University of Sri Jayewardenepura
Faculty of Graduate Studies

OPTICAL CHARACTER RECOGNITION FOR SINHALA TEXT

A Thesis
By
C. P. Amarajeewa

Submitted in partial fulfillment
of the requirements
for the Degree of
Master of Science
in
Industrial Mathematics.

ABSTRACT

In this study, the Feature Analysis Method is investigated to recognize Sinhala characters.

A scanned image of a character written in black on white paper is converted to a monochrome bitmap file. The bitmap file is then read into a 2-D matrix by using MathCAD software.

The elements of the 2-D matrix were made either 0 (representing black) or 1 (representing white). The matrix was analyzed, using several MathCAD algorithms to find out the features such as;

1. Height /Width ratio
2. Presence of long straight line section
3. Distances from the four corners of the character matrix to the path of the character.
4. Number of pixel curves crossed when a character is sliced along different directions.
5. Area and location of the center of gravity of the convex hull of the character.
6. Location and size of closed regions inside the character.

These features help in the identification of the characters. After testing 75 characters the results were stored in a database.

Results show that by using this database, a scanned image of an unknown character can be identified with a good degree of probability. The identification can be made fool proof by using matrix matching of the scanned image with a template of the most probable character.

CONTENTS

Chapter 3

Features of the Sinhala Characters Used in the Analysis