A Pre-processing Method for Printed Tamil Documents: Skew Correction and Textual Classification

M. Ramanan[1], A. Ramanan[2] and E.Y.A. Charles[2]

[1]Department of Computer Science, Trincomalee Campus, Eastern University, Sri Lanka
[2]Department of Computer Science, University of Jaffna, Jaffna, Sri Lanka
mramanan1981@gmail.com, { a.ramanan, charles.ey}@jfn.ac.lk

Abstract

An Optical character recognition (OCR) consists of the phases: preprocessing and segmentation, feature extraction, classification and post-processing. This paper focuses on pre-processing and segmentation tasks which plays a major role in the subsequent processes of an OCR. The objective of pre-processing and segmentation is to improve the quality of the input image. In addition this phase removes unnecessary portions of the input image that would otherwise complicate the subsequent steps of OCR and reduce the overall recognition rate. Preprocessing and segmentation step consists many sub processes namely, image binarisation, noise removal, skew detection and correction, page segmentation, text or non-text classification, line segmentation, word segmentation and character segmentation. This paper proposes a new approach to calculate the skew angle, segment and classify the blocks as text or non-text. The skew angle is calculated on the scanned document using Wiener filter, smearing technique and Radon transform. Document image is segmented into blocks using run length smearing algorithm and connected component analysis. Features such as basic, density and HOG are extracted from each block for text and non-text classification. The proposed methods are tested on 54 documents. The testing results show a recognition rate of 96.30% for skew detection and correction whereas the recognition rate is 99.18% for text or non-text classification with binary SVMs using RBF kernel.

Author Keywords

Printed Tamil documents; Skew correction; Textual classification