INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

University of Westminster, Coat of Arms

Design and Implementation of a Cloud-based Application for Sinhala and Tamil Manuscript Recognition

A dissertation by

Mr. Nimesh Ekanayake

w1867890 | 20210625

Supervised by

Ms. Niwarthana Kariyabadhuge

Submitted in partial fulfilment of the requirements for the MSc in Advanced Software Engineering degree at the University of Westminster.

July 2023

ABSTRACT

Computer vision has made major advances in recognizing and comprehending handwritten and printed documents, transforming human-computer interactions. However, when it comes to Asian languages, such advances confront significant obstacles. Character identification is particularly difficult in Sinhala and Tamil, two significant Asian languages. While Sinhala is Sri Lanka's official language, Tamil is extensively spoken in other nations, including India, Sri Lanka, and Singapore.

Sinhala and Tamil characters, compared to other regularly spoken languages, have complicated circular geometries with interwoven components. Sinhala contains 60 basic (non-cursive) characters that can be combined to form a larger character set. Because of these distinguishing characteristics, recognizing Sinhala and Tamil handwritten characters is a complex procedure.

Pattern matching and image processing techniques are used in traditional methods of character segmentation and recognition. These techniques, however, fail to adapt to the particular peculiarities of the Sinhala and Tamil scripts, limiting their usefulness.

To overcome these issues, the goal of this research project is to create a Convolutional Neural Network (CNN)-based system for identifying handwritten Sinhala and Tamil letters. Unlike older approaches, CNN extracts features automatically throughout the training phase, allowing it to adapt to the complex nature of these scripts. The major goal of this study is to use deep learning to develop an accurate technique for identifying Sinhala and Tamil handwritten characters.

The researcher expects to make major advances in character segmentation and recognition by exploiting this study initiative. The suggested CNN model would enable accurate recognition of Sinhala and Tamil scripts by precisely identifying individual characters inside handwritten text.

Key Words – OCR, CNN, Image Processing, Deep Learning, Sinhala and Tamil Handwritten Manuscript Recognition