Abstract

Text and not-text segmentation and text line extraction from document images are the most challenging problems of information indexing of Arabic document images such as books, technical articles, business letters and faxes in order to successfully process them in systems such as OCR. Researches on Arabic language related to documents digitization have been focusing on word and handwriting recognition. Few approaches have been proposed for layout analysis for Arabic scanned/captured documents. In this paper we present a page segmentation method that deals with the complexity of the Arabic language characteristics and fonts using the combination between two algorithms. The first method is the Run length Smoothing. The second method is the Connected Component Labeling algorithm for text and non-text classification using SVM. The combination of the two methods is based on Anding and Oring operations between the outputs of the two methods based on certain conditions. Then, dynamic horizontal projection based on dynamic updating of the threshold to commensurate with the noise associated with different documents and in between text lines. The performance evaluation is performed using manually generated ground truth representations from a dataset of Arabic document images captured using cameras and a hardware built for this purpose.
Evaluation and experimental results demonstrate that the proposed text extraction method is independent from different document size, text size, font, shape, and is robust to Arabic document segmentation and text lines extraction.

References

- H. E. Abed and V. Margner, "ICDAR 2009–Arabic handwriting recognition competition," Int. Journal on Document Analysis and Recognition, vol. 14, pp. 3–13, 2011.
- K. Y. Wong, R. G. Casey and F. M. Wahl, "Document analysis system," IBM J. Res. Devel., Vol. 26, NO. 6,111). 647-656, 1982.
- K Zagoris, N Papamarkos, Text Extraction Using Document Structure Features And Support Vector Machines, in Proceedings of the 11th IASTED International Conference on Computer Graphics and Imaging, (2010)
- N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Systems, Man, and Cybernetics, 1979, 9, pp. 62-66.
- K. S. Kumar, S. Kumar, and C. Jawahar, "On segmentation of documents in complex scripts," in 9th Int. Conf. on Document Analysis and Recognition, Brazil, Sep. 2007, pp. 1243–1247.
- Bukhari, S. S.; Shafait, F.; Breuel, T. M.; , "High Performance Layout Analysis of Arabic and Urdu Document Images," Document Analysis and Recognition (ICDAR), 2011 International Conference on , vol. , no. , pp. 1275-1279, 18-21 Sept. 2011
- G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," Computer, vol. 7, no. 25, pp. 10–22, 1992.
- H. S. Baird, "Background structure in document images," inDocument Image Analysis, H. Bunke, P. Wang, and H. S. Baird, Eds. World Scientific, Singapore, 1994, pp. 17–34.
- T. M. Breuel, "Two geometric algorithms for layout analysis," in Proc. Workshop on Document Analysis Systems, Princeton, NY, USA, Aug. 2002, pp. 188–199.
- L. O’Gorman, "The document spectrum for page layout analysis," IEEE TPAMI, vol. 15, no. 11, pp. 1162–1173, 1993.
- K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area Voronoi diagram," Computer vision and Image Understanding, vol. 70, no. 3, pp. 370–382, 1998.
- F. Shafait, D. Keysers, and T. M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," IEEE TPAMI, vol. 30, no. 6, 2008.
- F. Shafait, A. Hasan, D. Keysers, and T. M. Breuel, "Layout analysis of Urdu document images," in 10th IEEE Int. Multitopic Conference, INMIC’06, Islamabad, Pakistan, Dec. 2006.
- T. M. Breuel, "Two geometric algorithms for layout analysis," in Proc. Workshop on Document Analysis Systems, Princeton, NY, USA, Aug. 2002, pp. 188–199.
- W. Boussellaa, A. Zahour, H. E. Abed, A. Benabdellahfied, and A. M. Alimi, "Unsupervised block covering analysis for textline segmentation of arabic ancient handwritten document images," in ICPR, Istanbul, Turkey, 2010, pp. 1929–1932.
- Chih-Chung Chang, and Chih-Jen Lin, LIBSVM : a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
- P. Shivakumara, G. Hemantha Kumar, D. S Guru, P. Nagabhushan, Skew Estimation of Binary Document Images Using Static and Dynamic Thresholds Useful for Document Image Mosaicing. National Workshop on IT Services and Applications (WITSA2003) Feb 27-28, 2003.

- Effective Text Extraction from Video Scenes. E. H. Shaheen, K. M. El Sayed, S. H. Ahmed, ... p119-p134. Genetic Programming Scheme for Optimizing Register Allocation.
- J. Liang, R. Rogers, R. M. Haralick, and I. T. Phillips. UW-ISL document image analysis toolbox: An experimental environment. In In Proc. 4th Int’l Conf. on Doc. Analysis and Reco., pages 984–988, 1997.

**Index Terms**

Computer Science  
Pattern Recognition

**Keywords**

Arabic document layout analysis  
Arabic text line extraction