Abstract

Optical Character Recognition (OCR) technology in converting an image containing text to an editable text format is of high sense in document image processing. Input to OCR could be a scanned document, or a simple newspaper cut-out. Supervised Learning using Neural Networks yield the output with greater accuracy. Unlike English, Kannada Language has a huge set of characters as it includes kaagunithas, vattaksharas, etc. This makes recognition of the characters much more complex. The paper mainly concentrates on OCR for the Kannada Text which goes through a threshold as a first step converting input image into binary image, making segmentation easier. Characters can be extracted from the documents using various Segmentation methods. The vattaksharas are extracted/differentiated from the words by using base-line technique. When the characters are recognized, they are compared with Unicodes available on the system and then printed. In the above method, CNN plays a pivotal role in reading the character and comparing it with the Unicode look up table values to print the output.
This system has been tested with varying fonts. A total number of 37 sample documents are used for experimentation. The system has been developed for only printed Kannada Text.

References

1. HR Mamatha, S Sucharitha, Srikanta Murthy, “Multi-font and Multi-size Kannada Character Recognition based on the Curvelets and Standard Deviation”, International Journal of Computer Applications, Foundation of Computer Science, New York, USA, 2011.
2. R Prajna, VR Ramya, HR Mamatha “A study of different text line extraction techniques for multi-font and multi-size printed kannada documents”, International Journal of Computer Applications, Foundation of Computer Science, 2015.
3. M.K Jindal, R. K. Sharma & G.S. Lehal, “Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts”, International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.3, No.4 (2007), pp. 277–286
4. Ashwin T.V and P.S Sastry, “A font and size independent OCR system for printed Kannada using SVM”, Sadhana, vol. 27, Part 1, February 2002, pp. 35–58.
5. Anil. K. Jain, “Feature Extraction methods for Character Recognition – A survey”, Pattern Recognition Volume 29, Issue 4, April 1996, Pages 641-662
6. K. Indira, S. Sethu Selvi, “Kannada Character Recognition System: A Review”, InterJRI Science and Technology, Vol. 1, Issue 2, July 2009
7. Netravati Belagali, Shanmukhappa A. Angadi, “OCR for Handwritten Kannada Language Script”, International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 08; August – 2016.
8. C V, Aravinda, “Kannada handwritten character recognition using multi feature extraction techniques”. International Journal of Science and Research (IJSR). Vol 10, 2014
9. Shashikala Parameshwarappa1, B.V.Dhanda, “Basic Kannada Handwritten Character Recognition System using Shape Based and Transform Domain Features”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015
10. M. Vishwaas, M. M. Arjun and R. Dinesh, "Handwritten Kannada character recognition based on Kohonen Neural Network," 2012 International Conference on Recent Advances in Computing and Software Systems, Chennai, 2012, pp. 91-97.
11. G. Keerthi Prasad, I. Khan, N. R. Chanukotimath and F. Khan, "On-line handwritten character recognition system for Kannada using Principal Component Analysis Approach: For handheld devices," 2012 World Congress on Information and Communication Technologies, Trivandrum, 2012, pp. 675-678.
12. Gururaj mukarambi, dhanda b.v, mallikarjun hangarge, “recognition system for handwritten and printed kannada numerals and vowels”, International Journal of Machine Intelligence ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 4, 2011, pp-259-262.
13. CS231n Convolution Neural Networks for Visual Recognition; http://cs231n.github.io/convolutional-networks/

Index Terms

Computer Science

Artificial Intelligence
Keywords

Base-line Identification, CNN, Kannada, Neural Network, Optical Character Recognition, Pre-processing, Python, Segmentation.