AN APPROACH TO DEVELOP AN OPTICAL CHARACTER RECOGNIZER FOR PRINTED SINHALA TEXT USING C#

M.G.P.Karunasiri
Postgraduate Institute of Science,
University of Peradeniya,
Srilanka.

This paper describes an effort of developing a character recognizer for Sinhala printed text using C#. Many approaches have been made before to develop a character recognizer for Sinhala printed text by using mathlab as a tool which has all the image processing and classification methods so that it solves many of the implementation issues by the software itself that can arise during the development. When come to practical implementation of an OCR in a commercially developed application, all these implementation issues must be highlighted. But most of the previous works failed to highlight these implementation issues. C#.NET is a commercial framework developed by Microsoft Corporation which provides a bunch of built in methods plus many of the third party classes which are freely and commercially available and it is one of the widely used frameworks in the present IT industry. AForge.NET is a third party framework for C# which provides many of the image processing and AI related functionalities that can be used effortlessly. Digitizing of image involves several steps mainly preprocessing, classification and recognition. Many of the preprocessing methods are tested using AForge built in methods and some of the C# built in methods. Most of the implementation issues that can occur in implementing a digitization algorithm are highlight. Classification is based by taking decimal equivalent numbers for each pixel lines by treating it as binary numbers in row wise and column wise. All the characters in the alphabet are grouped in to four groups in the preliminary classification which yields a recognition rate over 90% of accuracy. And then each group is separately analyzed for further classification to recognize each individual character separately as the secondary classification which yields 70% of accuracy for fonts FS Sumudu, FM Derana and MI Ridhma. Fonts like DL Araliya, DL Kusumi, DL anupama and DL Paras yealds an accuracy less than 70%. A hidden markov model based alphabet training approach is used to give further insight for the secondary classification.