ISSN: 1391-8796
Proceedings of 10th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
January 18, 2023

Automatic extraction and recognition of Sinhala text from images with complex backgrounds

Dasunika A.K.P.J.[1*] and Wijerathna E.H.M.P.M.[1]

[1]*Department of Information Communication Technology, Faculty of Technology, University of Ruhuna, Kamburupitiya, Matara, Sri Lanka*.

Sinhala is a unique and national language spoken only in Sri Lanka. Sinhala characters are difficult to recognize in images with complex backgrounds, and some visually impaired people cannot read/write properly due to their eye problems. This study is mainly focused on extracting and recognizing Sinhala text from images with complex backgrounds using a Convolutional neural network. Different 33 Sinhala characters (10 images per each character) are used as a training dataset. 400 images of bus destination name boards with Sinhala characters are collected as a testing dataset. The character recognition model is trained using CNN with the collected training dataset. The model is trained 10 times until an accuracy of 81% is achieved. The collected dataset of bus destination name board images is used to extract and recognize Sinhala characters and pre-processing is performed on them to check the availability of text in the images. After the recognition process, the non-text regions are removed. Three types of segmentation such as line, word, and character are performed on preprocessed images to segment each character on the image. The segmented characters are used as input for the recognition model, and it is successfully identified as Sinhala characters. Furthermore, there are numerous types of Sinhala optical character recognition and Sinhala handwritten character recognition research, however, there is no research on recognizing Sinhala text on images with complex backgrounds. Moreover, people with eye problems/issues can easily read Sinhala characters in images with a complex background as an important benefit of this proposed methodology.

Keywords: Convolutional Neural Network, Sinhala text, Text Extraction, Text recognition, Complex background

*Corresponding author: jithmiabewickrama@gmail.com