525 E1

A statistical classifier for Sinhala handwriting recognition using structural features

In spite of new technological advances, handwriting has continued to persist as a preferred means of communication and information recording. Due to its presence in various human transactions, a computer system that can recognize handwriting has practical significance and has numerous applications in postal address recognition, processing manually filled-out forms, bank cheques processing, etc. Much of the emphasis on handwriting recognition had been concentrated on Latin script, which is used by languages such as English. Very little work is seen in the literature on Sinhala handwriting recognition, and this is the first time that a statistical approach is used.

A subset of characters and modifiers of the Sinhala alphabet, consisting of a total of 34 symbols were chosen for the study. Documents written by different writers were scanned, and binarized to obtain two-tone images. Each image was then automatically separated into text lines, words and finally into individual character and modifiers.

The classification process consists of two stages. First each unknown character image was pre-classified into one of six groups of character classes considering some structural properties in the text lines. A feature vector, consisting of geometric as well as structural features, was extracted from each character image. In the second stage, a statistical classifier based on confidence interval estimation was used for the final recognition. The system produced the best 3 matches for the unknown character image as output.
Thirty sample images from each of the 34 symbol classes were used to train the classifier. The system was tested using a separate set of 500 handwritten characters, written by different writers. Results have shown 74% recognition accuracy for the first choice and up to 94.4% recognition accuracy for the top three choices.