Abstract

Automatic handwritten script identification from document images facilitates many important applications such as sorting, transcription of multilingual documents and indexing of large collection of such images, or as a precursor to optical character recognition (OCR). In this paper, we investigate a texture as a tool for determining the script of handwritten document image, based on the observation that text has a distinct visual texture. Further, K nearest neighbour algorithm is used to classify 300 text blocks as well as 400 text lines into one of the three major Indian scripts: English, Devnagari and Urdu, based on 13 spatial spread features extracted using morphological filters. The proposed algorithm attains average classification accuracy as high as 99.2% for bi-script and 88.6% for tri-script separation at text line and text block level respectively with five fold cross validation test.
Reference

- B.B.Chaudhuri and U.Pal,” An OCR system to read two Indian language scripts: Bangla
  and Devnagari (Hindi),” Proc. of 4th ICDAR, Uhn. 18-20 August 1997.
- Santanu Chaudhury, Gaurav Harit, Shekar Madnani, R.B.Shet,” Identification of scripts of
  Indian languages by Combining trainable classifiers,” Proc. of ICVGIP, Dec-20-22, Bangalore,
  India, 2000.
- D.Dhanya, A.G Ramakrishnan and Peeta Basa pati, “Script identification in printed
- J. Hochberg, P. Kelly, T Thomas and L Kerns, “Automatic script identification from
  document images using cluster-based templates,” IEEE Trans. on Pattern Analysis and
- Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, “Script and
  Words from a Bilingual Document Employing Gabor Features and Radial Basis Function Neural
- M.C.Padma and P. Nagabhushan,” Script Identification and separation of text words of
  Kannada Hindi and English languages through discriminating features,” Proc. of NCDA2003,
- G.S.Peake and Tan, “Script and language identification from document images,” Proc. of
- S. Basavaraj Patil and N.V.Subbareddy, “Neural network based system for script
- Peeta Basa pati, S. Sabari Raju, Nishikanta Pati and A.G. Ramakrishnan, “Gabor filters
- P. Nagabhushan, S.A. Angadi and B.S. Anami,” An Intelligent Pin code Script
  Identification Methodology Based on Texture Analysis using Modified Invariant Moments,” Proc. of
- T.N.Tan, “Rotation invariant texture features and their use in automatic script
  identification,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, pp.751-756,
  1998.
- S. Wood. X. Yao. K.Krishnamurthi and L.Dang ”Language identification for printed text

Index Terms

Computer Science  Image Analysis

Key words
Script Identification offline handwritten
documents
Optical character reader

cross validation