Recognition of Printed Sinhala Characters by Direction Fields

Hemakumar Lalith Premaratne

Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg
Sweden

School of Information Science, Computer and Electrical Engineering
Halmstad University
Box 823
SE-801 18 Halmstad
Sweden

PhD defence will be held in Wigforssalen, Visionen,
Halmstad University, Halmstad
on Friday, May 20, 2005 at 13.15

The thesis is available at the Department of Signals and Systems, Chalmers University of Technology, and at the School of Information Science, Computer and Electrical Engineering, Halmstad University

Abstract

Although substantial research has been carried out on Optical Character Recognition (OCR) where a printed or a handwritten document of script is read as an image and converted to the editable text format, for various languages during the last 30 years, majority of Brahmi descended south Asian scripts are yet to achieve a commercial OCR system.

The Sinhala language and the Sinhala script are used by over 75% of the 19 million population in Sri lanka. The Sinhala alphabet has been evolved from the ancient Brahmi characters over a two millennia. The Sinhala character consists of a large number of different symbols making the segmentation a complex issue. Due to the generally rounded shape of the Sinhala script, Sinhala characters lack many features such as strokes, loops, vertical and horizontal segments, junctions that are widely observed in other member scripts of the 'Brahmi' family. There have been a very few published research on the recognition of the Sinhala script. Hence, an in-depth research on an OCR system is yet to be done.

In this thesis, a novel algorithm for the recognition of the printed Sinhala script has been proposed. The direction tensors that have been used in many applications during the past two decades, are used for the first time for the purpose of character recognition. This algorithm which focuses mainly on a family of similar fonts, overcomes many of the complex issues such as segmentation and feature extraction. In addition, it addresses the issue of confusing characters to a greater extent. The components associated with OCR such as skew correction and text segmentation have also been addressed by the use of direction fields.

Further, a novel statistical approach has been proposed to optimize the recognized script using the Hidden Markov Models (HMMs). In this algorithm words are effectively optimized using the confusing and transition probabilities between different characters.

The recognition algorithm performs at 92% accuracy in character-level. The initial word-level accuracy which stands at 80% is further enhanced to 87% by the proposed optimization algorithm.

Keywords: Optical Character Recognition (OCR), Skew Angle, Direction Tensors, Dynamic Thresholds, Word Optimisation, Hidden Markov Models (HMMs), Viterbi Algorithm.