Abstract

There are immense efforts to design a complete OCR for most of the world’s leading languages, however, multilingual documents either of handwritten or of printed form. As a united attempt, Unicode based OCRs were studied mostly with some positive outcomes, despite the fact that a large character set slows down the recognition significantly. In this paper, we come
out with a method to classify words to a language as the word segmentation is complete. For the purpose, we identified the characteristics of writings of several languages and utilized projecting method combined with some other feature extraction methods. In addition, this paper intends a modified statistical approach to correct the skewness before processing a segmented document. The proposed procedure, evaluated for a collection of both handwritten and printed documents, came with excellent outcomes in assigning words to languages.

Reference

Barbara, CA, USA, 2006.

Index Terms

Computer Science
Pattern Recognition

Key words
OCR
Multilingual OCR
MOCR
Classification