INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

AKSHARA

Deep Learning based Optical Character Recognition System

for Sinhala Handwritten Text with

Tri-Level Segmentation including Overlapped and Touched Character Segmentation

A Dissertation by

Mr. Nathindu Himansha Dias
2018455 | W1761740

Supervised by

Prof. Nihal Kodikara

May 2023

Submitted in partial fulfillment of the requirements for the

BSc in Computer Science degree at the University of Westminster.

ABSTRACT

Computer vision has been expanded to distinguish handwritten and printed characters to enhance interaction between humans and computers. For Asian languages, however, this is still a subject of much debate. Since Sri Lanka is the only country in Asia where Sinhala is the official language, identifying Sinhala language characters still needs to be answered. The majority of character recognition past research works use pattern matching and image processing approaches. However, these methods could be more capable of adapting to variances. Also, most past research works did not address the gaps in text segmentation in the Sinhala language. Segmentation can consider the key factor in text recognition because a single error in the segmentation process will lower the accuracy of the entire recognition process.

By fulfilling the current gaps in Sinhala text recognition, **Akshara – Sinhala HCR System** is capable of recognizing Sinhala handwritten text data from image input accurately and efficiently. This system uses pixel-based algorithms to segment the text input by line, word, and character(tri-level) wise, including overlapped characters. The author proposed a novel algorithm for the touched character segmentation in this research. After the segmenting paragraph into character-level, utilizing a Convolutional Neural Network (CNN) based architecture, the system recognizes the image character as a digital text.

This system performs text segmentation by archiving an average of 91.8% and character recognition with a 93.37% accuracy rate with 65 supported character classes. Segmenting the touched characters using a novel approach and architecture of the CNN is the novelty of this research which covers a gap that needs to be down-scale of Sinhala handwritten recognition.

KEY WORDS:

Image Processing, Optical Character Recognition, Sinhala Handwritten Character Recognition, Sinhala Handwritten Character Segmentation, Line Segmentation, Word Segmentation, Character Segmentation, Overlapped Character Segmentation, Touched Character Segmentation, Deep Learning, Convolutional Neural Network (CNN).

ACM SUBJECT DESCRIPTIONS:

Human-centered computing → Accessibility → Accessibility technologies

Human computer interaction (HCI) → Interaction techniques → Text input

Artificial intelligence → Computer vision → Computer vision tasks → Activity recognition and understanding