UNCONSTRAINED CURSIVE OFF-LINE SINHALA HANDWRITTEN WORD RECOGNITION
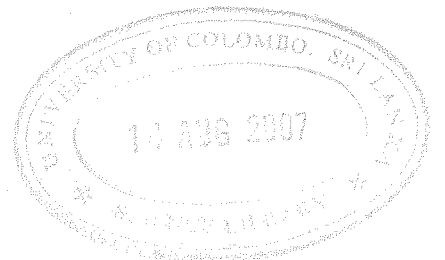
BY

M.L.M KARUNANAYAKA

MPhil in Computer Science

540283

FACULTY OF SCIENCE

UNIVERSITY OF COLOMBO

SRI LANKA

2007

ABSTRACT

Off-line handwriting recognition has been a popular research field for many decades and it is still an open area for research in the character recognition domain. Many researchers in industry and the academic field paid attention to that area with the commercial sector providing a large number of applications related to handwriting recognition systems.

This thesis focuses on unconstrained cursive Sinhala handwritten word recognition with reference to the application of handwritten postal city name recognition. Four major processing stages were used in this recognition system namely preprocessing, segmentation, character recognition and postprocessing. The preprocessing was carried out in six processing phases namely binarization, noise removal, underline removal, skew correction, slant correction and thinning. The accuracy of the segmentation directly affects the quality improvement of the handwritten word. The proposed preprocessing stages improve the quality of the handwritten word significantly.

The segmentation was achieved using the water reservoir concept, vertical projection profile and the labeling algorithm based hierarchical segmentation algorithm. After the segmentation, modifier removal technique was used to improve the recognition rate of characters. Prior to recognition, characters were grouped in to 17 categories using similarity. Eight characters were identified without using recognition approaches. Remaining character groups, having two, three and five characters have to use the recognition process. Two neural networks used in the recognition process were based on the unsupervised algorithm and supervised algorithm. The parallel recognition approach is proposed in this research. After the recognition, the Levenstein distance algorithm was used for the postprocessing. The postprocessing algorithm identifies the most suitable city name and recognizes using the best out of both recognition approaches. The accuracy of the identifying city name using this system is 61.2%. All the components of the system were tested on the real Sinhala postal city name database available in the National Science Foundation (NSF) of Sri Lanka, and some handwriting samples collected from the university students and staff.