A Performance Comparison and Post-processing Error Correction Technique to OCRs for Printed Tamil Texts

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

Optical Character Recognition (OCR) deals with automated recognition of characters that are in the format of digital image. OCR refers to the process by which scanned images are electronically processed and converted to an editable document. Handwritten and printed texts are the primary research areas of an OCR. Many OCR systems are commercially available for English and Arabic characters but there is still no recognition system available which yields higher recognition rate even though the scanned images are of high quality. The general framework of a Tamil OCR in the literature involves: preprocessing, line segmentation, word segmentation, character segmentation, feature extraction and recognition of characters. OCR for printed Tamil documents poses challenge owing to: one line may have different font styles, presence of pictures, multi columns, touching of adjacent characters, presence of broken characters, low print quality and complex layout. Furthermore, when comparing 26 alphabets in English, Tamil language has 247 alphabets which makes the recognition more difficult. There are few OCRs for Tamil language that are freely available with a moderate recognition rate as the performance comparisons of such OCRs are not available on a benchmark dataset. In this paper we compare OCRs for printed Tamil texts on four different types of documents: books, magazines, newspapers and pamphlets. Furthermore we propose a post-processing error correction technique to the tested OCRs which reduces the overall mean error rate by nearly 10% on those four categories.

Author Keywords

Google tesseract, i2ocr, newocr, ponvizhi, TamilOCR