Sinhala Character Recognition using Tesseract OCR

Manisha U.K.D.N.[a], Liyanage S.R.[b]

[a]Department of Computer Systems Engineering, Faculty of Computing and Technology, University of Kelaniya, Sri Lanka.
[b]Department of Software Engineering, Faculty of Computing and Technology, University of Kelaniya, Sri Lanka.

In Sri Lanka, there are many fields that uses Sinhala scripts, such as newspaper editors, writers, postal and application processes. In these fields there have only a scanned or printed copies of Sinhala script, where they have to enter them manually to a computerized system, which consumes much time and cost. The proposed method was consisted of two areas as image pre-processing and training the OCR classifier. In Image pre-processing, the scanned images were enhanced and binarized using image processing techniques such as grayscale conversion and binarization using local thresholding. In order to avoid distortions of scanned images such as water marks and highlights was removed through the grayscale conversion with color intensity changes. In the OCR training, the Tesseract OCR engine was used to create the Sinhala language data file and used the data file with a customized function to detect Sinhala characters in scanned documents. OCR engine was primarily used to create a language data file. First, pre-processed images were segmented (white letters in black background) using local adaptive thresholding where performing Otsu's thresholding algorithm to separate the text from the background. Then page layout analysis was performed to identify non-text areas such as images, as well as multi-column text into columns. Then used detections of baselines and words by using blob analysis where each blob was sorted using the x-coordinate (left edge of the blob) as the sort key which enables to track the skew across the page. After the separation of each character, then labeled manually into Sinhala language characters. By using the Sinhala language data file into OCR function, it returns the recognized text, the recognition confidence, and the location of the text in the original image. By considering the recognition confidence of each word it is possible to control the accuracy of the system. The classifier was trained using 40 characters sets with 20 images from each character and tested using over 1000 characters (200 words) with variations of font sizes and achieved approximately 97% of accuracy. The elapsed time was less than 0.05 per a line with more than 20 words, which was a higher improvement than a manual data entering. Since the classifier can be retrained using testing images, it can be developed to achieve active learning.

Keywords: Optical Character Recognition; Computer Vision; Image Processing; Image Segmentation