pyramid

Instead of using a multi-resolution pyramid and multiple mahalanobis algorithm for scale invariant face detection [18], face size is estimated from the segmented image. The problem is formulated as a classification problem that is face and non-face. Initially, a training set of faces is gathered. Each Training face image is scaled to 20x20 and normalized using max-min value normalization. According to vectorize way, affected much face detection performance. Searched for vectorization method that is optimized by an experiment, as a result, raised good performance. We could improve performance through efficient vectorization and dimension decrease. Training image contains 3,000 grayscale face pictures of size $20 \times 20$. We have
vectorized on each training image for each cell, we create a vector of 54 dimension that face image’s multi-resolution and blocked area average value. Training data have vectorized that can display performance by maximum doing not lose face characteristic information. This vectorization method have enhanced face detection rate face’ feature don’t reduce and face detection result is enhancement. Training data and vector of face to train are shown Fig. 2.

![Image of source images and block vectorization]

**Fig. 2.** Training face data and vectorization: mean value per each block

Squared mahalanobis distance[4] is one useful methods that measures free model course free pattern relationship similarity degree. The center of cluster is determined by the mean vector, and the shape of the cluster is determined by the covariance matrix mahalanobis distance shows Eq. 1.

\[ r^2 = (\mathbf{x} - \mu_x)\Sigma^{-1}(\mathbf{x} - \mu_x) \]  

(1)

Where \( \mu_x \) is average vector of face’s vector, \( \Sigma \) is Covariance matrix of independent elements. Eigenvalues determine the length of these axes. In order to determine how a single boundary relates to the collection of boundaries in the learning set, a distance measure needs to be defined. In the active shape model literature, the evaluation of boundary feature values is performed in terms of the mahalanobis distance, i.e. the distance to the average normalized by the variation in each dimension. Following Cootes et al., we use a mahalanobis distance model to compute the distance of a feature function to the average of the learning set[4, 5]. We divided only face image space into several probability clusters. We compute the mahalanobis distance from face image space.

The generation seen therefore forms model in 3,000 face images and computed mahalanobis distance, and the computed result decides face and non-face. Multi-resolution consists of nine steps by an experiment, and offset established by four pixels for face detection of various size. These models gathered and categorized manually are used to construct initial object detector with their possible pattern variations in a high-dimensional image vector space.

Fig. 4 shows result that mahalanobis distance of face and non-face image by Eq. 1. The distance is normalized between 0 and 255. The distance of face have lower reaction value and distance of non-face have higher reaction value. When the system is given an image in which to find the faces, the image is rescaled to multi-resolution, it takes each sub-window of the image, rescales them to size of 20 × 20, applies the
preprocessing, and computes the mahalanobis distance. The distance leading to a classification of each window(area) as a face or non-face. If the mahalanobis distance of sub-window is lower than threshold, is a face area. We estimate whether test image is face or not by.

Fig. 3. Scale-Invariance by Pyramid

Fig. 4. The comparison of Mahalanobis distance of face and non-face image

The above experiment result mahalanobis distance is effective to face detection, but is sensitive in pose. Therefore, pose classification is important. In this paper, integrate FuzzyART and RBF for pose's classification.

3 Pose Classification Combining Fuzzy ART and BP

In the proposed approach, pose classification is dealt with as local targets to be detected in an image. We formulate the object detector generator to learn to discriminate object pattern witch can be measured formally, not intuitively as most previous approaches did. The pose estimator classifies a face pattern into one of the view (pose) set. The facial pose estimation can be used to verify the detection of multi-view faces. Gong and colleagues use non-linear support vector regression to estimate facial pose. In this paper, we propose how combined supervised learning and
unsupervised learning. The proposed learning based classifier generator has advantages over previous approaches:

### 3.1 Pose Estimation by FART+RBF

The RBF networks, just like MLP networks, can therefore be used in classification and/or function approximation problems. In the case of a RBF network, we usually prefer the hybrid approach, described below \[6,7,8,9,10,11\]. The RBFNs, which have a similar architecture to that of MLPs, however, achieve this goal using a different strategy.

One cluster center is updated every time an input vector \( x \) is chosen by FuzzyART from the input data set. The cluster nearest to \( x \) has its position updated using

\[
W_j(t+1) = \beta (I \land W_j(t)) + (1-\beta)W_j(t)
\]  

FuzzyART is a variant of ART system derived from the first generation of ART, namely ART1. It is a synthesis of ART algorithm and Fuzzy operator. ART1 can only accept binary input pattern, but FuzzyART allows both binary and continuous input patterns\[3,12\]. The feature space of object instance with multiple viewing angles must be clustered properly so that the location error can be minimized. However, the classification of multiple viewing points is very subjective and ambiguous. Thus, we adopt FuzzyART and RBF methods for achieving an optimal pose classification architecture. Executed step is as following to FuzzyART\[3,12,13,14\]. In this Paper, Clustering’s performance improves by studying repeatedly about done data.

The cluster center is moved closer to \( x \) because this equation minimizes the error vector. Each hidden unit calculates the distance of the input vector from the corresponding Gaussian:

\[
\phi_j(x) = \exp \left\{ -\frac{||x - \mu_j||^2}{2\sigma_j^2} \right\}
\]  

In this paper, centers are obtained from unsupervised learning (clustering), FuzzyART Algorithm. The weights between the hidden units and the output layer, denoted by \( w_{kj} \), are regular multiplication weights (as in a MLP).

\[
y_k(x) = \sum_{j=1}^{M} w_{kj} \phi_j(x) + w_k0
\]  

Where \( x \) is the input vector, \( m_j \) is the \( j \)th prototype vector, \( \sigma_j \) is the width of the Gaussian of that prototype or cluster centre.

There are various approaches for training RBF networks. In this paper, centers are obtained from unsupervised learning (clustering), FuzzyART algorithm. Clustering (FuzzyART algorithm) and LMS are iterative. This is the most commonly used procedure. Typically provides good results. After finding a suitable cluster using clustering algorithm, do laying center on this. The winning node \( \mu_j \) is what FuzzyART is its best match for the input pattern. Hidden node’s center determined by unsupervised learning, FART.
As showed Fig. 5, the idea is to train the network in two separate stages - first, we perform an unsupervised training (FuzzyART) to determine the Gaussians’ parameters ($\eta, \mu$). In the second stage, the multiplicative weights $w_{ij}$ are trained using the regular supervised approach. Input pattern is vectorized for grayscale image size of 20x20 pixels, input node had mosaic of size of 10x10 pixels. The transformation from the input space to the hidden unit space is non-linear, whereas the transformation from the hidden-unit space to the output-space is linear. RBF classifier expand input vectors into a high dimensional space. RBF network has architecture that of the traditional three-layer back-propagation. In this paper, hidden units is trained using FuzzyART network and basis function used are Gaussians. The proposed network input consists of n normalized and rescaled size of 1/2 face images fed to the network as 1 dimension vector. And input unit has floating value [0,1]. The vector value is normalized. In case learn by FuzzyART, performance is best in case used picture itself by input node vectorized.
Training images shows Fig. 7 and this is multi-view based face images, the face have left (-10, -20) face images and right (0, 10, 20) images.

![Yale Face Dataset](image)

**Fig. 7. Yale Face Dataset**

The advantages of the two-stage training are that it is fast and very easy to interpret. It is especially useful when labeled data is in short supply, since the first stage can be performed on all the data (not only the labeled part).

## 4 Experiment

The experiment of the proposed method for the face detection has been performed with images captured in various environments. 600 images are captured and considered in the experiments. The superiority of the proposed method is discussed in the following. We should note that to pose estimation, Fig. 8 shows face detection in real-time image by using multi-resolution.

![Face Detection Result in the Real-Time](image)

**Fig. 8. Face Detection Result in the Real-Time**

Face detection result shows table1. As explained in frontal image and the test images have size of 320 x 240 pixels and encoded in 256 gray scale levels. We resized to various size using multi-resolution of 9 steps. Rescaled images is transferred. Each face is normalized into a re-scaled size of 20x20 pixels and each data – training image and test images – is preprocessed by histogram equalization and max-min value normalization.
Table 1. Test images for face detection

| Total Image | Accept | FAR |
|-------------|--------|-----|
| 120         | 118    | 2   |

The clustering results of face pose using only FuzzyART and integration FuzzyART and RBF is given in the following Table 2 and Table 3. We used 1000 images by train data and 600 images by test data of MIT +UMIST database and ORL database. We separated face images through recursive learning of 4 times. Result about pose data classification shows in Table 2. By we were going to do studying repeatedly, clustering's performance improved.

Table 2. Pose Estimation Result-Laboratory DB

| Data          | RBFN+FART | FART   | BP    |
|---------------|-----------|--------|-------|
| Front Image   | 597/600   | 591/600| 497/600|
| Left Image    | 597/600   | 593/600| 498/600|
| Right Image   | 600/600   | 599/600| 485/600|

Table 3. UMIST & MIT Database

| Data          | Accept | FAR  |
|---------------|--------|------|
| Front Image   | 292/300| 8/300|
| Left Image    | 48/49  | 1/49 |
| Right Image   | 298/300| 2/300|

Table 4. ORL Database: 40 Person * 9 Pose

| Data          | Accept | FAR  |
|---------------|--------|------|
| Front Image   | 39/40  | 1/40 |
| Left Image (4 Pose) | 118/120 | 2/120 |
| Right Image(4 pose) | 117/120 | 2/120 |

As seen in Tables, error in front side face appears in tilted image. We must consider about tilted image. We could improve pose classification’s performance for face through recursive studying like experiment result. The combined effect of eye lasses is also investigated. In this experiment, the factor of glasses and illumination was considered and experimental images were classified by the factor of glasses and illumination. We classified bad illumination images into the image including a partially lighted face, good images into that including a nearly uniformly lighted face.
5 Concluding Remarks

In this paper, we propose a novel generic appearance-based object detection, especially for variance in scale and viewing points. Even though much research has been done for object detection, it still remains a difficult issue. The proposed system could accumulate the object models by autonomous learning process. In appearance-based approach, the object can only be approximated with many scales and views, where the partitioning of multiple scales and viewpoints is very subjective and ambiguous. A tradeoff representation of pose and its accuracy is treated with the FuzzyART and the RBF Network method. The feature space for object detection with multiple scale and viewpoints are partitioned into subspaces so that the location error can be minimized. mahalanobis distance computing module explores the subspaces of multiple viewpoints for an efficient face detecting structure, and integrated the FuzzyART and RBF module resolves the subjective and ambiguous problem of partitioning boundaries. The experiment shows very encouraging results.

Acknowledgement

The work was supported (in part) by Biometrics Engineering Research Center, (KOSEF).

References

[1] H. Murase and S. Nayar : Visual learning and recognition of 3D objects from appearance. Int'l Computer Vision, Vol.14, No.1. (1995) 5-24
[2] Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images by Components, IEEE Trans. on PAMI, vol. 23, No. 4. April (2001) 349-361.
[3] T.Kasuba : Simplified Fuzzy ARTMAP. AI Expert, November (1993) 18-25.
[4] J. Duckworth : What is chemometrics. In Proceedindgs of International Conferenceon Near Infrared Spectroscopy (1995)
[5] T.F. Cootes, G.J. Edwards, and C.J Taylor. Active appearance models. IEEE Trans. Pattern Analysis and Machine Intelligence, 23(6). July (2001) 681-685.
[6] Shakunaga, T. : An Object Pose Estimation System Using A Single Camera. Proc. of IEEE international conference, Vol.2. (1992) 7-10
[7] Paulino, A. Araujo, H. : Pose Estimation for central catadioptric system : an analytic Approach. Proc. Of Pattern recognition (2002) 969-699
[8] Huang, J. Shao, X Wechsler, H. : Face Pose discrimination using support vector machanism(SVM). Proc. of Pattern Recognition, Vol.1. (1998) 154-156
[9] M. H. Hassoun : Fundamentals of Artificial Neural Networks. MIT Press (1995)
[10] S. Haykin : Neural Networks -A Comprehensive Foundation (Second Edition). Prentice-Hall Inc. (1999)
[11] A.S. Pandya, R.B. Macy : Pattern Recognition with Neural Networks in C++. CRC Press (1996)
[12] G.A. Carpenter et al : Fuzzy ARTMAP : A neural network architecture for increcentral supervised learning of analog multidimensional maps. IEEE Trans. Neural Networks, Vol.3, No. 5. September (1992) 698-712
[13] Issam Dagher, Michael Georgiopoulos, Gregory L. Heileman, George Bebis: An Ordering Algorithm for Pattern Presentation in Fuzzy ARTMAP That Tends to Improve Generalization Performance. IEEE Trans. Neural Network, Vol.10, No.4, July (1999)

[14] Ramuhalli, P., Polikar, R., Udpa L., Udpa S.: Fuzzy ARTMAP network with evolutionary learning. Proc. of IEEE 25th Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP 2000), Vol. 6. Istanbul, Turkey (2000) 3466-3469
A Novel Image Preprocessing by Evolvable Neural Network

M.Y. Nam, W.Y. Han, and P.K. Rhee

Dept. of Computer Science & Engineering Inha University
Biometrics Engineering Research Center
Yong-Hyun Dong, Incheon, South Korea
{rera, parthia}@im.inha.ac.kr, pkrhee@inha.ac.kr

Abstract. This paper presents a novel and efficient preprocessing method to relieve the effect of changing illumination by restructuring itself under dynamically changing environments. It monitors situations and evolves its structure accordingly, stores its experiences in the form of artificial chromosomes. It performs adaptive preprocessing by reorganizing its structure using the knowledge in the chromosomes matched to an operation environment. Introducing the concept of combining situation-awareness using the evolvable neural network and the evolutionary computing using the Genetic algorithm, the proposed method not only achieves highly efficient preprocessing for object recognition in varying illumination environments, but also solves the time-consuming problem of the evolutionary computing method. The proposed method has been tested and applied successfully to the preprocessing of face images. Face images are in spatially well-defined object class, and the features of face images are represented by multiple fiducial points, each of which is described by the Gabor wavelet transform. The superiority of the proposed preprocessing method is proven by showing the improvements of object recognition accuracy of face dataset: our lab, the AR, and the Yale.

1 Introduction

Robust object recognition needs high invariance with regard to those variations viewpoint, viewing distance, illumination, etc. However, most object recognitions today can operate successfully only under strongly constrained images captured in controlled illumination environments[1,2]. Ultimate goal of object recognition technology should include recognizing objects under varying illuminations. Recently, several researchers have tried to attack the problem of changing lighting conditions[3,4]. In this paper, we will focus on a limited range of the general object recognition, however it can be readily extended to more general object recognition. We will deal with image objects the spatial boundaries of which can be well estimated in prior, called “spatially well-defined object classes” [5]. Face images are in the class of well-defined image objects, the spatial boundaries of which can be well estimated in prior. Liu and Wechsler [6] have introduced EP(Evolutionary Pursuit) for face image encoding, and have shown its successful application.

In this paper, the Gabor wavelet guided by an evolutionary approach has been employed to adapt the system for variations in illumination[7]. Even though the Gabor
wavelet provides some nice properties in representation of face images as discussed above, it cannot provide sufficiently reliable solution in changing environments such as variations in illumination and facial expression. In this paper, the image preprocessing is optimized under the evolutionary constraint of the evolutionary computing (Genetic Algorithm). The role of evolutionary computing is to construct optimal filter bank for the optimal preprocessing in order to achieve best performance in varying environment. Evolutionary computing is inspired by biological evolution process [8,9]. Evolutionary computing initiates a population which is a set of members. Each member is described by a vector called a chromosome. The fitness of each member is evaluated, and only a portion of the population is selected as the population of next generation. It is said survival of the fittest with analogy to biological evolution. The evolutionary computing approach can adapt the system for varying illumination environments. However, it is usually very much time-consuming in reaching an optimal system structure. In this sense, the evolutionary computing can’t be used alone in real time applications. We introduce a novel concept of the situation-aware preprocessing for solving this problem. The changes of illumination environment can be detected by either analyzing the input images. We assume that the illumination environment changes continuously. That is, we exclude an impulse style of illumination changes in this paper. The recognition system monitors the changing environment, the system adapts itself to the changing environment by restructuring its filter bank reflecting the monitored environment information.

2 Description of the Novel Preprocessing Architecture

In this paper, the evolvable neural network is employed for a brittleness object recognition under varying illumination. Adaptive and intelligent preprocessing is performed for improving image quality as much as possible using image filtering techniques under the control of the evolutionary computing method and the evolvable neural network. The image preprocessing from an input image is usually a crucial task in a robust object recognition system. Most previous approaches for the object recognition determine their filter or filter set at their design phases, and preserve them during operations. However, such approaches may show severe limitations under dynamically changing illumination environments. It is almost impossible or very difficult to design an optimal preprocessing at the design step considering all possible factors coming from illumination changes. We employ the strategy that the employed filter set is allowed to be reorganized dynamically during operation depending on illumination variations. The filter set is called as evolvable filter bank. Variations in illumination include lighting direction, brightness, contrast, and spectral composition. In this session, a novel preprocessing technology using the evolvable neural network and the GA (Genetic Algorithm) are proposed for solving a brittleness of an object recognition under varying illuminations as well as time-consuming problem usually incurred from the GA.

The Proposed Preprocessing Architecture

The image filter bank can be optimized using the evolutionary computing method as will be described in the section 3. The image filters employed here are the lighting
compensation, the histogram equalization, and the homomorphic filter, but can be readily extended to other image filtering methods [10]. The evolutionary part evolves the system structure for each environment and stores the corresponding structure chromosome in the chromosome knowledge base. The filter manager searches for the chromosome knowledge base for proper chromosome in a given situation or environment. The action part is restructured using the information stored in the searched chromosome, and an optimal action is prepared and performed for the monitored illumination environment. The proposed recognition system consists of the evolvable neural network for the situation-awareness (illumination environment monitoring and categorization here), the evolvable filter bank for evolvable image preprocessing, the classification module, and the chromosome knowledge base (see Fig. 1).

![Fig. 1. The proposed preprocessing combining the situation awareness and the evolvable filter bank](image)

3 The Situation-Awareness Using the Evolvable Neural Network

The evolvable neural network[11] analyzes and decides the category of a given illumination environment. The system searches for the best matched filter and feature chromosomes and reconstructs its structure using the information in the chromosomes. The changes of illumination environment can be decided by either analyzing the input images or monitoring the system performance. The evolvable neural network is trained by supervised learning. The illumination environments is divided into nine categories, modeled by distinguishing the brightness level (i.e., High, Medium, Low here) and the coarse lighting direction (Left, Front, Right here). But, it can be readily extended to any number of combinations.

3.1 Evolvable Neural Network

The evolvable neural network is described in Fig. 2. It is basically a back propagation neural network, where the interconnection of the evolvable neural network is restructured by the guide of the evolutionary computation (GA).

Each connection between neurons “i” and “j” is given an evolved weighting factor \( W_{ij} \) which takes the form of a binary fraction, expressed with typically 5 bits. The weights of the evolvable neural network should be real number between \( U_{\min} \) and \( U_{\max} \), and the value of weight chromosome is converted to the real number using the follow Eq. (1) and (2).

\[
\text{Binary string} = \lceil \log_2 \left( \frac{U_{\min}}{U_{\max}} \right) + 1 \rceil + 1
\]

\[
\text{Real number} = (\text{Binary string})10 \pi + U_{\min}
\]

![Fig. 2. The structure of the evolvable neural network](image)
for example, the 3 bit binary fraction 0.111 is equal to 0 + 0.5 + 0.25 + 0.125 = 0.875 in decimal notation.

These $W_{ij}$ weightings are also given an evolved sign (+/-). the signal $I_{ij}$ coming from neuron “i” to neuron “j” ranges in value between -1 and +1. this $S_{ij}$ value is multiplied by the weighting factor $W_{ij}$ of the connection that the signal travels down.

These $V=E+G$ value is input to sigmoid output function $F(v)$ where $F$ is of the from

$$F(V) = \frac{2.0}{(1.0 + \exp(-V))} - 1.0$$

which ranges between +1.0 and -1.0.

The fitness value of the evolvable neural network is calculated output error of back propagation neural network with Eq (4). Where $a$ and $b$ are random variables.

$$Fitness = \sum_{i=1}^{k} f'(n_i(t))(o_i - n_i(t))$$

$$f(n) = \frac{a}{n_{\text{max}} - n_{\text{min}}} (n - n_{\text{min}})$$

$$f(w) = \frac{b}{w_{\text{max}} - w_{\text{min}}} (w - w_{\text{min}})$$

It is determined $a$ and $b$, they are initialized value 0.5 in Eq (5) and (6). The global solution is calculated for the evolvable neural network and Neuron that gathering together of input connection strength is less than norm can have number of minimum node being erased and keeps performance. The weighted node is threshold.

### 3.2 Training of the Evolvable Network

We define nine category of illumination environments in the proposed illumination model by distinguishing the brightness level (i.e., high, medium, low here) and the coarse lighting direction (left, front, and right here) as shown in Fig. 2. This is the synthesized images using the nine category illumination model.

![Synthesized images using the image modeling](image)

Fig. 2. Synthesized images using the image modeling
We have tested three methods for illumination discrimination: the simple rule based discrimination (SR), the back propagation neural network based discrimination (BP), and the evolvable neural network based discrimination (E-NN). The training of evolutionary neural network is done 100 original face images accumulated in our lab, and 800 virtually generated mosaic face images using the image synthesis method described above. We assume that illumination variation in face images can be represented by the noise model. The modeling face image reflecting an brightness variation can be done by the additive, the multiplicative, and the hybrid functions [12]. Directional illumination variations are modeled by the sine and the cosine weight function.

The additive, multiplicative, and hybrid functions are defined in Eq (7), (8), and (9), respectively.

\[
A(x, y) = f(x, y) + a(x, y) \\
M(x, y) = m(x, y) \cdot f(x, y) \\
H(x, y) = m(x, y) \cdot f(x, y) + a(x, y)
\]

where \( m(x, y) \) is the operand value of the multiplicative function at \((x, y)\), the \( a(x, y) \) is the operand value of the additive function, and \( f(x, y) \) is the pixel value of the original image at \((x, y)\). Illumination compensation is use that don't know illuminant direction. but, If we know state of lighting, we may achieve improving filtering for image.

4 Evolvable Filter Bank

Fig. 3 shows that adaptive filter selection step for varying illumination environments using the evolutionary computing method. The filter bank consists of three filters and the interconnection between them.

![Evolution (Genetic Algorithm)](image)

**Fig. 3.** The training process of the evolvable filter bank

**Homomorphic filter:** The Homomorphic filter [13] reduces brightness and emphasizes contrast in frequency domain in order to improve a reflectance effect and decrease a lighting effect.

**Illumination compensation filter:** The illumination compensation filter is the high pass filter using the local brightness, which means the average of difference between the brightness of central pixel point and the brightness of total window [3].

Illumination compensation is performed by a ratio of original image to background illumination modeling function. The final compensated image is obtained by multiplying the ratio and weight value as shown in Eq. (10).

\[
G(x, y) = \max \left\{ \frac{I(x, y)}{f(x, y)} \times \omega_1, 0 \right\}
\]
**Histogram equalization**: To improve contrast of image, histogram equalization is used. Ultimate purpose of histogram equalization is to create histogram that have fixed distribution. As a result, dark image is known well and so light image gets dark a little and hold reasonable delivery value. This normalization is done with stored histogram by gray scale mapping function. Therefore if the distribution of gray level was biased to one direction or scaled value was not uniformly distributed, histogram equalization is a good solution for image enhancement.

In the evolution mode, the evolve filter bank accumulates the knowledge of optimal filter bank structure for each illumination environmental category as follows:

1. Given an image set and, its illumination environmental category,
2. Begin the filter bank evolution until a criterion is met, where the criterion is the filter fitness that does not improve anymore or the predefined maximum trial limitation is encountered as follows:
   
   1) Generate a new filter bank combination and its parameters.
   2) Evaluate the filter fitness function of the filter bank using the newly derived filter population in the geno space. If the criterion is met, go to Step 3.
   3) Search for the filter banks from the filter population that maximize the fitness function and keep those as the best chromosomes.
   4) Applying GA’s genetic operators to generate new population of filter bank. Go to Step 2.2).
3. Update the knowledge base of corresponding chromosome for the given illumination environmental category.
4. Do Step 1 to Step 3 for other image sets and environment categories.

The fitness function for the evolvable filter bank is given in Eq. (11).

\[ \text{Fitness} = 1 - \sum_{i=1}^{k} \left( \frac{\sum_{j} (T_i - \text{OrgT})^2}{m} - \frac{\sum_{j} (T_i - \text{NotT})^2}{k - m} \right) \]  

(11)

\( T \) : Result that passes filter about registered image
\( \text{OrgT} \) : T and registration ID among registered images are same image
\( \text{NotT} \) : Image that T and registration ID among registered images differ
\( k \) : Whole number of image that is registered
\( m \) : Number of image that is registered a person

Each illumination environment category has its own filter chromosome, respectively. The format of the filter chromosome is given in Table 1.

**Table 1.** The format of the filter chromosome

|   | Homomorphic filter |   |   |
|---|-------------------|--|--|
| 1 | Illumination filter |
| 2 | Histogram filter   |
| 3 | Homomorphic filter parameter |
|   | 1 2 3 4 5 6 7 8 9 10 |

The action mode of the evolvable filter bank performs the preprocessing task as follows:
1) Given an image and its environmental category from the situation-aware phase.
2) Search for the filter chromosome representing optimal filter combination and parameters corresponding the detected illumination environment (environmental category).
3) Perform the preprocessing using the optimized filter combination.

5 Experimental Results

The feasibility of the proposed preprocessing has been tested our lab, Yale [14], AR[15] data set. We used 1000 images of 100 persons from our lab data set, 330 images of 33 persons excluding 99 images of wearing sunglasses from AR face data set, 60 images of 15 persons from Yale Face DB. We first examine the discriminating performance of the evolvable neural network on various illumination environments using Yale and AR data sets. The performance test of proposed face recognition system is followed using our lab, Yale, AR data sets.

There are several related preprocessing for face recognition subproblem is general image filtering, adaptive image filter selection by using Genetic algorithm, illuminant environment analysis of Back Propagation, Statistical illumination evaluation, and illuminant environment analysis of Evolutionary Accumulated Knowledge and filtering using Genetic Algorithm.

In face recognition, we extract feature 25 feature points using Gabor transform and similarity measure use cosine distance.

![Fig. 4. Our Lab dataset](image)

5.1 Discrimination Performance of Illumination Environments

The performance of the evolvable neural network for discriminating illumination environments has been tested. We have compared three methods: the simple rule based discrimination(SR), the back propagation neural network based discrimination(BP), and the evolvable neural network based discrimination(E-NN). The training of evolutionary neural network is 100 original face images accumulated in our lab, and 800 virtually generated mosaic face images using the image synthesis method described in section 3.1.
We devise the following definitions for providing the decision rules of the simple rule based discrimination, training the back propagation neural network, and training the evolvable neural network. If the average pixel value of the 16 mosaic cells is 100 and 150. As a first test, we used 500 face images accumulated in our Lab which is not overlapped with the training data set, and the results of the experiment are given in Fig. 7.
The AR dataset offers face image groups of the left directional with bright illumination (LB), the right directional with bright illumination (LR), and the normal bright illumination (B). Each group includes normal face image, face wearing muffler, and face wearing sunglasses. We exclude the sunglasses wearing images in order to concentrate the problem of illumination variations. The remaining 6 types of face images are synthesized in this experiment. Fig. 8 shows the low performance of the proposed evolvable neural network for AR dataset. The overall performances is dropped those achieved from our lab data set due to the effect of wearing muffler.

![Fig. 8. Discrimination performance of illumination environment using the AR dataset](image)

In our experiment, we used the Yale data set containing single light source images of 15 people under 4 viewing conditions. Fig. 9 shows the testing performance for Yale dataset. The enhanced discrimination rates are achieved by the Yale data set comparing to the AR dataset since the Yale dataset has more clear images and no muffler wearing image is included. Major source of error is those images that do not belong to any defined illumination groups.

![Fig. 9. Discrimination performance of illumination environment use Yale dataset](image)

5.2 The Preprocessing Performance Evaluation via Object Recognition Accuracy

Object Feature Space

The characteristic vectors is generated from 25 fiducial points(see Fig. 10). Each fiducial points generates 40 vector elements, and thus total 1000 vector elements are generated.
The Accuracy of Object Recognition

Object recognition is performed in the action mode which consists of four phases: the situation-aware phase, the preprocessing phase, the Gabor feature space phase, and the class decision phase. The input image is analyzed into one of the environmental categories as described in the session 3. The input image is preprocessed by the restructured filter bank using the filter chromosome corresponding to the detected environmental category as described in the session 4. The preprocessed image is transformed into the Gabor feature vector. Finally, the class is decided by the class decision phase. A-nn (Approximate Nearest Neighbor) algorithm is employed for the recognition.

The first experiment for the face recognition is performed using the data set accumulated by our lab, AR dataset and Yale dataset.

Table 2, 3 and 4 show a recognition rate of proposed method. It is 97.98 % for our Lab DB, 96.97 % for AR dataset and 95.56 % for Yale dataset. The experimental result of proposed method shows the average recognition rate of 97% and a image filtering recognition rate of 93.06%. From Tables, it becomes apparent that selected image filter bank by genetic algorithm method shows good recognition performance while general illuminant filter single filter do. Evolvable neural network improved performance over the two during testing. This can interpret use existence and nonexistence and parameter of each image filter using genetic algorithm, because general filtering may appear result that flow image filter unconditionally, and drops preferably quality of original above zero because suitable parameter control is impossible.

Table 2. Our Lab dataset

| Data                  | Number of images | Number of person | Success | Reject | Successful rate |
|-----------------------|------------------|------------------|---------|--------|-----------------|
| Histogram Only        | 4500             | 100              | 4385    | 29     | 97.44%          |
| Illumination compensation | 4500         | 100              | 4170    | 39     | 92.66%          |
| Simple rule           | 4500             | 100              | 4395    | 28     | 97.66%          |
| BP                    | 4500             | 100              | 4319    | 42     | 95.98%          |
| ENN                   | 4500             | 100              | 4409    | 2      | 97.98%          |
A Novel Image Preprocessing by Evolvable Neural Network

Table 3. AR dataset

| Data                  | Number of images | Number of person | Success | Reject | Successful rate |
|-----------------------|------------------|------------------|---------|--------|-----------------|
| Histogram Only        | 198              | 33               | 165     | 13     | 83.33%          |
| Illumination          | 198              | 33               | 158     | 16     | 79.79%          |
| compensation          | Simple rule      | 198              | 33      | 175    | 7               | 88.38%          |
| BP                    | 198              | 33               | 182     | 2      | 91.92%          |
| ENN                   | 198              | 33               | 192     | 0      | 96.97%          |

Table 4. Yale dataset

| Data                  | Number of images | Number of person | Success | Reject | Successful rate |
|-----------------------|------------------|------------------|---------|--------|-----------------|
| Histogram Only        | 45               | 15               | 39      | 3      | 86.66%          |
| Illumination          |                  |                  |         |        |                 |
| compensation          | Simple Rule      | 45               | 15      | 43     | 0               | 95.56%          |
| BP                    | 45               | 15               | 41      | 1      | 91.11%          |
| ENN                   | 45               | 15               | 43      | 0      | 95.56%          |

6 Concluding Remarks

In this paper, we propose a novel adaptive and evolutionary technique for robust image preprocessing, especially for varying illumination environments. Much research has been done for object image enhancement, however, most existing technologies are vulnerable under changing illumination. The proposed method for image preprocessing can adapt itself to varying illumination environments. The major contribution of the proposed method is to achieve highly robust object recognition in varying environments by adapting evolving in both the preprocessing and feature representation steps. Furthermore, it solves the time-consuming problem of the evolutionary computing method by combining the evolvable neural network and the Genetic algorithm approaches. The whole object recognition system using our novel preprocessing method divides operation into the evolutionary and action modes. The evolutionary mode accumulates the knowledge from varying illumination environments, and the accumulated knowledge is stored in terms of artificial chromosomes. The action mode executes the task of identification using the preprocessing knowledge accumulated in the evolutionary mode. The proposed method has been tested to the popular object recognition problem, face recognition using four datasets: our lab, AR, and Yale. Extensive experiments shows that the performance of the proposed method is superior to those of most popular methods.

Acknowledgement

The work was supported (in part) by Biometrics Engineering Research Center, (KOSEF).
References

[1] Peter N.Belhumeur, David J. Kriegman, “What is the Set of Images of an Object Under All Possible Lighting Conditions?,” Proc.of the Computer Vision and Pattern Recognition 1996.

[2] Peter N.Belhumeur, Joao P. Hespanha, Davis J.Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection,” European Conference on Computer Vision.1996

[3] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From Few to Many: Illumination CONe Models for face recognition under Variable Lighting and Pose," IEEE Trans. on PAMI, vol. 23 no. 6, pp. 643-660, June 2001.

[4] C. Liu and H. Wechsler, "Evolutionary Pursuit and Its Application to Face recognition," IEEE Trans. on PAMI, vol. 22, no. 6, pp. 570-582, 2000.

[5] Sung, K.-K. and Poggio, T., "Example-based learning for view-based human face detection Pattern Analysis and Machine Intelligence," IEEE Transactions on , Volume: 20 Issue: 1, pp.39-51, Jan. “1998.

[7] M. Potzsch, N. Kruger, and C. Von der Malsburg, "Improving Object recognition by Transforming Gabor Filter responses," Network: Computation in Neural Systems, 7(2) pp. 341-347.

[8] J. H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[9] D. Goldberg, “Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[10] C. Liu and Harry Wechsler, “Evolutionary Pursuit and Its Application to Face Recognition,” IEEE Trans. on Pattern Analysis and Machin Intelligent, Vol.22, No.6, pp570~582, June 2000.

[11] Chengjun Liu and Harry Wechsler, “Evolutionary Pursuit and Its Application to Face Recognition,” IEEE Trans. on Pattern Analysis and Machin Intelligent, Vol.22, No.6, pp570~582, June 2000.

[12] A. Mohan, C.Papageorgiou, and T.Poggio, “Example-based object detection in images by components,” IEEE Trans. Paattern Anal. Machine Intell., vol.23, pp.349-361, Apr. 2001

[13] J. Jones and L. Palmer, “An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex.” J. Neurophysiology, pp. 1233-1258, 1987.

[14] L. Wiskott J.M.Fellous N.Kruger and Christoph von der Malsburg “Face Recognition by Elesic Graph matching”, In intelligent Biometric Techniques in fingerprint and face recognition , CRC Press, Chapter 11, pp355-396, 1999

[15] K. Eternad and R.Chellappa, “Discriminant Analysis for recognition of Human Face Images”.
Transition Properties of Higher Order
Associative Memory of Sequential Patterns

Hiromi Miyajima, Noritaka Shigei, and Yasuo Hamakawa
Kagoshima University, 1-21-40 Korimomto, Kagoshima 890-0065, Japan
miya@eee.kagoshima-u.ac.jp

Abstract. This paper describes some properties of storage capacities and robustness of higher order associative memory of sequential patterns. First, storage capacities of higher order correlation and differential correlation models are shown from the prediction using transition properties. Further, the robustness for the rate of correlation is shown. As a result, it is shown that the higher order differential correlation model is superior (inferior) in the robustness (storage capacity) to the higher order correlation model. Further, it is shown that higher order models are superior in the pattern selection ability to one dimension model.

1 Introduction

Many models processing static patterns such as auto-correlation associative memory, have been studied with associative memory[1,2]. Further, associative memory of sequential patterns is a very important problem in the cases where we construct associative memory as a model of the brain and associative memory of sequential patterns is desired to apply to various types of applications. Therefore, some models for associative memory of sequential patterns have also been proposed. Amari and Meier have shown that the storage capacity of the conventional correlation model is about $0.27N^{3}$, where $N$ is the number of neurons. Further, Yanai has shown the relation between the correlation for the patterns and the storage capacity in one dimension correlation and differential correlation associative memory[4]. In general, it is well known that higher order neural networks are a generalized model of the conventional neural networks whose potential is represented as the linear sum of weights and input vector and are effective in many applications than the conventional ones. From this point, it is suggested that higher order neural networks have higher ability than the conventional ones even in associative memory of sequential patterns. Hence, we have proposed generalized models to higher order ones and shown the transition properties[5][6]. This paper describes some properties of storage capacities and robustness for higher order associative memory of sequential patterns. As a result, it is shown that the higher order differential correlation model is superior (inferior) in the robustness (storage capacity) to the higher order correlation model. Further, it is shown that higher order model is superior in the pattern selection ability to the one dimension model.
2 Higher Order Associative Memory of Sequential Patterns

2.1 Higher Order Neural Networks

The internal potential for higher order neural element with the order $k$ is represented by

$$u_i(t) = \sum_{k'=1}^{k} \sum_{[l_{k'}]} v_{i[l_{k'}]} x_{l_1}(t) x_{l_2}(t) \cdots x_{l_{k'}}(t) + bx_i(t) - \theta_i,$$

(1)

$$\sum_{[l_k]} = \sum_{l_1} \sum_{l_2} \cdots \sum_{l_k},$$

(2)

$$l_{a-1} + 1 \leq l_a \leq N - k + a,$$

(3)

where $[l_k] = l_1, \cdots, l_k$, $v_{i[l_k]}$ is the weight for products of input $x_{l_1}, \cdots, x_{l_k}$, to the $i$-th neuron, $a = 1, \cdots, k$, $l_0 = 0$, $\theta_i$ is the threshold of the $i$-th neuron, and $b$ is the weight for self-loop of each neuron. In this paper, the following model is used:

$$u_i(t) = \sum_{[l_k]} v_{i[l_k]} x_{l_1}(t) \cdots x_{l_k}(t) + bx_i(t) - \theta_i,$$

(4)

$$x_i(t+1) = \text{sgn}(u_i(t)),$$

(5)

where $\text{sgn}(u) = 1$ for $u \geq 0$ and $-1$ for $u < 0$. A higher order neural network is composed of $N$ higher order neural elements mutually connected.

2.2 Memorizing and Recalling of Sequential Patterns

Let us consider sequential patterns as follows:

$$S^1 \rightarrow S^2 \rightarrow \cdots \rightarrow S^P \rightarrow S^1 \rightarrow \cdots,$$

(6)

where $S^\mu = (s^\mu_1, \cdots, s^\mu_N)^T$ ($\mu = 1, \cdots, P$) and $s^\mu_i = +1$ or $-1$. Each element of the patterns is selected randomly. For sequential patterns, each weight of higher order associative memory models is defined as follows:

$$v_{i[l_k]} = \begin{cases} 
\frac{1}{k} \sum_{\mu=1}^{P} (s^\mu_i - \bar{s})(s^\mu_{l_1} - \bar{s}) \cdots (s^\mu_{l_k} - \bar{s}) & \text{The correlation model} \\
\frac{1}{k} \sum_{\mu=1}^{P} (s^\mu_i - \bar{s})(s^\mu_{l_1} - s^\mu_{l_1-1}) \cdots (s^\mu_{l_k} - s^\mu_{l_k-1}) & \text{The differential correlation model}
\end{cases}$$

(7)

Let us consider the rough transition properties of the models shown in the Eq. (7). If an input pattern $X(0)$ similar to the memorized pattern $S^\nu$ is given,
then the internal potential of \( X(1) \), which is the transition pattern of \( X(0) \), is obtained as follows:

\[
\begin{align*}
    u_i(0) &= \begin{cases} 
        s^{\nu+1}_i + b \cdot s^{\nu}_i + N_i - \theta_i & \text{The correlation model} \\
        2s^{\nu+1}_i - s^{\nu}_i - s^{\nu+2}_i + bs^{\nu}_i + N_i - \theta_i & \text{The differential correlation model,}
    \end{cases}
\end{align*}
\]

where \( N_i \) is the crosstalk term for the pattern \( S^{\nu+1} \). If the absolute of \( N_i \) is sufficiently small, \( \theta_i = 0 \), and \( b = 0 \) for the correlation model and \( \theta_i = 0 \), \( b = 1 \) for the differential correlation model, then \( x_i(1) = s^{\nu+1}_i \). That is, recalling of sequential patterns is performed. Specifically, if sequential patterns are mutually orthogonal, perfect recalling is performed. It means that the models can realize associative memory.

In order to get the transition properties of the networks, three assumptions for the models are made as follows:

1. Each element \( s^{\mu}_i \) of the sequential patterns is as follows:
   \[
   \Pr\{s^{\mu}_i = -1\} = p,
   \]
   where \( \Pr\{\cdot\} \) is the probability of the event \( \{\cdot\} \). If \( p \neq 0.5 \), each pattern is correlative to each other.

2. \( P \) and \( N \) are sufficiently large.

3. All states \( s^{\mu}_i \)s with different values of \( i \) and \( \mu \) are mutually independent.

Let \( \bar{s} = 1 - 2p \). The value \( \bar{s} \) means the rate of correlation. For example, if \( p = 0.5 \), \( p = 0.4 \) and \( p = 0.3 \), then \( \bar{s} = 0 \), 0.2 and 0.4, respectively.

Let the pattern ratio be defined as follows:

\[
    r_k = \frac{P}{\binom{N}{k}},
\]

where \( k \) is dimension. The pattern ratio means the ratio of the number of memorized patterns per one weight. The storage capacity of the model is defined as critical pattern ratio. It means how many patterns are memorized in the model.

In this paper, the following transition properties of the models are considered under the assumptions 1, 2 and 3:

\[
    d_t = \frac{1}{N} \sum_{i=1}^{N} s^{\mu+t}_i x_i(t),
\]

where \( S^{\mu+t} = (s^{\mu+t}_1, \cdots, s^{\mu+t}_N) \), \( X(t) = (x_1(t), \cdots, x_N(t)) \). Then \( d_0 \) means the distance between the pattern \( X(0) \) and the memorized pattern \( S^{\nu} \) and, \( d_1 \) means the distance between the pattern \( X(1) \) and the memorized pattern \( S^{\nu+1} \) and so on.

In this paper, let \( \theta_i = 0 \) without loss of generality, and the variable \( t \) as step is neglected if there does not exist any misunderstanding.

\[1\] As shown later, the Eq.8 for the differential correlation model holds only for the odd number \( k \).
3 Some Properties of the Proposed Models

3.1 Transition Properties of the Models
First, let us consider the transition property $d_1$ under the assumptions 1, 2 and 3. Let $X(0)$ be an input pattern similar to the memorized one $S^\nu$. Then, we can get the result by computing the internal potential of $X(1)$, which is the transition pattern of $X(0)$. The following relation holds for the correlation model \[5\]:

$$d_1 = \frac{1}{4} \sum_{\alpha, \beta \in \{-1, 1\}} (1 + \alpha \bar{s})(1 + \beta d_0) \times \Phi \left( \frac{(1 - \alpha \bar{s})d_0^k + \alpha \beta \bar{b}}{\sqrt{1 - \bar{s}^2}^{1-k} r_k} \right),$$  \hspace{1cm} (12)

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} \exp(-t^2/2) dt, \hspace{1cm} (13)$$

where $s_{i+1}^\nu = \alpha$, $x_i(0) = \beta$, $\bar{b} = b/(1-\bar{s})^2$.

From the Eqs. \[4\] and \[7\] we can get the following result for the differential correlation model \[6\]:

$$d_1 = \frac{1}{8} \sum_{\alpha, \beta, \delta \in \{-1, 1\}} (1 + \alpha \bar{s}^2)(1 + \beta \bar{s}^2)(1 + \delta d_0) \times \Phi \left( \frac{\{2 - \alpha(1 + \beta)\}d_0^k + \bar{b}\alpha\delta}{\sqrt{1 - \bar{s}^2}^{1-k} r_k 2^{k+1}} \right)$$

if $k$ is odd,

$$d_2 = \frac{1}{4} \sum_{\alpha, \delta \in \{-1, 1\}} (1 + \alpha \bar{s}^2)(1 + \delta d_0) \times \Phi \left( \frac{(1 - \alpha)d_0^k + \bar{b}\alpha\delta}{\sqrt{1 - \bar{s}^2}^{1-k} r_k 2^{k+1}} \right)$$

$k$ is even.  \hspace{1cm} (14)

The even case of the Eq. \[14\] does not always hold theoretically, because the assumption 3 does not hold in the transition of two steps. Hence, we perform numerical simulation of recalling of $S^{\nu+1}$ formally and compare the theoretical result with one of numerical simulation.

3.2 Numerical Simulations
For the correlation and the differential correlation models, we have already shown that the results in numerical simulations are of fairly general agreement with the theoretical ones of the Eq. \[12\] and Eq. \[14\]. For example, Fig.1 shows the agreement of results of numerical simulation (wavy curve) and theoretical results (smooth curve). With $k = 1$ and 3, the same results are obtained. Further, from the comparison among the graphs for $k = 1, 2$ and 3, we made clear that higher order models are superior in the pattern selection ability to one dimensional model.

From the transition properties, we have predicted the storage capacities for both models as following method: First, let us assume that the case where $d_1$ for
$d_0 = 0.93$ in the Eq.[14] (see Fig.1) is greater than 0.93, is successful in recalling. Further, let us define that the storage capacity of the model is $r_k$ for the critical case in these successful cases. The Table 1 shows the result for each model. The results show that as the dimension increases, the storage capacity of models becomes low. However, given the number $N$, it holds $0.059N$ for $k = 1$ and $0.023\left(\frac{N}{2}\right)$ for $k = 2$. For example, if $N = 100$, then the numbers of memorized patterns are about 6 for $k = 1$ and about 114 for $k = 2$. It shows that higher order models are superior in the number of memorized patterns to one dimensional case.

3.3 The Robustness of the Model

In order to consider the robustness of the proposed model, let us compute the probability of the sum of $Pr\{u_i > 0\}$ when $s_i^{\nu+1}(s_i^{\nu+2}) = 1$ and $Pr\{u_i < 0\}$ when $s_i^{\nu+1}(s_i^{\nu+2}) = -1$ for $d_0 = 1$. It means the probability recalling the pattern correctly after one step when one memorized pattern is input. (i) The case of correlation model

$$
\frac{1 + \bar{s}}{2}\Psi\left(\frac{(1 - \bar{s}) + \bar{b}x_i(0)}{\sqrt{(1 - \bar{s}^2)^{1-k}r_k}}\right) + \frac{1 - \bar{s}}{2}\Psi\left(\frac{(1 + \bar{s}) + \bar{b}x_i(0)}{\sqrt{(1 - \bar{s}^2)^{1-k}r_k}}\right),
$$

where $\Psi(u) = \frac{1}{2}(1 + \Phi(u))$
Fig. 2. Probability of associative memory for correlation $\bar{s}$ of sequential patterns

Table 2. Some properties on higher order associative memory of sequential patterns

|                   | The correlation model | The differential correlation model |
|-------------------|-----------------------|-----------------------------------|
| Transition Property| Eq.12                 | Eq.14                             |
| Storage Capacity   | High                  | Low                               |
| Robustness         | Low                   | High                              |

(ii) The case of differential correlation model

\[
\frac{1 - \bar{s}^2}{4} \psi\left(\frac{4 - \bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right) + \frac{1 + 3\bar{s}^2}{4} \psi\left(\frac{\bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right)
\]

\[
+ \frac{1 - \bar{s}^2}{4} \psi\left(\frac{2 + \bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right) + \frac{1 - \bar{s}^2}{4} \psi\left(\frac{2 - \bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right)
\]

if $k$ is odd.

\[
\frac{1 - \bar{s}^2}{2} \psi\left(\frac{2 - \bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right) + \frac{1 + \bar{s}^2}{2} \psi\left(\frac{\bar{b}}{\sqrt{(1 - \bar{s}^2)1 - k2^k + 1\tau_k}}\right)
\]

if $k$ is even.

Fig. 2 shows the result of recalling probabilities for the rates $\bar{s}$ of correlation. It shows that the case of the differential correlation model is fixed for $0 \leq |\bar{s}| \leq 0.4$ but the case of correlation model is not. Fig. 3 shows the result of recalling probabilities for the memory ratio $\tau_k$. It shows that the differential correlation model is superior in robustness to the correlation model. Table 2 shows the relation between the results of both models.
4 Conclusions

This paper describes some properties of storage capacities and robustness of higher order associative memory of sequential patterns. First, storage capacities of higher order correlation and differential correlation models are shown from the prediction using transition properties and simulation results. Further, the robustness for the rate of correlation was shown. As a result, it has been shown that the higher order differential correlation model is superior (inferior) in the robustness (storage capacity) to the higher order correlation model.

References

1. J. Hertz, A. Krogh, and R. G. Palmer, Introduction to the Theory of Neural Computation, Perseus Books Publishing, 1991.
2. C.T. Lin, and C.S.G. Lee, “Neural Fuzzy Systems”, Prentice Hall PTR, 1996.
3. S. Amari, “Statistical Neurodynamics of Various Versions of Correlation Associative Memory”, Proceedings of IEEE conference on Neural Networks, pp.I-633–I-640, 1988.
4. H. Yanai, and Y. Sawada, “On Some Properties of Sequence-Association Type Model Neural Networks”, IEICE Trans., Vol.J73-D-II, No.8, pp.1192–1197, 1990 (in Japanese).
5. Y. Hamakawa, H. Miyajima, N. Shigei and T. Tsuruta, “On Some Properties of Higher Order Correlation Associative Memory of Sequential Patterns”, Journal of Signal Processing, vol.8, no.2, 2004 (in print).
6. H. Miyajima, N. Shigei and Y. Hamakawa, “Higher Order Correlation Associative Memory of Sequential Patterns”, IJCNN 2004 (in print).
Morphological Blob-Mura Defect Detection Method for TFT-LCD Panel Inspection

Young-Chul Song, Doo-Hyun Choi, and Kil-Houm Park

School of Electrical Engineering and Computer Science
Kyungpook National University, Daegu, Korea
{songyc03,dhc,khpark}@ee.knu.ac.kr

Abstract. The current paper proposes a blob-Mura defect detection method for TFT-LCD panel inspection. A new constraint function that can grow and shrink is defined. Specially, a morphology-based preprocessing method is proposed to improve the detecting capacity of a blob-Mura-defect-detecting algorithm, whereby a test image with blob-Mura defects is expanded to facilitate the defect detection. Plus, in the case of defects with a diameter over 49 pixels, which are hard to detect due to the non-uniformity of their interior, the proposed method changes the image size instead of the constraint function size. The proposed method enables superior defect detection and the use of a simple detecting algorithm.

1 Introduction

Detecting blob-Mura defects in a TFT-LCD panel can be difficult for two main reasons [1]: first, the defect interior or background has a non-uniform brightness level, and second, there can be small-sized blob-Mura defects with only slightly different brightness levels between the blob-Mura defect region and the background region. In the former case, a preprocessing method, such as smoothing, is needed to remove the non-uniformity of the defect interior. Without such a process, it is difficult for a blob-Mura-defect-detecting algorithm to extract a defect perfectly. Meanwhile, in the latter case, if the brightness difference between a defect and its neighbor is very small, such defects can be hard to detect, and even more difficult in the case of small-sized defects. As such, an algorithm for detecting blob-Mura defects in a TFT-LCD panel must be able to overcome these problems. Accordingly, this paper uses an area scan camera to create a low-resolution image for blob-Mura defect detection. The obtained images had 400\(\mu\)m spatial resolutions for each pixel and 8-bit brightness resolutions. Fig. 1 shows the flow chart for detecting blob-Mura defects.

2 Morphology-Based Preprocessing

This paper proposes a morphology-based preprocessing method to solve the above mentioned problems. Fig. 2(a) shows a test image including circular-shaped blob-Mura defects, where 24 defects with diameters ranging from 5~15
pixels and a large-sized defect with diameter 150 pixels have been created by a signal generator. Meanwhile, Fig. 2(b) shows a real image including blob-Mura defects. These defects have various differences between the defect and the background ranging from 3 to 15 pixels. As shown in the lower part of Fig. 2(a), the small-sized defects cannot be seen in detail (first and second rows in the lower part of Fig. 2(a)), which inevitably becomes more extreme if the brightness difference between a defect region and its surrounding region is very small. Therefore, to solve this problem, an erosion process, which is a morphology-based operation, is applied to the blob-Mura defects in Fig. 2(a). In this paper, the brightness values of the defect region are assumed to be darker than those of the surrounding region. Thus, dark blobs can be expanded by an erosion process [2]. Note that the proposed method can also be applied in the opposite case such as white blobs. However, in this case, dilation is required to expand the defect.
Fig. 3. Example of morphology-based preprocessing method, (a) Lower part of Fig. 2(a), (b) after application of erosion, (c) detected defects, (d) after application of erosion

Fig. 3(a) is to present the lower part of Fig. 2(a). Fig. 3(b) shows the result of applying erosion five times to Fig. 3(a). Some of the advantages of erosion, as shown in Fig. 3(b), include the removal of non-uniformities from the defect interior or background, due to the filtering effect of the erosion, and easy detection of the expanded defects using a simple detecting algorithm.

3 Blob-Mura Defect Detection

3.1 CSCF

As shown in Fig. 2, since the appearance of a regular and irregular shaped blob-Mura defect is similar to that of a circle, in this paper, the circular-shaped constraint functions (CSCF) are defined and used to detect blob-Mura defects with different sizes and brightness levels. Initially, smaller CSCFs are used to detect smaller defects, then gradually larger CSCFs are applied to detect larger defects. This method is very efficient for detecting blob-Mura defects with diameters ranging from 5~15 pixels, which is assumed to include most blob-Mura defect sizes. As such, this means that most defects can be detected by changing the CSCF size in sequence. As shown in Fig. 4(b), CSCF is defined as

$$m = l \cdot (1 + \alpha), \quad F = \left( \sum_i D_i + \sum_j B_j \right) / m$$

(1)

where $l$ is a diameter of defect, $\alpha$ is the weighing factor controlling the size of the background region, and $F$ is the brightness value of the CSCF interior in Fig. 4(b). For this study we chose to use an isotropic Gaussian function centered on the seed point location $O(\mu_x, \mu_y)$ with a variance $\sigma_i^2$ as CSCF. Note that various CSCF sizes can be obtained, such as $n_i = \{5, 7, 9, \ldots, 53\}$, by changing variance $O(\mu_x, \mu_y)$. To detect a blob-Mura defect, the brightness values between
the input image \( f(x, y) \) and \( F \) are compared. Thus, the current coordinate \((x, y)\) is regarded as a blob-Mura defect when \( f(x, y) \) is less or equal to \( F \). Most defects can be detected, as they satisfy the condition \( \max(D) \leq F \leq \min(B) \). Here, \( D \), \( F \), and \( B \) mean the brightness values of the defect, the kernel, and the background, respectively. However, if the area of the background region is much larger than that of the defect region, the relationship of \( \max(D) \leq F \leq \min(B) \) can not always be guaranteed. In this paper, this problem is solved by fixing \( \alpha \) at \( 3/5 \), then, if a defect is enclosed or perfectly matched by a CSCF, it can be detected, regardless of its shape. Here, a perfect matching means that at least three of \( X_1, X_2, Y_1, \) and \( Y_2 \) (refer to Fig. 5) are equal to \( n/2 \). Here \( n \) is the CSCF size. It is important to note that defects over 49 pixels in diameter (the largest circular-shaped defect in Fig. 2(a)) can be detected by reducing the image size by 1/8, instead of the CSCF size. As a result, the reduced defect can be detected using the same processes mentioned above based on changing the CSCF size.

### 3.2 Changeable Seed Point and Expansion of CSCF

In this paper, an adaptive multilevel-threshold method is used to extract a seed point, as employed in previous research by the current authors [3]. As shown in Fig. 5, the extracted seed point coordinate \( O(\mu_x, \mu_y) \) should be replaced with \( O'(\mu_x, \mu_y) \), when the following conditions are satisfied: First, horizontal shift: \((X_1 \text{ is equal to } n/2 \text{ and } X_2 \text{ is less than } n/2) \text{ or } (X_1 \text{ is less than } n/2 \text{ and } X_2 \text{ is equal to } n/2)\). Second, vertical shift: \((Y_1 \text{ is equal to } n/2 \text{ and } Y_2 \text{ is less than } n/2) \text{ or } (Y_1 \text{ is less than } n/2 \text{ and } Y_2 \text{ is equal to } n/2)\). Here, \( n \) is the CSCF size, as shown in Fig. 4(b). In the process of detecting a defect, if larger CSCFs are required, the detecting algorithm is repeated with a new seed point \( O'(\mu_x, \mu_y) \) determined as:

\[
\begin{align*}
    \text{Shift}_x &= X_1 - X_2, \\
    \text{Shift}_y &= Y_1 - Y_2 \\
    O'(\mu_x, \mu_y) &= O(\mu_x - \text{Shift}_x, \mu_y - \text{Shift}_y)
\end{align*}
\]
The changing of the seed point is performed at the same time as changing the CSCF size, and stops with perfect matching. However, since it is very difficult to determine whether the detecting process should be repeated or ended for a specific defect. The current paper solves this problem by providing a decision criterion. Once perfect matching has occurred, the detecting algorithm is then applied to the neighboring four directions of $O'(\mu_x, \mu_y)$, i.e. $O'(\mu_{x-1}, \mu_y)$, $O'(\mu_{x+1}, \mu_y)$, $O'(\mu_x, \mu_{y-1})$, and $O'(\mu_x, \mu_{y+1})$. Here, if perfect matching occurs in more than one direction, the detecting process is repeated, otherwise the detecting process is ended.

3.3 Non-uniform Defect

A morphology-based preprocessing can’t perfectly remove the non-uniformities in the defect interior. After all, this can create problems for the defect-detecting algorithms. As such, with the proposed method, if there is a non-uniformity in the defect interior, even when the CSCF is perfectly shifted to the defect interior, the detected defect area may be smaller than the CSCF area. Thus, in effect, a larger CSCF is required to detect a defect that is equal to the current CSCF size. Note that this case is only considered when at least three of $X_1$, $X_2$, $Y_1$, and $Y_2$ are less than $n/2$. To solve this problem, a criterion is presented as follows:

$$\text{ratio} = \frac{n-2}{n} \times 100\%$$

$$\text{repeat} = \begin{cases} 
\text{Yes}, & \text{if } l_D < w(n-2) \\
\text{No}, & \text{otherwise}
\end{cases}$$

where $$w = \begin{cases} 
1.0, & \text{if } \text{ratio} < 67 \\
0.7, & \text{otherwise}
\end{cases}$$

where $n$ is the CSCF size, $w$ is the weighting factor used to determine whether or not the detecting process is repeated, $l_D$ means the detected defect size, and $\text{ratio} \leq 67$ includes defects with a diameter of less than 11 pixels. Simulations verified that a defect can be detected by a CSCF with an identical size or at most two sizes larger.
## 4 Experimental Results

Fig. 2(c) shows the result after applying the proposed method to Fig. 2(b). Due to the expanded defect size and removal of non-uniformity by erosion, the smaller and larger-sized defects were almost all perfectly detected. In addition, the defects with a small brightness difference between the defect and the background were also well detected. Fig. 2(d) shows the result after applying erosion five times to Fig. 2(c). Although the morphology-based expansion and shrinking are irreversible, as shown in Fig. 2(d), the detected defect sizes were almost equal to the original defect sizes in Fig. 2(a). Thus, it was verified that the proposed a morphology-based preprocessing improved the detecting capacity of a blob-Mura-defect-detecting algorithm in a TFT-LCD panel.

Figs. 6(a) and (b) show the detection results for the artificial blob defects (upper part of Fig. 2(a)) and real Mura defects (Fig. 2(b)). As shown in Fig. 6(a),

![Image](image1.jpg)

**Fig. 6.** Resulting images for the upper part of Fig. 2(a) and Fig. 2(b). (a) Artificial blob-Mura defects, (b) real blob-Mura defects

![Image](image2.jpg)

**Fig. 7.** Blob-Mura defect detection verifying capability to shift seed point and change kernel size. (a) Real blob-Mura defects, (b) detected blob-Mura defects
the large-size defect with a badly non-uniform interior was accurately detected by reducing the image size instead of the CSCF size, confirming the ability of the proposed method to detect larger-sized defects, even if its shape is not smooth due to magnification by linear interpolation. Thus, the new defined CSCF based on the shape of the defect was found to be very effective for detecting blob-Mura defects with regular or irregular shapes (similar to a circle), different brightness levels, and different sizes.

Fig. 7(a) shows a real Mura defect with an approximate diameter of 42 pixels, with the resulting image represented in Fig. 7(b). Here, the seed points were manually determined to verify the capability of changing the seed point and changing the CSCF size. As shown in Fig. 7(b), the two proposed methods were successful in detecting real Mura defect.

5 Conclusion

The proposed blob-Mura defect detection method can overcome the two main difficulties encountered when detecting defects in an LCM panel: a non-uniform brightness in the defect interior or background and small brightness difference between the defect and the background in the case of small-sized defects. Specialy, by employing a morphology-based preprocessing, the proposed method could more improve the detecting capacity of a blob-Mura-defect-detecting algorithm in a TFT-LCD panel.

References

1. Kim, J.H., Ahn, S., Jeon, J.W., Byun, J.E.: A High-speed High-resolution Vision System for the Inspection of TFT LCD, Proceedings. ISIE 2001. IEEE International Symposium, 1, (2001) 101 -105.
2. Jain, R., Kasturi, R., Brain G.S.: Machine Vision, McGrawHill (1995).
3. Oh, J.H., Kwak, D.H., Song, Y.C., Choi, D.H., Park, K.H.: Line Defect Detection in TFT-LCD Using Directional Filter Bank and Adaptive Multilevel Thresholding, 11th APCNDT, (2003) 61.
A Recommendation System for Intelligent User Interface: Collaborative Filtering Approach

Ju-Hyoung Yoo, Kye-Soon Ahn, Jeong Jun, and Phill-Kyu Rhee

Dept. of Computer Science and Engineering Inha University
Biometrics Engineering Research Center
Yong-Hyun Dong, Incheon, South Korea
yjh9270@hanmail.net, {ahn, pkrhee}@inha.ac.kr

Abstract. We present a framework of recommendation system by organize users into different data groups and performing collaborative filtering on each groups to overcome problems that traditional recommendation systems have. Extensive experiment shows that recommendation system can observe user's behavioral characteristics better than previous approaches and can provide more accurate recommendation.

1 Introduction

As content on the Web increases rapidly, it is difficult to find contents appropriate to each user's interests [1][7][11]. Recently, much research for recommendation expert systems that find useful content to each user's needs is carried out to solve these problems [15]. Recommendation systems provide recommendations to potential customers or users.

There are two major approaches for providing recommendation service: content-based approaches [10][8][9][13][6] and collaborative filtering approaches [5][18][14][11][17]. Content-based approaches select content based on the correlation between user's preference and contents. It recommends contents that are similar to what user has been interested in his/her past. However, content-based filtering approaches have several drawbacks. Most contents are represented by multimedia type information, so it is difficult to analyze contents to be used for the content-based method. Some domains such as multimedia data type item (sound, image, video, etc.) are not amenable to any useful feature extraction methods with current technologies. This method can only recommend content items scoring highly against user's profile, the user is restricted to observe items similar to those already rated [10][7][9][13][6]. Therefore, it is necessary to develop improved recommendation system to overcome mentioned drawbacks of traditional systems.

Collaborative filtering approaches have been proposed to solve the problems that content-based filtering method has [5][18][14][12][16]. Main idea of this method is to automate the process of “Word of Mouth”. Collaborative filtering approaches can
overcome the weak points of content-based filtering that content of the items don't have to be analyzed.

Most collaborative filtering approaches gather user’s ratings from available content. They provide users content that similar users have interests. Similarities are measured from the statistical or probabilistic analysis of whole users. However, collaborative filtering has a weak point in finding similar users only based on their rating behaviors and relevant statistical or probabilistic analysis. It doesn’t reflect common characteristics of users (age, gender and so on) to measure similarities among them.

In this paper, we present a framework of recommendation system by organize users into different data groups and performing collaborative filtering on each groups to overcome problems that traditional ones have. Recommendation system can see user more meaningfully and offer useful recommendation.

2 Related Work

The content-based approaches in the information retrieval (IR) society have been adapted to recommendation systems [10][8][9][13][6]. Text documents are recommended based on a comparison between their content and a user profile. The weights for the words extracted from the document are added to the weights for the corresponding words in the user profile, if the user is interested in a page. Example of such systems are WebWatcher [6], NewsWeeder [9], InfoFinder [7], and client side agent Letizia [10]. This approach has several shortcomings. Generally, in some domains, the items are not applicable to any useful feature extraction methods with current technology such as motion pictures, music, etc. Also, the system can only recommend items scoring highly against a users’ profile, the user is restricted to see items similar to those already rated [17].

The collaborative filtering usually requires users to explicitly input ratings about a pieces of information. recommendations are based on the similarity among users. This approach does not consider any analysis of content. This characteristic enables collaborative filtering based system to naturally be applied to such domains as music, image, sound, etc. Examples of such systems include Tapestry [5], Gingo [19], Grouplens [12] and so on.

As the first collaborative filtering system, Tapestry helps user to co-work efficiently within the group [5]. This system delivers mail to mailing lists and users rate the information by annotations, numeric rating values and Boolean rating values. By using content-based filtering and subjective rating, it provides links, items, rating values, etc. to small number of users. If users want to be recommended the items, the users explicitly annotate what they need in SQL-like language, TQL. However, Tapestry has several drawbacks. It provides no mechanism for determining similar users. It requires users to actively annotate all information they receive to be effective. It only applies to small number of users. It adopts difficult language to use.

Ringo is a music recommendation system based on user profiles generated by explicit feedback information on artists through Web and email [19]. The term ”Social Information Filtering” was first introduced in Ringo system. It considers explicit user
ratings on a scale of 1 to 5 and find nearest neighbors by employing constraint Pearson r correlation coefficient algorithm. It decides any pair of users with a similarity more than a given threshold to be close, and predictions are generated using weighted average of the predictions made by the user's neighbors. GroupLens is a distributed system that collects, supplies, and utilizes users ratings to predict other users' preference [12]. This system includes "Better Bit Bureaus" that collects user ratings and predicts as well as news reading client. The goals of GroupLens design are providing ways of integrating with the traditional news reading clients, convenient rating scheme and prediction of user rating. It makes a user rate new or check time spending for user to read the news to analyze user's preference. It adopts the Pearson r algorithm to calculate user similarities.

3 Recommendation System

The proposed recommendation system expands traditional collaborative filtering system by applying data grouping. First, it organizes users into different data groups using users' common characteristics (age, gender and so on). Second, it performs collaborative filtering on each group and verifies the fitness for the group by performance of collaborative filtering. If it is below the threshold, the group is improper, recommendation system organize users in groups that satisfy threshold. Finally, recommendation list that users prefer is generated.

![Figure 1. Recommendation Engine Architecture](image-url)
3.1 Data Partitioning

Data partitioning in database creates a set of tuples that satisfy specific condition [16]. For example, tuples of user table inside database can be classified by distinction of sex or age attribute. A traditional collaborative filtering uses rating information of item to search for similar user group. In this case, similar user group can’t reflect users' inclination to select item. User’s rating for item appears differently in same item by distinction of sex difference. Data partitioning based on user's information can solve this problem and have influence on result.

Data partitioning based on user information can create data groups by singleness of various attribute and several union. This paper can apply data groups which just consider distinction of sex attribute. Also, it can consider distinction of sex and age attribute. Furthermore, it can consider more attributes. Example is depicted in (Figure 2). Figure 2 shows that data by distinction of sex and age attribute is classed.

![Diagram of Data Partitioning]

Fig. 2. Data partitioning

3.2 Collaborative Filtering

In this paper, a collaborative filtering method based on Singular Value Decomposition (SVD) is used. This requires the singular value decomposition to construct a
A Recommendation System for Intelligent User Interface

preference vector space that can be used to represent conceptual user-contents associations [3][4].

In this method, a matrix of users by contents is constructed. Once users and items are determined, a user by contents incidence matrix \( A = (a_{ij}) \) can be determined, where \( a_{ij} \) is the preference of user \( i \) for content \( j \). Matrix \( A \) is analyzed by SVD into three matrices of special form.

In collaborative filtering, SVD constructs approximate model that has reduced dimension. Similarities between users are roughly estimated in reduced dimension in this model. Cosine distance is represented in a geographical space corresponding to the estimated similarities.

SVD in collaborative filtering is considered a technique to eliminate set of unconsidered content. Users are represented by vector of ratings. As effect of reduced dimension, users that have more and less different profile are located as same vector. This is a feature for achieving improvement of unreliable data.

Any \( u \times i \) matrix \( A \) whose number of rows \( u \) is greater than or equal to its number of columns \( i \), can be written as the product of an \( u \times i \) column-orthogonal matrix \( U \), and an \( i \times i \) diagonal matrix \( S \) with positive or zero elements (the singular value), and the transpose of an \( i \times i \) orthogonal matrix \( I \). The various shapes of these matrices will be made clearer by the following tableau.

\( A \) can be rewritten to express any matrix \( A_{ij} \) as a sum of outer product of columns of \( U \) and rows \( V^T \), with the “weighting factors” being the singular value \( s_k \),

\[
A_{ij} = \sum_{k=1}^{i} s_k U_{ik} I_{jk}
\]

(1)

If a situation where most of the singular values \( s_j \) of a matrix \( A \) are very small, then \( A \) will be well-approximated by only a few items in the sum. This means that we have to store only a few columns of \( U \) and \( I \) (same \( k \) ones) and we will be able to recover, with good accuracy, the whole matrix.

Note also that it is very efficient to multiply such an approximated matrix by a vector \( x \): We just dot \( x \) with each of the stored columns of \( I \), multiply the resulting scalar by the corresponding \( s_k \), and accumulate that multiple of the corresponding column of \( U \). If your matrix is approximated by a small number \( K \) of singular values, then this computation of \( A \cdot x \) takes only about \( K(M+N) \) multiplications, instead of \( MN \) for the full matrix.

The dot product between row vectors of \( ^\wedge A \) reflects the extent to which two users have a similar pattern of occurrence across the set of items. The matrix \( ^\wedge A \cdot ^\wedge A \) is the square symmetric matrix containing all these user-to-user dot products. Since \( S \) is diagonal and \( I \) is orthonormal. It is easy to verify that:

\[
^\wedge A \cdot ^\wedge A = US^2U^T
\]

(2)

Note that this means that the i, j cell of \( ^\wedge A \cdot ^\wedge A \) can be obtained by taking the dot product between the i and j rows of the matrix \( US \). That is, if one considers the rows of \( US \) as coordinates for users, dot products between these points give the comparison
between users. Note that the relation taking $U$ as coordinates is simple since $S$ is diagonal. The positions of the points are the same except that each of axes has been stretched or shrunk in proportion to the corresponding diagonal element of $S$.

Also, Similarities between two users is estimated by simply using normal cosine distance. Taking $A$ and multiplying by feature item matrix $I$, yields $US$, which is the user feature response matrix multiplied with the feature importance matrix. Feature space comparison will determine nearest neighbor based on their underlying reaction to the features in the users. If the least important feature is thrown, it is possible to reduce the dimension of the comparison and make better use of correlation.

\[
R = US
\]  

(3)

3.3 Partitioned Collaborative Filtering

Partitioned collaborative filtering method searches for similar users from partitioned data groups. Original data is divided by user information with many attributes. Each data group can be used for collaborative filtering. Collaborative filtering calculates similar users in each data group.

Also, some data groups can be divided more detail and can be used for searching more similar users than previous data groups. Figure 3 shows that collaborative filtering uses partitioned data groups to search similar users.

![Fig. 3. Partitioned collaborative filtering](image)

3.4 Recommendation Algorithm

In this phase, we introduce basic algorithm to generate recommendation list for users. The previous subsection provided an introduction to data partitioning
and partitioned collaborative filtering. The general procedure is summarized in (Figure 4).

4 Experiment

In this section, we report the experimental evaluation of our proposed recommendation system. We describe the data set used. The experimental methodology, as well as performance measures we consider appropriate for this task.

Algorithm: Recommendation System.
Predict recommendation list that users prefer.

Input: (i) UR, a user rating. (ii) UI, user information, (iii) CI, content information. (iv) UGT, user generation threshold.
Output: R, a recommendation list for users
Method: The method is outlined as follows.
(1) DG organize_data_group(UI, CI); // DG create data groups for users.
(2) DR reduce_dimension (UR, DG); // DR hold reduced-dimension representation for UR at current data groups
(3) US user_similarity(UR, DG); // US hold similarities among users.
(4) RL recommendation_list(US); // RL hold recommendation list for each users.
(5) If the quality of recommendation is below UGT, go back (1).

![Fig. 4. Basic algorithm for recommendation](image)

4.1 EachMovie Data set

The EachMovie Data set is a real rating data from EachMovie Recommendation service by the DEC Systems Research Center. Over a period of 18 months 72,916 users entered a total of 2,811,983 numeric ratings for 1,628 different movies (films and videos) [2].

User information in the EachMovie Data set has only age, gender, zip code and it is unclear that these data is valid. Therefore, in this paper, we selected users whose age is between 10 and 50 and gender is existed. Such a data consists of 42,324 users and 1,778,950 numeric ratings for 1,628 different movies.

4.2 Experimental Methodology

First, we carried out experiments to decide optimal parameters for a collaborative filtering using SVD. These parameters include the dimension of reduced space and
the threshold for users' similarity. Because the size of total data is very large, we randomly extracted 10 set consisting 1000 users from total data and removed 20% of these data for performance evaluation. We measured performance by mean absolute error (MAE) between the actual and predicted ratings. Similarities between users were measured by cosine distance. We repeated the entire process 2~10 for reduced dimension and 0.5, 0.6, 0.7, 0.8, 0.9 for similarities. We found 9 for reduced dimension and 0.8 for similarities to be optimal dimension (Figure 5).

Second, we created data groups for user. We performed SVD-based collaborative filtering at each data group. If the error of collaborative filtering is more than UGT at a data group, the data group that has most discrete values is specialized. This process is repeated until the error of all data groups is lower than UGT (Figure 6).

![Fig. 5. Determination of optimal values for reduced dimension and cosine distance](image)

Third, we carried out experiments applying partitioned collaborative filtering. We performed collaborative filtering at each data group. The result is depicted in Figure 7. For example, the result showed that users whose age is between 11 and 20 and gender is male need to be specialized in data group (A). Those users were specialized by more detail data group. We performed again at each data group in partial detailed data groups (B). This process was repeated until the error of all data groups was lower than UGT(0.230) (C). We found that if users were female and age was between 11 and 30, the data group need to be specialized some more detail such as (E)(F).
Conclusively, performance from whole data group that is no longer partitioned appears as is good.

5 Conclusions

We introduced a recommendation system using SVD-based partitioned collaborative filtering. We presented a framework of organizing users into different data groups and performing collaborative filtering on each data group to overcome problems that traditional recommendation systems have. Recommendation system can see user more meaningfully and offer useful recommendation.

Fig. 6. Data partitioning process

Our future work in this area is to carried out experiments in user having detail information. Since the proposed framework of recommendation system utilizes only age and gender, it is necessary to consider diverse user information for more effective recommendation services.

Acknowledgement

The work was supported (in part) by Biometrics Engineering Research Center, (KOSEF).
Fig. 7. The result of partitioned collaborative filtering

References

1. Belkin, N.J., Croft, W.B. : Information filtering and information retrieval-two sides of the same coin. Communications of the ACM 35(12):29-38 (1992).
2. DEC : Eachmovie collaborative filtering data set. http://www.research.digital.com/SRC/eachmovie/.
3. Deerwester, S., Dumais, S.T., Landauer, T. K., Furnas, G.W., Harshman, R.A. : Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407 (1990).
4. Foltz, P.W. : Using latent semantic indexing for information filtering. Proceedings of the conference on Office information systems, Pages 40 – 47 (1990).
5. Goldberg, D., Nichols, D., Oki, B., Terry, D. : Using collaborative filtering to weave an information tapestry. Communication of the ACM, 35(12):61-70 (1992).
6. Joachims, T., Freitag, D., Mitchell, T.: Webwatcher: A tour guide for the world wide web. In the 15th International Conference on Artificial Intelligence, Nagoya, Japan (1994).
7. Krulwich, B., Burkey, C.: Learning user information interests through extraction of semantically significant phrases. In Proceedings of the AAAI spring Symposium on Machine Learning in Information Access, Standford, California (1996).
8. Krulwich, B.: LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data. Artificial Intelligence Magazine 18(2), 37-45 (1997).
9. Lang, K.: NewsWeeder: Learning to filter netnews. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, California (1995).
10. Leiberman, H.: An agent that assist web browsing. In Proceeding of the International Joint Conference on Artificial Intelligence, Montreal, Canada (1995).
11. Maes, P.: Agents that reduce work and information overload. Communication of the ACM, Vol. 37, No 7, pp. 31-40 (1994).
12. Paul, R., Neophytos, I., Mitesh, S., Peter, B., John, R.: An open architecture for collaborative filtering of netnews. In Proceedings of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pages 175-186 (1994).
13. Pazzani, M.: A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review, chapter A (1999).
14. Pryor, M.H.: The Effects of Singular Value Decomposition on Collaborative Filtering. Dartmouth College Technical Report PCS-TR98-338 (1998).
15. Schafer, J.B., Konstan, J.A., Riedl, J.: Recommender Systems in E-Commerce. In ACM Conference on Electronic Commerce (EC-99), pages 158-166 (1999).
16. Schallehn, E., Sattler, K., Saake, G.: Advanced Grouping and Aggregation for Data Integration. Proc. 4th Int. Workshop on Engineering Federated Information Systems, EFIS'01, Berlin, Germany (2001).
17. Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating ‘word of mouth’. In Proceedings of the Conference on Human Factors in Computing Systems, pages 210-217 (1995).
18. Soboroff, I.M.: Collaborative Filtering with LSI. Department of Computer Science and Electrical Engineering, University of Maryland, Technical Report TR-CS-98-01 (1998).
19. Upendra, S.: Social Information Filtering for Music Recommendation. S.M. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology (1994).
Fast Half Pixel Motion Estimation Based on the Spatial Correlation

Hyo Sun Yoon and Guee Sang Lee*

Department of Computer Science, Chonnam National University, 300 Youngbong-dong, Buk-gu, Kwangju 500-757, Korea estheryoon@hotmail.com, gslee@chonnam.chonnam.ac.kr

Abstract. Motion estimation is an important part of video encoding systems, because it can significantly affect the output quality of an encoded sequence. Most of the advanced techniques for motion estimation consist of two stages, the integer pixel motion estimation and the half pixel motion estimation. Although many fast integer pixel motion estimation algorithms have been developed, the performance of current methods for half pixel motion estimation still has room for improvement. In this paper, we propose a new algorithm for half pixel motion estimation which exploits the spatial correlation between integer and half pixel motion vectors. The proposed algorithm adaptively decides the search range of half pixel points considering the correlation of the motion vectors around the point of interest. The experiment results show that the proposed method outperforms most of existing methods in computation time by reducing the number of search points with little or no degradation in image quality. Even when compared to the full half pixel search method, it shows the search point reduction up to 95% with only 0.01 ∼ 0.09 (dB) degradation of image quality in terms of PSNR.

1 Introduction

Recently, great interest has been devoted to the study of different approaches in video compressions. The high correlation between successive frames of a video sequence makes it possible to achieve high coding efficiency by reducing the temporal redundancy. Motion estimation (ME) and motion compensation techniques are an important part of most video encoding, since it could significantly affect the compression ratio and the output quality.

Generally, ME is made of two parts, the integer pixel motion estimation and the half pixel motion estimation. For the first part, the integer pixel motion estimation, many search algorithms such as Diamond Search (DS) [1, 2], Three Step Search (TSS) [3], New Three Step Search (NTSS) [4], Four Step Search (FSS) [5], Two Step Search (2SS) [6], Two-dimensional logarithmic search algorithm [7], HEXagon-Based Search (HEXBS) [8], Motion Vector Field Adaptive Search

*corresponding author.
Fast Half Pixel Motion Estimation Based on the Spatial Correlation Technique (MVFAST) [9], and Predictive MVFAST (PMVFAST) [10] have been proposed to reduce the computational complexity. Some fast integer pixel motion estimation algorithms of these algorithms can find an integer pixel Motion Vector (MV) by examining less than 10 search points. For the second part, the half pixel motion estimation, the Full Half pixel Search Method (FHSM) that is a typical method, examines eight half pixels around the integer motion vector to determine a half pixel motion vector. This method takes nearly half of the total computational complexity in the ME that uses fast block matching algorithms for the integer motion vector. Therefore, it becomes more important to reduce the computational complexity of the half pixel motion estimation. For these reasons, Horizontal and Vertical Direction as Reference (HVDR) [11], the Parabolic Prediction-based, Fast Half Pixel Search algorithm (PPHPS) [12], Chen’s Fast Half Pixel Search algorithm (CHPS-1) [13] and the methods [14, 15, 16] have been proposed to accelerate the half pixel motion estimation.

In this paper, we propose the fast method based on the spatial correlation among integer pixel motion vectors and half pixel motion vectors. According to the spatial correlation among integer pixel motion vectors and half pixel motion vectors, The proposed method selects the search pattern adaptively and decides whether the half pixel motion estimation is skipped or not for the half pixel motion vector. Experiments results show that the proposed method preserves the image quality and reduces the computational complexity when compared to that of FHSM.

This paper is organized as follows. Section 2 describes the previous works. The proposed method is described in Section 3. Section 4 reports the simulation results and conclusions are given in Section 5.

2 The Previous Works

In Motion Estimation and Compensation, the half pixel motion estimation is used to reduce the motion prediction error and to improve image quality. The full half pixel search method (FHSM) that is a typical method, examines eight half pixels around the integer motion vector ‘c’ illustrated in Fig. 1. The cost function value of each eight half pixels is calculated to find the best matching point. Finally, the half pixel motion vector is obtained by comparing the cost function value of the best matching point with that of the point ‘c’ that is pointed by the integer MV shown Fig. 1. This method takes nearly half of the total computational complexity in the ME that uses fast block matching algorithms for the integer motion vector. Therefore, it becomes more important to reduce the computational complexity of the half pixel motion estimation. For these reasons, the fast half pixel motion estimation algorithms have been proposed.

In HVDR which is one of the fast half pixel motion estimation algorithms, 2 neighboring half pixel points in vertical direction and 2 neighboring half pixel points in horizontal direction around the integer motion vector ‘c’ illustrated in Fig. 1. are examined. And the best matching point in each direction is decided. Then, a diagonal point between these two best matching points is also examined.
the point having the minimum cost function value among these 5 points and the point 'c' shown Fig. 1. is decided as the half pixel motion vector. In HVDR, only 5 half pixel points are checked to find the half pixel motion vector.

CHPS-1 examines 4 horizontal and vertical half pixels '2', '4', '5', '7' shown Fig. 1. The best matching point is decided as the half pixel motion vector by comparing the cost function values of these 4 half pixel points and the point 'c'. In CHPH-1, only 4 half pixel points are checked to find the half pixel motion vector.

PPHPS predicts the possible optimal half pixel point by using the cost function value of 5 integer pixel points 'A', 'B', 'C', 'D', 'E' shown Fig. 1. The cost function values of the optimal half pixel point and its nearest points are calculated to find the best matching point. And the point of the minimum cost function value is decided as the final half pixel MV by comparing the cost function value of this best matching point with that of the point 'c'. In PPHPS, only 3 half pixel points are checked to find the half pixel motion vector.

3 The Proposed Method

In order to reduce the computational complexity of the half pixel motion estimation, the proposed method predicts the motion of the current block in the half pixel motion estimation by using the spatial correlation among integer motion vectors and half pixel motion vectors. And also the proposed method use the following observations shown in the Table 1. When the integer pixel MV is (0,0), the probability of that its half pixel MV is (0,0) is about 77% ~ 96%.

In other words, the proposed method exploits spatially correlated motion vectors depicted in Fig. 2. and the above observation to predict the motion
Table 1.

| Prob. | Akiyo | Claire | Container | Foreman | M&D | News | Salesman | Silent | Stefan | Suzie | Table |
|-------|-------|--------|-----------|---------|-----|------|----------|--------|--------|-------|-------|
|       | 96    | 95     | 97        | 78      | 91  | 93   | 95       | 92     | 77     | 84    | 93    |

Table 1.

| MV1_Integer \((dx1,dy1)\) | MV1_Half \((dxh1,dyh1)\) |
|---------------------------|---------------------------|
| MV2_Integer \((dx2,dy2)\) | MV2_Half \((dxh2,dyh2)\) |
| MVC_Integer \((dxc,dyc)\) |

MV1_Integer \((dx1,dy1)\) : integer pixel MV of above block
MV2_Integer \((dx2,dy2)\) : integer pixel MV of left block
MVC_Integer \((dxc,dyc)\) : integer pixel MV of current block
MV1_Half \((dxh1,dyh1)\) : half pixel MV of above block
MV2_Half \((dxh2,dyh2)\) : half pixel MV of left block

Fig. 2. Blocks for Spatial Correlation Information

of the current block in the half pixel motion estimation. The block diagram of the proposed method appears in Fig. 3. According to the spatial correlated information, the proposed method selects the search pattern adaptively and decides whether the half pixel motion estimation is skipped or not. The proposed method is summarized as follows.

**Step 1** If MVC_Integer \((dxc,dyc)\), the integer MV of the current block shown in Fig. 2., is equal to \((0,0)\), go to Step 2. Otherwise, go to Step 3.

**Step 2** I. If MV1_Integer \((dx1,dy1)\) which is the integer MV of the above block shown in Fig. 2., and MV2_Integer \((dx2,dy2)\) which is the integer MV of the left block shown in Fig. 2., are equal to \((0,0)\), go to II. Otherwise, go to III.

II. If MV1_Half \((dxh1,dyh1)\) which is the half MV of the above block shown in Fig. 2., and MV2_Half \((dxh2,dyh2)\) which is the half MV of the left block shown in Fig. 2., are equal to \((0,0)\), \((0,0)\) is decided as the half pixel MV of the current block. In other words, the half pixel motion estimation is skipped. Otherwise, go to III.

III. CHPS-1 is selected as the search pattern for the half pixel motion estimation

**Step 3** I. If MV1_Integer \((dx1,dy1)\) and MV2_Integer \((dx2,dy2)\) are equal to MVC_Integer \((dxc,dyc)\), go to II. Otherwise, go to III.

II. If MV1_Half \((dxh1,dyh1)\) is equal to MV2_Half \((dxh2,dyh2)\), \((dxh2, dyh2)\) is decided as the half pixel MV of the current block. In other words, the half pixel motion estimation is skipped. Otherwise, CHPS-1 is selected as the search pattern for the half pixel motion estimation.

III. HVDR is selected as the search pattern for the half pixel motion estimation
4 Simulation Result

In this section, we show the experiment results for the proposed method. The proposed method has been evaluated in the H.263 encoder. Eleven QCIF test sequences are used for the experiment: Akiyo, Carphone, Claire, Foreman, Mother and Daughter, News, Salesman, Silent, Stefan, Suzie and Table. The mean square error (MSE) distortion function is used as the block distortion measure (BDM). The quality of the predicted image is measured by the peak signal to noise ratio (PSNR), which is defined by

\[
MSE = \left( \frac{1}{MN} \right) \sum_{m=1}^{M} \sum_{n=1}^{N} [x(m,n) - \hat{x}(m,n)]^2 \tag{1}
\]

\[
PSNR = 10 \log_{10} \frac{255^2}{MSE} \tag{2}
\]

In Eq. (3), \(x(m,n)\) denotes the original image and \(\hat{x}(m,n)\) denotes the motion compensated prediction image. For integer pixel motion estimation, Full Search algorithm and the unrestricted center-based diamond search (UCBDS) [1] are adopted. For half pixel motion estimation, we compared FHSM, HVDR, CHPS-1, and PPHPS to the proposed method in both of image quality and search speed. The simulation results in Table 2 and 3 show that the search speed of the proposed method is faster than the other methods (FHSM, HVDR, CHPS-1, and PPHPS) while its PSNR is similar to them except for FHSM. In computational complexity, the proposed method is about 2.5 ~ 20 times faster than is FHSM.
Table 2. Average PSNR for half pixel motion estimation algorithms

| Integer-pel ME method | Full search | UCBDS |
|-----------------------|-------------|-------|
|                       | FHSM | HVDR | CHPS-1 | PPHPS | Proposed | FHSM | HVDR | CHPS-1 | PPHPS | Proposed |
| Akiyo                 | 34.5  | 34.41 | 34.46  | 34.16  | 34.43  | 34.39  | 34.28 | 34.34  | 34.03  | 34.28  |
| Carphone              | 30.88 | 30.85 | 30.86  | 30.88  | 30.87  | 30.48  | 30.45 | 30.46  | 30.47  | 30.48  |
| Claire                | 35.05 | 35.02 | 35.03  | 34.87  | 35.04  | 34.85  | 34.81 | 34.83  | 34.66  | 34.84  |
| Foreman               | 29.54 | 29.52 | 29.50  | 29.47  | 29.50  | 28.65  | 28.64 | 28.61  | 28.60  | 28.61  |
| M&D                   | 31.54 | 31.50 | 31.54  | 31.44  | 31.46  | 31.45  | 31.39 | 31.41  | 31.29  | 31.34  |
| News                  | 30.59 | 30.49 | 30.54  | 30.49  | 30.50  | 30.25  | 30.20 | 30.24  | 30.14  | 30.16  |
| Salesman              | 32.7  | 32.64 | 32.67  | 32.53  | 32.65  | 32.62  | 32.55 | 32.59  | 32.48  | 32.55  |
| Silent                | 31.81 | 31.80 | 31.76  | 31.72  | 31.73  | 31.61  | 31.61 | 31.57  | 31.51  | 31.52  |
| Stefan                | 23.89 | 23.85 | 23.86  | 23.69  | 23.81  | 22.78  | 22.74 | 22.72  | 22.58  | 22.7   |
| Suzie                 | 32.19 | 32.17 | 32.15  | 32.14  | 32.15  | 31.9   | 31.84 | 31.81  | 31.78  | 31.84  |
| Table                 | 26.52 | 26.48 | 26.46  | 26.35  | 26.46  | 25.72  | 25.69 | 25.65  | 25.6   | 25.65  |

Table 3. The Number of Search points per half pixel MV

|        | FHSM | HVDR | CHPS-1 | PPHPS | Proposed |
|--------|------|------|--------|-------|----------|
| Akiyo  | 8    | 5    | 4      | 3     | 0.4      |
| Carphone | 8    | 5    | 4      | 3     | 3.2      |
| Claire | 8    | 5    | 4      | 3     | 1.2      |
| Foreman | 8    | 5    | 4      | 3     | 3.9      |
| M&D    | 8    | 5    | 4      | 3     | 1.6      |
| News   | 8    | 5    | 4      | 3     | 1.1      |
| Salesman | 8    | 5    | 4      | 3     | 0.7      |
| Silent | 8    | 5    | 4      | 3     | 1.4      |
| Stefan | 8    | 5    | 4      | 3     | 3.2      |
| Suzie  | 8    | 5    | 4      | 3     | 3.1      |
| Table  | 8    | 5    | 4      | 3     | 3.2      |
| Average | 8    | 5    | 4      | 3     | 2.1      |

And in PSNR, the proposed method is about 0.01 ~ 0.09 (dB) worse than FHSM. Thus, the proposed method is a good alternative for FHSM in the half pixel motion estimation.

5 Conclusion

Based on the spatial correlation among integer pixel MVs and half pixel MVs, and the correlation between the integer MV and its half MV, a fast method for the half pixel motion estimation is proposed in this paper. According to the correlated information, the proposed method selects the search pattern adaptively and decides whether the half pixel motion estimation is skipped or not. Experimental results show that the average speedup improvement of the proposed method over FHSM can be up to 2.5 ~ 20 times faster and the average image quality degradation is about 0.01 ~ 0.09 (dB).
Acknowledgement

This work was supported by grant No.R05-2003-000-11345-0 from the basic Research Program of the Korea Science & Engineering Foundation.

References

1. Tham, J.Y., Ranganath, S., Kassim, A.A.: A Novel Unrestricted Center-Biased Diamond Search Algorithm for Block Motion Estimation. IEEE Transactions on Circuits and Systems for Video Technology. 8(4) (1998) 369–375
2. Shan, Z., Kai-kuang, M.: A New Diamond Search Algorithm for Fast block Matching Motion Estimation. IEEE Transactions on Image Processing. 9(2) (2000) 287–290
3. Koga, T., Inuma, K., Hirano, Y., Iijim, Y., Ishiguro, T.: Motion compensated interframe coding for video conference. In Proc. NTC81. (1981) C9.6.1–9.6.5
4. Renxiang, L., Bing, Z., Liou, M.L.: A New Three Step Search Algorithm for Block Motion Estimation. IEEE Transactions on Circuits and Systems for Video Technology. 4(4) (1994) 438–442
5. Lai-Man, P., Wing-Chung, M.: A Novel Four-Step Search Algorithm for Fast Block Motion Estimation. IEEE Transactions on Circuits and Systems for Video Technology. 6(3) (1996) 313–317
6. Yuk-Ying, C., Neil, W.B.: Fast search block-matching motion estimation algorithm using FPGA. Visual Communication and Image Processing 2000. Proc. SPIE. 4067 (2000) 913–922
7. Jain, J., Jain, A.: Dispalcement measurement and its application in interframe image coding. IEEE Transactions on Communications. COM-29 (1981) 1799–1808
8. Zhu, C., Lin, X., Chau, L.P.: Hexagon based Search Pattern for Fast Block Motion Estimation. IEEE Transactions on Circuits and Systems for Video Technology. 12(5) (2002) 349–355
9. Ma, K.K., Hosur, P.I.: Report on Performance of Fast Motion using Motion Vector Field Adaptive Search Technique. ISO/IEC/JTC1/SC29/WG11.M5453 (1999)
10. Tourapis, A.M., Liou, M.L.: Fast Block Matching Motion Estimation using Predictive Motion Vector Field Adaptive Search Technique. ISO/IEC/JTC1/SC29/WG11.M5866 (2000)
11. Lee, K.H., Choi, J.H., Lee, B.K., Kim. D.G.: Fast two step half pixel accuracy motion vector prediction. Electronics Letters 36(7)(2000) 625–627
12. Cheng, D., Yun, H., Junli, Z.: A Prabolic Prediction-Based, Fast Half Pixel Search Algorithm for Very Low Bit-Rate Moving Picture Coding. IEEE Transactions on Circuits and Systems for Video Technology. 13(6) (2003) 514–518
13. Cheng, D., Yun, H.: A Comparative Study of Motion Estimation for Low Bit Rate Video Coding. SPIE 4067(3)(2000) 1239–1249
14. Sender, Y., Yano, M.: A Simplified Motion Estimation using an approximation for the MPEG-2 real time encoder. ICASSP’95,(1995) 2273–2276
15. Choi, W.I., Jeon, B.W.: Fast Motion Estimation with Modified diamond search for variable motion block sizes. ICIP 2003. (2003) 371–374
16. Li, X., Gonzles, C.: A locally Quadratic Model of the Motion Estimation Error Criterion Function and Its Application to Subpixel Interpolations. IEEE Transactions on Circuits and Systems for Video Technology. 6(1) (1996) 118–122
A New Vertex Selection Scheme Using Curvature Information

Byoung-Ju Yun¹, Si-Woong Lee², Jae-Soo Cho³, Jae Grak Choi⁴, and Hyun-Soo Kang⁵

¹ Dept. of Information and Communication, Kyungpook National University, Daegu, South Korea
bjisyun@ee.knu.ac.kr

² Divi. of Information Communication and Computer Engineering, Hanbat National University, Daejeon, South Korea

³ School of Internet Media Engineering, Korea University of Technology and Education, Cheonan, South Korea

⁴ Dept. of Computer Engineering, Dongeui University, Busan, South Korea

⁵ Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul, South Korea

Abstract. A new vertex selection scheme using the curvature information of contour points for polygonal contour approximation is presented. The proposed method consists of two-step procedure. At first, we compute curvature values for contour points on the curvature scale space (CSS) and select high curvature points as principal vertices. Selected principal vertices, thereby, divide an overall contour into several contour segments. In the second step, the vertices for each contour segment are selected using the progressive vertex selection (PVS) method in order to obtain minimum number of vertices under the given maximum distance criterion (Dₖₐₓ).

1 Introduction

Region-based image coding method [1] and object-based image coding method [2] introduced shape coding into image and video coding. A region is defined its homogeneous texture and described by texture and shape, and an object is defined by its uniform motion and described by motion, shape and colour parameters. The purpose of using shape was to achieve better subjective picture quality, increased coding efficiency as well as an object-based video representation. MPEG-4 visual is the first international standard allowing the transmission of arbitrarily shaped video objects (VO’s) [3]. MPEG-4 visual transmits motion, texture, and shape information of one VO within one bitstream. By MPEG-4 visual, the importance of shape for video objects is recognized. In MPEG-7, with other information such as colour, texture, and motion of an object, shape is also used as a Descriptor (D), which describes an object.

Shape information is very important in many applications, thus several researchers have been studying to efficiently encode shape information [4]～[7].
There are two major classes of shape coders: bitmap-based coders and contour-based coders. The former encodes each pixel whether it belongs to the object or not and the latter encodes the outline of the object. Our method belongs to the second type. We approximate the boundary by a polygon, which has merits in inherent quality control [8], and consider the problems of finding the polygon, which leads to the smallest distortion for a given number of bits. The curvature information has been widely used as a feature in the fields of object matching, pattern recognition, and computer vision. High curvature points, which are commonly called corners, are rich in information content and important to characterize the shape of an object. However, the curvature information has not been properly incorporated in the contour coding applications. Two schemes are commonly employed for vertex selection in polygon-based contour approximation: the progressive vertex selection (PVS) method [9] and the iterated refinement method (IRM) [6]. Both methods use the subjective criterion of the allowable maximum distance ($D_{\text{max}}^*$) in the vertex selection process, so the resultant polygon is apt to lose the intrinsic shape of the original contour.

For example, consider the case of Fig 1(a), which shows the result of vertex selection when the PVS method is employed. Thick line represents the original contour and thin one is the approximating polygon. As shown, there is an apparent shape deformation in the resultant polygon. The deformation is generated because the PVS adds vertices from the starting point when the distance between the approximating polygon and the original contour exceeds $D_{\text{max}}^*$. Fig 1(b) shows an alternative vertex selection where the vertices are mostly selected at the corner points of the contour. The resulting polygon represents the shape of the original contour more precisely than Fig 1(a), and much error region are reduced at the same time. In this paper, we propose a new polygon-based contour coding method where the curvature information is employed as an additive feature in vertex selection process.

This paper is organized as follows: In section 2, the proposed vertex selection method is presented. The experimental results are shown in section 3 and conclusions are in section 4.
2 Proposed System

2.1 System Overview

Fig. 2 shows a general framework for shape coding methods that rely on polygonal approximations of the shape [4].

Fig. 2. General block diagram of vertex-based shape information encoder

At first, an input segmentation mask is preprocessed to decrease spatial resolution to the desired level and reduce unwanted noise and edge jaggedness. Then, a contour is determined and extracted from the input mask. Vertex selection (VS) module selects vertices defining a polygonal approximation of the contour subject to some distortion criteria or a limit of the number of vertices. Finally, vertex encoding (VE) module encodes the lists of vertices for each contour approximation. Some vertex-based coding methods reconstruct an approximation of the object contour and encode the approximation error in the dashed elements in Fig. 2. The proposed method reorganizes the VS (Vertex Selection) module of Fig. 2 (See Fig. 3). It consists of two-step procedure. At first, it selects contour pixels, which have their maxima of absolute curvature, as the principal vertices using the CSS, thereby, dividing original contour into several contour segments. Each contour segment is considered as an open contour, where two end points are two consecutive principal vertices. In the second step, it approximates each contour segment using the PVS, because the PVS selects minimum vertices under the given maximum distortion.

Fig. 3. Block diagram of the proposed vertex selection method
2.2 Curvature-Based Principal Vertex Selection

The CSS technique is suitable for recovering invariant geometric features - curvature zero-crossing points and/or extrema - of a planar curve at multiple scales. To compute it, the curve is first parameterised by the arc length parameter and evolved by Gaussian of width $\sigma$ ($\sigma$ is referred as the scale parameter). And the curvature value of the evolved curve is calculated according to [10]. The corner points are defined as the local maxima of the absolute value of curvature. At a very low scale, there exist many such maxima due to noise on the digitised contour. As the scale is increased, the noise disappears and only the maxima corresponding to the real corners remain. The CSS corner detection method finds the corners at these local maxima. We name the corner points as the principal vertices of a polygon. The CSS computes the absolute value of curvature at the initial scale $\sigma_{high}$. The local maxima of the absolute curvature are the principal vertices. Because the principal vertices are detected at scale $\sigma_{high}$, the actual localizations of them change. So after the principal vertices are located, a tracking method is introduced to improve the localization of them. The CSS computes a curvature at a lower scale and examines the principal candidates in a small neighbourhood of the previous principals. And, the locations are updated, if needed, in this neighbourhood. The tracking is continued until scale is very low. This process gives a good localization and reduces the computational burden. The number of principal vertices is determined at the initial $\sigma_{high}$ and tracking only changes the localization, not the number of the principals. The principal vertices of Stefan image, detected by the CSS, are shown in Fig. 4 (a).

2.3 Segment-Wise Vertex Addition

Each contour segment is considered as an open contour, where two end points are two consecutive principal vertices (Fig. 4 (b)). From one principal vertex,

Fig. 4. Curvature-based vertex selection (a) Principal vertices selected by CSS (Stefan image) (b) Polygonal approximation of each contour segment using PVS
the PVS performs vertex selection. It traces the contour points from initial point and check whether the peak distance ($d_P$) between a straight line, whose two end points are initial point and current contour point, and the contour portion, which is approximated by the segment, is larger than given maximum distance criterion ($D_{max}^*$). If $d_P$ is larger than $D_{max}^*$, the previous point is selected as a vertex of a polygon and the process is continued from this vertex.

3 Experimental Results

For experiment, the distortion metric, which has been adopted as a performance measure in MPEG-4, is used to evaluate the performance of the proposed method:

$$D = \frac{\text{number of pixels in error}}{\text{number of interior pixels}}$$  \hspace{1cm} (1)

For calculating the bit rates, the octant-based coding [4] is used. Many kinds of MPEG-4 test sequences were used in the experiments, and, among them, the results of two sequences (Children kids and Stefan) are presented here.

Fig. 5 (a) and (b) show the rate-distortion curves of Children-kids and Stefan sequences, respectively. The proposed method showed outstanding performance compared with the conventional vertex selection methods over all distortion levels. It gave a maximum 54% and an average 45% error reduction than the PVS and a maximum 31% and an average 22% reduction than the IRM over Children-kids sequences. And it also gave a maximum 57% and an average 47% error reduction than the PVS and a maximum 33% and an average 18% reduction than the IRM over Stefan sequences. Table 1 shows the performance comparison between the proposed method and the conventional methods according to the
$D_{\text{max}}^*$ of children-kids sequence. As shown in the table 1, the proposed method shows excellent performance than the conventional methods. Fig. 6 shows the original mask and the error images of Stefan image; (a) is original mask and (b)-(d) are the error images which are the mismatched pixels between the original mask and the reconstructed one by the proposed, by the PVS, and by the IRM, respectively. The proposed method is able to reduce the error region with similar number of vertices.

**Table 1.** Performance comparison between the proposed method and the conventional methods (Children-kids sequence)

| $D_{\text{max}}^*$ | PVS |      | IRM |      | Proposed |
|---------------------|------|------|------|------|----------|
|                     | # of bits | D | # of bits | D | # of bits | D |
| 1.0                 | 637.7  | 0.026157 | 674.9  | 0.015431 | 584.7  | 0.0149265 |
| 1.5                 | 494.0  | 0.039820 | 537.1  | 0.024036 | 464.9  | 0.0237445 |
| 2.0                 | 400.4  | 0.057492 | 446.9  | 0.033056 | 393.0  | 0.0349420 |
| 2.5                 | 364.7  | 0.071747 | 405.0  | 0.040211 | 347.9  | 0.0451210 |
| 3.0                 | 309.6  | 0.088793 | 364.8  | 0.050069 | 303.9  | 0.0563130 |
| 3.5                 | 277.0  | 0.102020 | 324.9  | 0.061718 | 282.0  | 0.0641875 |

**Fig. 6.** Comparision of the error image of Stefan image (a) Original mask (b) Proposed method (# of vertices: 26, # of error pixels: 403)(c) PVS (# of vertices: 26, # of error pixels: 827) (d) IRM (# of vertices: 27, # of error pixels: 538)
4 Conclusions

This paper proposed a new curvature-based vertex selection method for polygonal contour approximation. The proposed method incorporated the curvature information into the vertex selection for the first time. The proposed method showed the outstanding performance than the conventional methods in the rate-distortion sense. However, the proposed method has some computational burden because it compute the curvature information before vertex selection.

The proposed method will offer the high speed solution to find optimal position of polygon’s vertices in the vicinity of contour including the contour of the object in the rate-distortion sense.

Acknowledgement. This work was supported by the IT Research Center (ITRC), Ministry of Information and Communication, Korea.

References

1. M. Kunt, A. Ikonomopoulos, and M. Kocher, “Second-generation image coding techniques,” *Proceedings of the IEEE*, vol. 73, no. 4, pp. 549 – 574, April 1985.
2. H. G. Musmann, M. Höttter, and J. Ostermann, “Object-Oriented Analysis-Synthesis Coding of Moving Images,” *Signal Processing: Image Communication*, vol. 1, pp. 117 – 138, October 1989.
3. R. Koenen, Ed., “Overview of the MPEG-4 standard,” International Standards Organization, Stockholm meeting, ISO/IEC/JTC1/SC29/WG11 N1730, July 1997.
4. Kevin J. O’Connell, ”Object-Adaptive Vertex-Based Shape Coding Method,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 251 – 255, February 1997.
5. J. Ostermann, E. S. Jang, J. S. Shin, and T. Chen, ”Coding of Arbitrarily Shaped Video Objects in MPEG-4,” *Proceedings of ICIP 97*, vol. 1, pp. 496 – 499, 1997.
6. A. K. Katsaggelos, L. P. Kondi, F. W. Meier, J. Ostermann, and G. M. Schuster, ”MPEG-4 and Rate-Distortion-Based Shape-Coding Techniques,” *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1126 – 1154, June 1998.
7. B. J. Yun, S. W. Lee, and S. D. Kim, ”Vertex adjustment method using geometric constraint for polygon-based shape coding,” *Electronics Letters*, 7th, vol. 37, no. 12, pp. 754 – 755, June 2001.
8. P. Gerken, ”Object-Based Analysis-Synthesis Coding of Image Sequences at Very Low Bit Rates,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol.4, pp. 228 – 235, June 1994.
9. Chung, J. W., Moon, J. H., and Kim, J. K.: ‘Shape information reduction based on contour prediction and shape coding type’, ISO/IEC/JTC1/SC29/WG11 MPEG95/0461, Dallas, TX, November, 1995.
10. Farzin Mokhtarian and Riku Soumela, ”Robust image corner detection through curvature scale space,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1376 – 1381, December 1998.
Author Index

Abe, Akinori     II-815
Abe, Jair Minoro     II-935, II-942
Abe, Koji     I-614
Abe, Norihiro     II-175, II-181
Aboudarham, J.     III-467
Achuthan, Manoj     II-469
Adachi, Yoshinori     II-71, II-77, II-89
Adaminides, Emmanuel     II-62
Adams, Rod     III-256
Adorno, Marcello Cattaneo     I-426
Aghdaei, Mohammad Hossein     III-278
Ahmad, Abdul Manan     II-587
Ahn, Byung-Ha     I-952
Ahn, Kye-Soon     III-869
Ahn, Seongjin     III-38
Aizawa, Teruki     II-156
Akamatsu, Norio     I-799, I-807, I-827, I-840, I-847, I-853, I-874, I-881, I-891, II-1045
Akashi, Takuya     I-799
Akima, Hisanao     I-1010
Akutagawa, Masatake     II-1060, II-1068, II-1074
Akutsu, Tomonori     II-920
Allan, J.     I-699
Alvarez, Julio César     I-1227
Alvarez, Luis     III-395
Amamiya, Makoto     I-124
Amamiya, Satoshi     I-124
Andreae, Peter     III-224
Anuchitkittikul, Burin     I-385
Aoe, Jun-ichi     I-530, I-541, I-549, I-558, II-567
Ao Ieong, Tony W.H.     II-572, II-594
Aoki, Keisuke     III-136
Aoki, Terumasa     I-140
Araki, Kenji     II-131
Areerak, Kongpan     I-1255, III-695
Areerak, Kongpol     III-695
Arentz, Will Archer     I-620
Arita, Daisaku     I-419
Ariton, Viorel     II-382
Asano, Akira     III-756
Asano, Chie Muraki     III-756
Ashidi, Nor     I-591
Asorey-Cacheda, R.     I-685
Athanassoula, Lia     III-404
Atlam, El-Sayed     I-530, I-541, I-549, I-567
Attkimmongkol, Kitt     III-643
Auli, Francesc     I-647
Aziz, Shalihatun Azlin     I-599
Baba, Norio     I-434, I-792
Bac, Le Hoai     II-708, II-1114
Baccigalupi, C.     III-426
Baek, Ihn-Han     II-219
Baek, Jonghun     III-610
Bajaj, Preeti     III-551
Baker, Jesse T.     I-232
Balachandran, Bala     II-469
Balducelli, Claudio     I-1242
Bandara, G.E.M.D.C.     II-698
Barnes, Stuart     II-28
Bar-Yam, Yaneer     III-809
Batty, Thayne     I-905
Bauer, Michael     I-1164
Bedini, L.     III-426
Belea, Radu     III-246
Bell, David     III-521
Benkhalil, A.     III-446, III-453, III-460
Bento, Carlos     III-678
Bertoli, Marcello     I-1171
Bertolotto, Michela     II-425
Bhowan, Urvesh     III-224
Bi, Yaxin     III-521
Bigelow, John     II-996
Bo, Feng     II-564
Bocaniala, C. D.     III-365
Bodyanskiy, Yevgeni     II-764, II-772
 Bölöni, Ladislaus     II-1121
Booyanunta, Natthaphan     III-529
Bosma, Albert     III-404
Bouvry, Pascal     I-727
Braun, K.     III-476
Bright, Glen     I-225
| Author          | Page Numbers |
|-----------------|--------------|
| Brumen, Boštjan | II-1039      |
| Brust, Matthias R. | I-734      |
| Burns, Alex     | I-148        |
| Caforio, Antonio| III-342      |
| Cairó, Osvaldo  | I-1227       |
| Cali, Andrea    | III-187      |
| Calvanese, Diego| III-187      |
| Camilleri, Guy  | II-9         |
| Cao, Cungen     | I-263, II-580|
| Caraman, Sergiu | III-246      |
| Carreiro, Paulo | III-678      |
| Ceravolo, Paolo | III-328, III-335|
| Cha, Jae Sang   | III-116      |
| Chan, Stephen   | III-232      |
| Chang, Jae-Woo  | II-975       |
| Chau, Rowena    | I-155        |
| Chen, Chun-Hua  | I-929        |
| Chen, Guanlong  | III-544      |
| Chen, Guoqing   | I-270, III-513|
| Chen, Junghuei  | I-25         |
| Chen, Kok Yeng  | III-357      |
| Chen, Yen-Wei   | II-337, II-352, II-359|
| Cheng, Jian     | II-344, II-352|
| Cheng, K.P.     | II-28        |
| Cheung, King Hong| II-572     |
| Chi, Sheng-Chai | II-1252      |
| Chiba, Masaya   | II-905       |
| Chiu, Chaochang | I-922, I-937 |
| Cho, Jae Hoon   | I-50         |
| Cho, Jae-Soo    | III-887      |
| Cho, Jundong    | III-573      |
| Cho, Yongjoo    | III-103      |
| Cho, Youngwan   | II-202       |
| Choi, Doo-Hyun  | III-862      |
| Choi, J. G.     | II-211       |
| Choi, Jae Gark  | II-253, III-887|
| Choi, Jaeseok   | III-166      |
| Choi, Woo-Kyoung| III-589      |
| Chuang, Zhang   | III-632      |
| Chun, Kwang Ho  | III-52       |
| Chung, Jinwook  | III-38       |
| Clark, David    | II-483       |
| Coghill, George | II-319       |
| Colucci, Simona | III-187      |
| Coppock, Sarah  | II-1136      |
| Corallo, Angelo | III-335, III-342|
| Corne, David    | I-952        |
| Cornelis, Chris | I-270, II-779|
| Cosgrove, Michael C. | I-232 |
| Costa-Montenegro, E. | I-685 |
| Cox, Robert     | III-210, III-566|
| Coyette, Adrien | I-1150       |
| Crampin, Edmund J. | II-329 |
| Cranefield, Stephen | II-1172 |
| Crespo, J.L.    | I-661        |
| Crippa, Paolo   | II-683       |
| Crowther, Patricia | III-210    |
| Czyzewski, Andrzej | I-743      |
| Dahalan, Azweeda| II-294       |
| Damiani, Ernesto| III-321      |
| D’Anjou, A.     | I-692        |
| Davey, Neil     | III-256      |
| Deaton, Russell | I-25         |
| De Cock, Martine| I-270, II-779|
| de Garis, Hugo  | I-905        |
| Deguchi, Toshinori| II-103     |
| de la Calleja, Jorge | III-411 |
| Denda, Tatsuaki | I-981        |
| De Oliveira, Marcos | II-1172 |
| Deselaers, Thomas| II-989      |
| Devedžić, Vladan| I-284, I-299 |
| Diegel, Olaf    | I-225        |
| Dietrich, Jens  | I-455        |
| Ding, Kai       | II-447       |
| Ding, Shuxue    | II-366       |
| Di Noia, Tommaso| III-187      |
| Dipoppa, Giovanni| I-1242     |
| Donini, Francesco M. | III-187 |
| Dote, Yasuhiko  | I-113        |
| Družovec, Marjan| II-1003, II-1039|
| Duro, R.J.      | I-661, I-669 |
| Dziubinski, Marek| I-743      |
| Eakins, John P. | I-614        |
| Eduards, Mark A. | I-793      |
| Eguchi, Hajime  | II-398       |
| Elia, Gianluca  | III-335, III-342|
| Emoto, Takahiro | II-1060      |
| Estévez, Pablo A. | I-140      |
| Falkowski, Bernd-Jürgen | I-675 |
| Faulkner, Stéphane | I-1150    |
Favela, Jesús     I-1181
Fernández-Redondo, Mercedes     I-677
Ferreira, José L.     III-678
Fleming, PJ     I-699
Flisar, Dušan     II-1018
Forgionne, Guiseppi A.     I-103
Fotouhi, F.     III-173
Frank, Celia     III-506
Frey, Hannes     I-718
Fuchino, Tetsuo     II-418
Fuentes, Olac     III-395, III-404, III-411
Fujii, Satoru     II-912, II-920
Fujita, Toshiharu     II-1208
Fujita, Yusuke     I-480, I-516
Fuketa, Masao     I-530, I-549, I-558, I-567,
Fukue, Yoshinori     I-501
Fukui, Shinji     II-632
Fukumi, Minoru     I-799, I-807, I-827, I-840, I-881, I-891, I-898, II-1045
Fukumoto, Taro     III-1
Fuller, N.     III-467
Funahashi, Kenji     II-110, II-632
Furumura, Takashi     II-898
Fyfe, Colin     I-16, I-74

Galitsky, Boris     III-314, III-307
Garcia, Fernando     I-647
Garzon, Max     I-18
Gašević, Dragan     I-284
Ghada, Elmarhomy     I-530
Gianini, Gabriele     III-321
Gil-Castiñeira, F.     I-685
Giripunje, Shubhangi     II-640
Glebocka, Agnieszka Dardzinska     II-1143
Goghill, George     II-284
Goh, John     III-795
Goh, T.T.     I-277
Gomes, Paulo     III-678
Gomez, Juan Carlos     III-404
Gonzales, Jorge     I-647
González-Castaño, F.J.     I-685
Goonesekera, Tharanga     II-1, III-772
Görgen, Daniel     I-734
Gorodetsky, Vladimir     I-1136
Gradecki, Joseph D.     I-637
Graham, Jonathan     III-388
Graña, M.     I-692
Grosky, W.I.     III-173
Grzymala-Busse, Jerzy W.     I-757
Gürgen, Daniel     I-718
Guan, Jiwen     III-521
Guan, Xiaoxiang     I-1017
Guinand, Frederic     I-727
Guitart, Pere     I-647
Güld, Mark O.     II-989
Guo, Wanwu     II-432, II-454
Guo, Ying     III-388

Ha, Sook-Jeong     II-219
Hagita, Norihiro     II-815
Håkansson, Anne     I-253
Hall, Richard     I-1023
Hamaguchi, Takashi     II-398
Hamakawa, Yasuo     III-855
Hamanaka, Hiroshi     I-1002
Hamamoto, Yoshihiro     I-608
Han, Hee-Seop     III-788
Han, Sun-Gwan     III-788
Han, W.Y.     III-843
Han, Youngshin     III-95
Hara, Akira     II-1089
Harada, Katsuyuki     II-181
Harada, Kouji     II-519
Hashimoto, Yoshihiro     II-398
Hassan, Halah     II-46
Hatanaka, Toshiharu     I-1189
Hatzilygeroudis, Ioannis     I-292, II-1106
Hayakawa, Yoshihiro     I-974, I-981
Hayama, Hiroshi     III-625
Hayashi, Hatsuho     I-967
Hayashi, Takahiro     III-180
He, Qiang     I-1037
Herd, Susanna     I-1023
Herrández-Espinosa, Carlos     I-677
Herranz, D.     III-426
Herzog, A.     III-476
Hien, Thai Duy     II-359
Higchi, Tetsuya     I-6
Hijazi, Mohd. Hanafi Ahmad     II-587
Hirakawa, Satoshi     III-756
Hirao, Masahiko II-418
Hirata, Toshiyuki I-322
Hirata, Yoshihiko II-898
Hirayama, Katsumi II-1215
Hirose, Kota II-188
Hogie, Luc I-727
Hong, S. H. II-211
Hong, Tzung-Pei II-1283
Hori, Koichi III-350
Hori, Satoshi II-175, II-181, II-188
Horie, Kenichi II-831
Horio, Yoshihiko I-988
Hoshikawa, Masafumi II-1074
Hou, Jia III-74
Howard, Daniel II-793, III-217
Hsu, Chao-Hsing I-929
Hsu, Chi-I I-922
Hsu, Han-Jen III-749
Hsu, Pei-Lun I-922
Hu, Hong I-772
Huang, Hung-Hsuan I-357
Huang, Jen-Peng II-1245
Huang, Jie II-366
Huang, Te Ming III-802
Hung, Shao-Shin II-1237
Hutter, Christian I-734
Hwang, Chang-Soon III-596
Hwang, Gi Yeon III-74
Hwang, Kyoung-Soon III-817
Ibrahim, Zuwairie I-32
Ichalkaranje, Nikhil I-71, I-80, I-110
Ichikawa, Teruhisa II-928
Ichimura, Hiroshi II-920
Ichimura, Takumi II-156, II-195, II-1081, II-1089, II-1097, II-1128
Ikai, Tomohiro I-113
Ikeda, Toshiaki I-337
Imai, Hideyuki I-1058
Imura, Takashi I-840
Inoue, Yoshio II-898
Inuzuka, Nobuhiro II-95
Ipsen, S.S III-446, III-453, III-460
Isa, Mat I-591
Iseyama, Yukari I-509
Ishida, Yoshiteru II-504, II-534
Ishigaki, Chuhei III-536
Ishii, Naohiro II-83, II-103, II-124
Ishikawa, Hiroshi I-178
Ishikawa, Ryuji II-954
Ishizaki, Masato I-330
Ismail, Ahmad Faris II-294
Isokawa, Teijiro III-491
Isomichi, Yoshinori II-1089
Ito, Shin-ichi I-853
Itoh, Toshiaki II-398
Iwahori, Yuji II-110, II-118, II-632
Iwamoto, Seiichi II-1201, II-1208
Iwao, Tadashige I-124
Iwata, Atsushi I-995
Iwata, Jun II-912
Iwata, Tomoharu II-624
Izworski, Andrzej III-740
Jain, Lakhmi I-74, I-80, I-786, II-949
Jang, Min-Soo III-196
Jefferies, Margaret E. I-232
Jeon, Hong-Tae III-589
Jeon, Inja II-227
Jeon, Jae Wook III-573
Jeon, Joongnam III-817
Jeremić, Zoran I-299
Jimbo, Takashi II-83
Jing, Ju III-419
Johansson, Christopher I-959
Jones, Harrison P. III-433, III-439
Jovanović, Jelena I-284
Jun, Guo III-632
Jun, Jeong III-869
Jun, Moon-Seog III-81
Jung, Mi Gyoung II-237, II-244
Juszczyszyn, Krzysztof II-1194
Kadirkamanathan, V. I-699
Kadmin, Ahmad Fauzan I-591
Kadoya, Yuki I-541, I-549, I-567
Kaewkasi, Chanwit I-1235
Kaewkasi, Pitchaya I-1235
Kailing, Karin II-982
Kakusho, Koh I-364
Kamal, M.A.S. I-1197
Kambayashi, Yasushi II-1010
Kameda, Yoshinari I-411
Kamiyama, Daisuke II-905
Kamiura, Naotake III-491
Kanda, Taki III-143
Kaneda, Yuji II-616
Kang, Hyun-Soo II-211, II-253, III-887
Kang, Ju-Hyun III-596
Kang, Seung Hwan II-261, III-8
Karacapilidis, Nikos II-62
Karatrantou, Anthi I-292
Karci, Ali I-946, II-268
Karunagaru, Stephen II-1045
Kasaev, Oleg I-1136
Kashiji, Shinkaku I-541, I-558, I-567
Katarzyniak, Radosław Piotr II-1150
Katayama, Karoru I-178
Kato, Tsuneaki II-148
Kato, Yoshikiko III-350
Kawaguchi, Tomoko I-434
Kawaguch, Masashi II-83
Kawahara, Daisuke I-393
Kawakatsu, Jun I-1100
Kawanaka, Haruki II-118, II-632
Kawaoka, Tsukasa III-650
Kawasaki, Hiroshi I-827
Kawata, Seiichi I-1107
Kawaura, Takayuki III-136
Kayano, Akifumi II-876
Kazienko, Przemysław II-1157
Kecman, Vojislav III-802
Kerre, Etienne I-270, II-779
Keskar, Avinash III-551
Kessoku, Masayuki I-523
Keyers, Daniel II-989
Khalid, Marzuki III-380
Khojulklang, Kasem I-1262
Khosla, Rajiv II-1, III-657, III-772
Kiguchi, Kazuo I-1092
Kim, Dong Hwa I-50, I-57, II-661
Kim, Dong-Hwee II-276
Kim, Dongwon II-716, III-596, III-603
Kim, Euntai II-202
Kim, Gwang-Hyun III-110
Kim, Hyeoncheol III-788
Kim, Hyuncheol III-38
Kim, Hyung Jin III-45, III-88
Kim, Jae-Bong III-788
Kim, Jong Tae III-573
Kim, Mi Young II-237, II-244
Kim, Min Kyung III-270
Kim, Nak-Hyun II-716
Kim, Sang-ho III-116
Kim, Sang-Jun III-203
Kim, Seong-Joo III-589
Kim, Soon-Ja II-276
Kim, Taewan III-45
Kim, Tai-hoon III-60, III-116
Kim, Yong-Guk III-196, III-203
Kimura, Naoki II-412
Kimura, Yuuya I-1189
Kinosita, Yosuke III-180
Kinouchi, Yohsuke II-1060, II-1068, II-1074
Kinshuk I-277
Kluchi, Yosuke II-961
Klein, Mark III-809
Klonis, Nectarios I-1023
Koda, Tomodo II-862
Koga, Takanori I-13
Kogure, Kiyoshi II-815
Kojima, Masanori II-898
Kokol, Peter II-1018, II-1025
Kolodyazhnii, Vitaliy II-764
Kolp, Manuel I-1150
Komatsu, Takanori I-371, I-378
Komeda, Takashi I-371
Komura, Kazunari II-110
Kondo, Tadashi II-1051
Kong, Chan Chi III-232
Konishi, Osamu III-780
Korekado, Keisuke I-995
Korkotyan, E. III-476
Kosaka, Takuya I-411
Kostek, Bozena I-750
Koutsojannis, Constantin II-1106
Koyama, Koji II-77
Koyama, Yukie II-110
Kozuma, Masafumi II-175
Kriegel, Hans-Peter II-982
Krishnamurthy, E.V. I-87, I-95
Krogh, Anders I-64
Król, Dariusz II-1165
Kryssanov, Victor I-364
Kube, K. III-476
Kubota, Naoyuki     I-1121
Kudo, Mineichi     I-1058, I-1065
Kudo, Yasuo     I-1079, I-1085
Kulworawanichpong, Thanatchai
I-1255, I-1262, III-695, III-710
Kumamoto, Satoru     II-1230
Kumamoto, Tadahiro     II-139
Kumsawat, Prayoth     III-643
Kunifuji, Susumu     I-337, I-322
Kuo, Huang-Cheng     II-1245
Kuo, Ting-Chia     II-1237
Kurano, Masami     II-1230
Kurihara, Masahito     I-1072
Kuroda, Chiaki     II-412
Kurohashi, Sadao     I-385, I-393
Kurosawa, Yoshiaki     II-156, II-1128
Kuruoğlu, E. E.     III-426
Kushiroyori, Noriyuki     II-807
Kusiak, Andrew     I-148
Kwon, Kisang     II-227
Lai, Chris     III-772
Lai, Hsin-Hsi     III-618
Lai, Wei-Cheng     II-1260
Lai, Weng Kin     II-284, II-294, III-357
Lai, Chris     III-657
Lam, H.F.     III-373
Lam, Toby H.W.     II-557
Lansner, Anders     I-959
Lau, K.H.     II-28
Lau, Sim Kim     II-261, III-8
Le, Kim     II-491
Lee, Byung-Joo     III-196
Lee, Chilgee     III-95
Lee, Dong Chun     III-110
Lee, Eric W.M.     III-373
Lee, Eun-ser     III-60
Lee, Geehyuk     III-610
Lee, Guee Sang     III-880
Lee, Huey-Ming     III-123
Lee, Hyo-Young     II-219
Lee, Hyun-Gu     III-196
Lee, Ickjai     I-196
Lee, Ji Hyong     III-573
Lee, Jong-Hee     III-81
Lee, Keon Myung     II-723, III-573, III-817
Lee, Keun-Wang     III-81
Lee, Kwang-Hyoung     III-81
Lee, Kwangyup     II-202
Lee, Kyoung Jun     II-646, II-668
Lee, Moon Ho     III-74
Lee, Raymond S.T.     II-549, II-557, II-594
Lee, Sang-Keon     III-88
Lee, Seung Wook     III-573
Lee, Shu-Yen     III-123
Lee, Si-Woong     II-211, II-253, III-887
Lee, Soek-Joo     III-196
Lee, Soobeom     III-45
Lee, Sung-Oh     III-203
Lee, Yangsun     II-202, III-826
Lee, Yeong-Chyi     II-1283
Lee, Yong-Hwan     III-67
Lehmann, Thomas M.     II-989
Lehnert, Johannes     I-718
Lenič, Mitja     II-1025
Letsche, Terry     I-148
Levachkine, Serguei     III-718
Li, Deren     III-513
Li, Deyi     III-513
Li, Gary C.L.     II-549
Li, Qiubang     III-657
Lim, Chee Peng     III-357
Lim, Myoung Seob     III-52
Lim, W.S.     II-305
Lin, In-Jou     II-1252
Lin, Wen-Yang     II-1276, II-1283
Lin, Yang-Cheng     III-618
Lin, Yi-Sen     II-1245
Lin, Zhongqin     III-544
Lindemann, Udo     I-1157
Litvan, Irene     II-1018
Liu, Chi     II-440
Liu, Damon Shing-Min     II-1237
Liu, Hugo     III-293
Liu, James N.K.     II-564, II-572
Liu, Min     II-71
Liu, Qingshan     II-344, II-352
Logan, Erica     I-1023
Loo, C.K.     II-305
López Ariste, Arturo     III-388
| Author                  | Pages          |
|------------------------|----------------|
| López-Peña, F.         | I-661, I-669  |
| Lovrek, Ignac          | I-1143         |
| Lu, Hanqing            | II-344, II-352 |
| Luo, Jun               | I-189          |
| Luo, Xiao              | III-498        |
| Ma, Bingxian           | II-580         |
| Ma, Songde             | II-344         |
| Ma, Xiaohang           | I-1051         |
| Ma, Zhiqiang           | II-454         |
| MacDonald, Bruce A.    | I-203          |
| Maghsoudi, Shahin      | II-36          |
| Malanushenko, Olena    | III-439        |
| Malowiecki, Michal     | II-1179        |
| Mao, Ching-Hao         | III-120        |
| Marinelli, Marco       | I-1242         |
| Martínez, Ana I.      | I-1181         |
| Matsumura, Takashi     | I-314          |
| Matsuda, Noriyuki      | II-175, II-181 |
| Matsugu, Masakazu      | I-995          |
| Matsui, Nobuyuki       | I-833, III-491 |
| Matsumoto, Hideyuki    | II-412         |
| Matsumoto, Hiroshi     | I-1114         |
| Matsumoto, Yoshiyuki   | III-159        |
| Matsumura, Naohiro     | II-839         |
| Matsumura, Yuji        | I-891          |
| Matsunaga, Naofumi     | I-608          |
| Matsushita, Mitsunori  | II-148         |
| Matsuyama, Hisayoshi   | II-375         |
| Maurer, Maik           | I-1157         |
| Mayiwar, Narin         | I-253          |
| Mazlack, Lawrence J.   | II-1136        |
| McClean, Sally         | I-171          |
| McSharry, Patrick E.   | II-329, III-483|
| Mera, Kazuya           | II-195, II-1128|
| Messom, Chris          | I-218          |
| Metzler, Richard       | III-809        |
| Michaelis, E.          | III-476        |
| Ming, Wu               | III-632        |
| Minoh, Michihiko       | I-364          |
| Mitani, Keiichiro      | I-480          |
| Mitani, Yoshihiro      | I-608          |
| Mitrovic, Antonija     | I-306          |
| Mitsukura, Kensuke     | I-807, I-827   |
| Mitsukura, Yasue       | I-807, I-827, I-847, I-853, I-874, I-881 |
| Miura, Hirokazu        | II-175, II-181 |
| Miura, Motoki          | II-883         |
| Miura, Yuka            | II-912         |
| Miyajima, Hiromi       | III-855        |
| Miyakoshi, Masaaki     | I-1058         |
| Miyawaki, Asuka        | II-800         |
| Mizukoshi, Noriyoshi   | I-487          |
| Mizuno, Tadanori       | II-898, II-912 |
| Mogami, Yoshio         | I-792          |
| Monavar, Hamid         | III-278        |
| Monroy, Raúl           | II-526         |
| Montero, Calkin A.S.   | II-131         |
| Mørch, Anders I.       | I-131          |
| Moreno, Marco          | III-718        |
| Mori, Koji             | I-988          |
| Morie, Takashi         | I-995          |
| Morihiro, Koichiro     | I-833          |
| Morishige, Hajime      | I-1205         |
| Morita, Kazuhiro       | I-530, I-541, I-549, I-558, I-567 |
| Morohashi, Kazuya      | I-1107         |
| Moshidi, Behzad        | III-559        |
| Motoda, Hiroshi        | II-800         |
| Mouhoub, Malek         | III-702        |
| Mrázová, Iveta         | I-1044         |
| Munemori, Jun          | II-869, II-876, II-891, II-905 |
| Murai, Tetsuya         | I-1079, I-1085 |
| Murase, K.             | II-968         |
| Murata, Junichi        | I-1197, I-1213 |
| Murata, Tadahiko       | I-1114, I-1128 |
| Murthy, V.K.           | I-87, I-95     |
| Na, Seungwon           | III-826        |
| Nævdal, Jan Eirik B.   | I-131          |
| Nagashino, Hirofumi    | II-1060, II-1068, II-1074 |
| Nakada, Kazuhiro       | II-920         |
| Nakagami, Jun-ichi     | II-1230        |
| Nakajima, Koji         | I-974, I-981, I-1010 |
| Nakamatsu, Kazumi      | II-954, II-961 |
| Nakamura, Yuichi       | I-401, I-411   |
| Nakano, Kenji          | I-472          |
| Nakano, Miyoko         | I-898          |
| Author Name               | Pages                              |
|--------------------------|------------------------------------|
| Nakano, Ryohei           | II-602, II-609                     |
| Nakao, Zensho            | II-359                             |
| Nakaoji, Kumiyo          | II-148                             |
| Nakasuka, Shin’ichi      | III-350                            |
| Nakaura, Kazuhiro        | I-840                              |
| Nakayama, Hirotaka       | I-441                              |
| Nam, J. Y.               | II-211                             |
| Nam, M. Y.               | III-833, III-843                   |
| Naoe, Yukihisa           | II-898                             |
| Nara, Yumiko             | II-823                             |
| Neel, Andrew             | I-18                               |
| Negoita, Mircea Gh.      | I-240, I-914                       |
| Ng, Vincent              | III-232                            |
| Ngah, Umi Kalthum        | I-599                              |
| Ngh, Nguyen Thanh        | II-1114                            |
| Nguyen, Ngoc Thanh       | II-1179                            |
| Nguyen, Tai              | I-1150                             |
| Niimi, Ayahiko           | III-780                            |
| Nishida, Toyoaki         | I-357, I-385, I-393                |
| Nishimoto, Kazushi       | I-314, I-330                       |
| Nishimura, Haruhiko      | I-833                              |
| Nishizaki, Takashi       | I-401                              |
| Nocerino, Maria Cristina | III-328                            |
| Nomura, Osamu            | I-995                              |
| Nonaka, Hidetoshi        | I-1072                             |
| Nowostawski, Mariusz     | II-1172                            |
| Ny, Bunna                | II-541                             |
| Oeda, Shinichi           | II-1097                            |
| Ogata, Ryo               | I-401                              |
| Ogawa, Tomoya            | I-95                               |
| Ogura, Kanayo            | I-330                              |
| Oh, Sun-Jin              | II-219                             |
| Ohno, Sumika             | II-869                             |
| Øhrn, Aleksander         | I-620                              |
| Ohsawa, Yukio            | I-11, II-786, II-807, II-823, II-831, II-839, II-847 |
| Ohta, Manabu             | I-178                              |
| Ohta, Yuichi             | I-401, I-411                       |
| Ohtsuka, Shoichiro       | I-371                              |
| Oka, Natsuki             | I-371, I-378                       |
| Okamoto, Masashi        | I-385, I-393                       |
| Okamoto, Takeshi        | II-534                             |
| Okuno, Takahide          | I-988                              |
| Omata, Sadao             | II-366                             |
| Omi, Takeshi             | I-178                              |
| Onai, Rikio              | III-180                            |
| Ong, C. W.               | III-16                             |
| Ong, M.                  | I-699                              |
| Ono, Masaki              | I-558                              |
| Ono, Osamu               | I-32                               |
| Ota, Yutaka              | II-398                             |
| Oysal, Yusuf             | III-581                            |
| Ozaki, Hiroshi           | II-869                             |
| Ozaki, Masahiro          | II-71, II-77, II-124               |
| Özen, Özgür              | I-583                              |
| Paiva, Paulo             | III-678                            |
| Palade, Vasile           | II-698, III-246, III-365           |
| Pan, Hongqi              | II-753                             |
| Panat, Anshish           | II-640                             |
| Pandya, Abhijit S.       | II-1051                            |
| Pappis, Costas P.        | II-62                              |
| Park, Gwi-Tae            | II-716, III-196, III-596, III-603  |
| Park, Hyun Seok          | III-270                            |
| Park, Kil-Houm           | III-862                            |
| Park, Kyoung S.          | III-103                            |
| Park, Seon Hee           | III-270                            |
| Park, Seong-Mo           | II-253                             |
| Park, Seon-Hee           | III-263                            |
| Park, Wonbae             | III-610                            |
| Park, Gwi-Tae            | III-203                            |
| Pedrycz, Witold          | I-807                              |
| Pensuwon, Wanida         | III-256                            |
| Penumatsa, Phani         | I-18                               |
| Pereira, Francisco C.    | III-678                            |
| Perry, Mark              | I-1164                             |
| Peters, James F.         | I-764                              |
| Peterson, Don            | III-314                            |
| Petrovsky, Nikolai       | III-566                            |
| Phillips-Wren, Gloria E. | I-71, I-103, I-110                 |
| Piattini, Mario          | I-1181                             |
| Pierrakeas, C.           | I-292                              |
| Ping, Chan Choyi        | I-599                              |
| Pirani, Massimiliano     | II-683                             |
| Polkowski, L.           | I-779                              |
| Popat, Deval             | II-691                             |
| Popescu, Theodor D.      | I-1220                             |
| Popov, S.                | II-772                             |
| Potgieter, Johan         | I-225                              |
| Author                        | Page(s)       |
|-------------------------------|---------------|
| Pousada Carballo, J.M.        | 685           |
| Povalej, Petra                | 1018, 1025    |
| Pritchard, David              | 240, 914      |
| Prügel-Bennett, Adam          | 64            |
| Puangdownreong, Deacha        | 710           |
| Puketa, Masao                 | 549           |
| Purvis, Martin                | 1172          |
| Purvis, Maryam                | 1187          |
| Qiu, Bin                      | 1017          |
| Qu, Ming                      | 419           |
| Quintero, Rolando             | 718           |
| Ra, Ilkeyun                   | 637           |
| Rajasekaran, Sanguthevar      | 189           |
| Ramanna, Sheela               | 764           |
| Ramli, Dzati Athiar           | 591           |
| Ranawana, Romesh              | 698           |
| Rao, M.V.C.                   | 305           |
| Rashidi, Farzan               | 653, 738, 745 |
| Rashidi, Mehran               | 653, 738, 745 |
| Rees, David                   | 388, 419      |
| Ren, X.                       | 699           |
| Resta, Marina                 | 426           |
| Rhee, Phill-Kyu               | 227, 833, 843 |
| Rhee, Sang-Surm               | 67            |
| Riaño, David                  | 1039          |
| Rodríguez, Oscar M.           | 1181          |
| Rodríguez-Hernández, P.S.     | 685           |
| Rose, John A.                 | 8, 40         |
| Rosengard, Jean-Marc          | 31            |
| Rothkugel, Steffen            | 734           |
| Roy, Debabrata                | 614           |
| Ruttkowski, Tomasz M.         | 364           |
| Ryu, Jeha                     | 210           |
| Sa da Costa, J.               | 365           |
| Sado, Nobuaki                 | 118           |
| Saito, Kazumi                 | 602, 616, 624 |
| Saito, Toshimichi             | 1002          |
| Sakai, Sanshiro               | 912           |
| Sakakibara, Tsuneki           | 831           |
| Sakamoto, Katsuhiro           | 847           |
| Sakamoto, Masaru              | 398           |
| Salami, Momoh-Jimoh E.        | 294, 312      |
| Salerno, E.                   | 426           |
| Salmenjoki, Kimmo             | 1032          |
| Samoilov, Vladimir            | 1136          |
| Sanada, M.                    | 1085          |
| Sasaki, Hiroshi               | 124           |
| Sato, Eri                     | 1100          |
| Sato, Hideaki                 | 847           |
| Sato, Shigeo                  | 1010          |
| Sato, Shigeyuki               | 504           |
| Sato, Y.                      | 1085          |
| Sato, Yoichi                  | 385           |
| Satoh, Hironobu               | 866           |
| Savarimuthu, Roy              | 1187          |
| Sawamura, Hajime              | 1-1           |
| Schnell, Santiago             | 329           |
| Schönauer, Stefan             | 982           |
| Schubert, Henning             | 989           |
| Schwitter, Rolf               | 711           |
| Scotney, Bryan                | 171           |
| Seco, Nuno                    | 678           |
| Seo, Sam-Jun                  | 716, 603      |
| Serneniuk-Polkowska, M.       | 779           |
| Serra-Sagrista, Joan          | 647           |
| Shadabi, Fariba               | 566           |
| Shafawi, Mohd                 | 380           |
| Shahjahan, Md.                | 968           |
| Shapcott, Mary                | 171           |
| Sharda, Hema                  | 691           |
| Sharma, Dharmendra            | 469, 476, 498 |
| Shih, Zhongzhi                | 772           |
| Shibata, Tomohide              | 393           |
| Shigei, Noritaka              | 855           |
| Shigenobu, Tomohiro           | 869, 876      |
| Shih, Frank                   | 419           |
| Shim, Choon-Bo                | 975           |
| Shimizu, Koichi               | 511           |
| Shimizu, Toru                 | 898           |
| Shimooka, Toshiyuki           | 511           |
| Shin, Chi-Hyun                | 88            |
| Shin, Jungpil                 | 165           |
| Shin, Kyung-shik              | 646, 668      |
| Shin, Miyounge                | 263           |
| Shin, Myong-chul              | 116           |
| Name                        | Pages |
|-----------------------------|-------|
| Shioya, Yasuo               | I-178 |
| Shiraki, Wataru             | I-441 |
| Shizuki, Buntaoru           | II-883|
| Shon, Min-Kyu               | I-1213|
| Shu, Wen-Lung               | II-1260|
| Shukri, Mohamad             | III-380|
| Si, Jin X                  | I-263, II-580|
| Siegel, Howard Jay          | II-17 |
| Sil, Jaya                   | III-24 |
| Sing, Push                  | III-293|
| Sinkovic, Vjekoslav         | I-1143|
| Sioutis, Christos           | I-80  |
| Smith, Kate A.              | I-155 |
| Soak, Sang-Moon             | I-952 |
| Sohn, Bong Ki               | III-573|
| Sokolov, Alexander          | II-731 |
| Solazzo, GianiLuca          | III-342|
| Son, Bongsoo                | III-45, III-88 |
| Song, Young-Chul            | III-862|
| Sospedra, Joaquin Torres    | I-677 |
| Spitzer, Klaus              | II-989 |
| Spravedlyyy, V.             | III-476|
| Squire, David McG.          | II-996 |
| Sreenath, D.V.              | III-173|
| Srikaew, Arhit              | III-643|
| Srithorn, Phinit            | I-1262|
| Stefanowski, Jerzy          | I-757 |
| Stiglic, Bruno              | II-1018|
| Štiglic, Gregor             | II-1018, II-1025|
| Stranieri, Andrew           | I-1171|
| Sturm, Peter                | I-718 |
| Suena, Shinya               | I-974 |
| Suetsake, Noriaki           | III-536|
| Sugiyama, Kozo              | I-314 |
| Sui, Yuefei                 | I-263 |
| Sujitjorn, Sarawut          | I-1255, III-643, III-695, III-710 |
| Suka, Machi                 | II-1081|
| Sumi, Yasuyuki              | I-357 |
| Sumitomo, Toru              | I-541, I-549, I-558 |
| Sun, Baiqing                | I-859 |
| Suzuki, Atsuyuki            | II-954, II-961|
| Suzuki, Shoji               | II-89 |
| Syed Mustapha, S.M.F.D.     | I-343, I-350 |
| Sztandera, Les M.           | III-506|
| Tachiki, Masato             | I-393 |
| Tadeusiewicz, Ryszard       | III-740|
| Taguchi, Masashi            | II-786 |
| Takada, Kenji               | I-1128 |
| Takagi, Masato              | III-166|
| Takahama, Tetsuyuki         | II-1089|
| Takahashi, Fumihiko         | II-1068|
| Takahashi, Hiroshi          | I-494 |
| Takahashi, Ken'ichi         | I-124 |
| Takahashi, Koichi           | II-839 |
| Takahashi, Masakazu         | I-487, I-523|
| Takahashi, Satoru           | I-494, I-509|
| Takahashi, Takehisa         | III-1 |
| Takeda, Atsushi             | II-165 |
| Takeda, Fumiaki             | I-859, I-866, I-891 |
| Takeda, Kazuhiro            | II-375 |
| Takeoka, Saori              | II-77 |
| Taki, Hirokazu              | II-175, II-181, II-188 |
| Takigawa, Ichigaku          | I-1058 |
| Takimoto, Hironori          | I-874 |
| Takimoto, Munehiro          | II-1010|
| Tamura, Hiroshi             | II-847 |
| Tanahashi, Yusuke           | II-602 |
| Tanaka, Akira               | I-1058 |
| Tanaka, Jiro                | II-883 |
| Tanaka, Katsuaki            | III-350|
| Tanaka, Koki                | II-53 |
| Tanaka, Shogo               | I-1205 |
| Tanaka, Takushi             | II-53 |
| Tanaka-Yamawaki, Mieko      | I-449 |
| Taniar, David               | II-691, III-795 |
| Taniguchi, Kazuhiko         | I-1121 |
| Taniguchi, Ken              | I-464 |
| Taniguchi, Rin-ichiro       | I-419 |
| Tanimoto, Satoshi           | II-609 |
| Tateyama, Takeshi           | I-1107 |
| Tay, J. C.                  | III-16 |
| ten Brinke, Walter          | II-996 |
| Terano, Takao               | I-464, I-472 |
| Terlevich, Roberto          | III-395|
| Thai, Le Hoang              | II-708 |
| Thatcher, Steve             | I-74  |
| Thompson, HA                | I-699 |
| Tilley, Leann               | I-1023 |
| Tomita, Shigeyuki           | II-405 |
| Tonazzini, A.               | III-426|
| Tony, Bastin                | II-1187 |
Torikai, Hiroyuki  I-1002
Torres, Miguel  III-718
Tran, Dat  II-476, II-498
Tronci, Enrico  I-1242
Tsao, Chanhsi  I-937
Tseng, Ming-Cheng  II-1276
Tsuboi, Yusei  I-32
Tsuda, Kazuhiko  I-480, I-487, I-494, I-501, I-509, I-516, I-523
Tsuge, Yoshifumu  II-375
Tsurusaki, Kazuyoshi  II-1201
Uchino, Eiji  III-536
Ueda, Atsushi  I-1121
Ueda, Kazuhiro  I-371, I-378, III-625
Ueno, Takayuki  II-1208
Umeno, Masayoshi  II-83
Uosaki, Katsuji  I-1189
Urlings, Pierre  I-80
Ursu, Marian F.  III-31, III-764
Usuki, Masao  I-314
Utsunomiya, Atsushi  I-378
Velásquez, Juan D.  I-140
Vemulapali, Balaji  III-506
Vera, Eduardo  I-140
Virginas, Botond  III-764
Viviani, Marco  III-328
Vizcaíno, Aurora  I-1181
Vlček, Miroslav  III-726, III-733
Voudouris, Chris  III-764
Wada, Takao  II-418
Wagenknecht, Michael  II-731
Walton, Chris  II-920
Wang, Changhui  III-702
Wang, Dianhui  I-1051
Wang, Haimin  III-419
Wang, Hong-Ming  III-749
Wang, Jhing-Fa  III-749
Wang, Min-Feng  II-1276
Wang, Shuliang  III-513
Wang, Xizhao  I-1037
Wang, Pei  III-285
Washida, Yuichi  II-847
Washio, Takashi  II-800
Watabe, Hirokazu  III-650
Watada, Junzo  III-129, III-136, III-151, III-159, III-166
Watanabe, Takayuki  II-534
Watanabe, Teruyuki  III-129
Watanabe, Yuji  II-504
Watman, Craig  II-491
Watson, Ian  I-575, I-1249, II-36, II-46, III-672
Weerasinghe, Amali  I-306
Wei, Daming  II-366
Wein, Berthold  II-989
Welzer, Tatjana  II-1003, II-1025, II-1032, II-1039
Wilk, Szymon  I-757
Wills, Anna  I-575
Wojcik, Jaroslaw  I-750
Won, Kyoung-Jae  I-64
Wong, Ka Yan  I-654
Wszelek, Wieslaw  III-740
Wu, Annie S.  II-17
Wu, Chen-Cheng  II-1260
Wu, Chih-Hung  II-1268
Wu, Huanrui  I-1030
Wun, Chian-Huei  II-1268
Xiao, Jitian  II-461
Xie, Nengfu  II-263, II-580
Xu, Baishan  II-447
Yabuuchi, Yoshiyuki  III-151
Yada, Katsutoshi  II-800
Yama, Hisaaki  II-405
Yamada, Kunihiro  II-898, II-920
Yamada, Masafumi  I-1065
Yamaguchi, Takashi  II-831
Yamaguchi, Toru  I-1100
Yamakami, Toshihiko  II-855
Yamakawa, Takeshi  I-13
Yamamoto, Yasuhiro  II-148
Yamashita, Yoshiyuki  II-391
Yamato, Kazuharu  III-491
Yamawaki, Shigenobu  I-786, II-949
Yan, Min-Chuan   II-1252
Yan, Peng   I-270
Yang, Zhiyi   I-630
Yasuda, Hiroshi   I-140
Yasuda, Masami   II-1230
Yasukata, Fumiko   I-898
Yeap, Wai-Kiang   I-232
Yee, Paul   II-319
Yeh, Chung-Hsing   I-155, II-753, III-618
Yip, Angela Y.N.   III-665
Yip, Chi Lap   I-654
Yip, Daniel C.Y.   II-28
Yoneyama, Mika   I-113
Yong, Shereen   II-284
Yoo, Ju-Hyoung   III-869
Yoo, Seong-Joon   I-164
Yoo, Seung-Jae   III-110
Yoon, Hyo Sun   III-880
Yoon, Jungwon   I-210
Yoon, Min   I-441
Yoshida, Hajime   II-891
Yoshida, Jun   I-516
Yoshida, Katsumi   II-1081, II-1097
Yoshida, Kenichi   I-516
Yoshida, Koji   II-898
Yoshida, Kouji   II-912, II-920
Yoshida, Motoharu   I-967
Yoshida, Yuji   II-1222, II-1230
Yoshimori, Seiki   I-881
Yoshino, Takashi   II-869, II-876, II-891, II-905
Yoshioka, Hitoshi   II-405
You, Il-Sun   III-67
You, Jane   II-572
Yu, Ge   I-1030
Yu, Han   II-17
Yu, Hui   II-440
Yu, Zhiwen   I-630
Yuan, Fang   I-1030
Yuan, Hanning   III-513
Yue, Xiaoli   I-263
Yuen, David C.K.   I-203
Yuizono, Takaya   II-876
Yun, Byoung-Ju   II-211, II-253, III-610, III-887
Yun, Yeboon   I-441
Yusuf, Rubiyah   III-380
Zabidi, Suriza Ahmad   II-312
Zadeh, Lotfi A.   I-1
Zahradnik, Pavel   III-726, III-733
Zalaket, Joseph   II-9
Zeephongsekul, Panlop   III-529
Zeng, Xiang-Yan   II-337
Zhang, Chunxia   II-580
Zhang, Mengjie   II-541, III-224
Zhang, Qinyu   II-1074
Zhang, Yansong   III-544
Zharkov, S.I.   III-446, III-453, III-460
Zharkova, V.V.   III-446, III-453, III-460
Zheng, Zheng   I-772
Zhong, Guoqiang   I-124
Zhou, Min   II-425
Zhou, Xingshe   I-630
Zilli, Antonio   III-335
Zincir-Heywood, A. Nur   III-498
Zyzalo, Jonathan R.   I-225