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


**Index Terms**

Computer Science  
Image Analysis

**Key words**

Script Identification  
offline handwritten documents
Optical character reader

cross validation