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
Multilingual OCR (MOCR): An Approach to Classify Words to Languages

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

  Handwritten Line Segmentation”, in Document Recognition and Retrieval XIV, Proceedings of
- Richard O. Duda, Peter E. Hart: “Use of the Hough Transform to detect lines and curves
  in pictures”, Technical Note 36, AI Center, April 1971.
- Srihari, S.N. and V. Govindaraju: “Analysis of textual images using the Hough
- D.S. Le, G.R. Thoma and H. Wechsler: “Automatic page orientation and skew angle
- Mohammad Abu Obaida, Tanay Kumar Roy, Md. Abu Horaira and Md. Jakir Hossain:
  “Skew Correction Function of OCR: Stroke-Whitespace based Algorithmic Approach”,
- Akiyama, T. and N. Hagita: “Automated entry system for printed documents”, Pattern
- Omar, K., A. Ramli, R. Mahmod and M. Sulaiman: “Skew detection and correction of jawi
- Hashizume, A., P.S. Yeh and A. Cosenfeld: “A method of detecting the orientation of
- Yan, H.: “Skew correction of document images using interline cross correlation”,
- Laurence Likforman-Sulem, Anahid Hanimyan, Claudie Faure: “A Hough Based Algorithm
  for Extracting Text Lines in Handwritten Documents”, Proceedings of the Third International
- G. Louloudis, K. Halatsis, B. Gatos, I. Pratikakis: “A Block-Based Hough Transform
  Mapping for Text Line Detection in Handwritten Documents”, 10th International Workshop on
  Frontiers in Handwriting Recognition (IWFHR 2006), La Baule, France, pp. 515-520, October
  2006.
- Y. Pu and Z. Shi: “A Natural Learning Algorithm Based on Hough Transform for Text
  Lines Extraction in Handwritten Documents”, Proceedings of the 6th International Workshop on
- Esra Ataer, Pinar Duygulu: “Retrieval of Ottoman Documents”, Proceedings of 8th ACM
  SIGMM International Workshop on Multimedia Information Retrieval , October 26-27, Santa
Barbara, CA, USA, 2006.
Multilingual OCR (MOCR): An Approach to Classify Words to Languages


Index Terms

Computer Science

Pattern Recognition

Key words

OCR
Multilingual OCR
MOCR

Classification