

1958  
10/02/2010

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.

I grant the University of Sri Jayewardenepura the nonexclusive right to use this work for the University's own purposes and to make single copies of the work available to the public on a not-for-profit basis, if copies are not otherwise available.

Library - USJP



195854

Signature: ... C. P. Amarajeewa

Name: C. P. Amarajeewa

195854

# 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

	Page
• Abstract	i
• Contents	ii
• List of Tables	v
• List of Figures	vi
• Acknowledgment	vii
• Abbreviations.	viii

## Chapter 1

### Introduction

1.1	Optical Character Recognition	1
1.2	Background	1
1.3	The Study	2
1.4	Objectives	3
1.5	Methodology	3
1.6	Scope of the Study	4
1.7	Limitations	5
1.8	Organization of the Thesis	5

## Chapter 2

### Optical Character Recognition

2.1	Historical Perspective	7
2.2	OCR System	8

	Page
2.2.1 Image Scanners	9
2.2.2 OCR Software and Hardware	9
2.2.2.1 Document Analysis	9
2.2.2.2 Character Recognition	10
2.2.3 Output Interface	10
2.3 Element of Successful OCR Application	11
2.4 Reasons for Using OCR Application	11
2.5 Application of OCR to Non-English Languages	11
2.6 Handwritten Address Interpretation	12

## Chapter 3

### Features of the Sinhala Characters Used in the Analysis

3.1 Features that Relate to Outside Limit of the Character	13
3.1.1 Outside Limit of the Character	13
3.1.1.1 Inverse Aspect Ratio of a Character	13
3.1.1.2 Distance from the Four Corners of the “Character Matrix” to the Path of the Character	16
3.2 Longest Straight Line Section of a Character	
3.2.1 Longest Array of Black Pixels	21
3.3 Number of Pixel Curves Crossed when a Character is Sliced Along Different Directions	24
3.3.1 Slicing Code	24
3.3.2 Slant Slicing Code of a Character	26
3.3.3 Rotation of a Character	27
3.4 Feature Relating to Closed Regions in the Character	29
3.4.1 The Number of Enclosed Blank Regions	29
3.4.2 The Location and the Size of the Closed Regions of the Character	30



3.5	Convex hull	31
3.5.1	Convex hull of a Character	31

## **Chapter 4**

### **Results**

4.1	Experimental Results	34
4.2	Formation of Probability Distribution Table	35
4.3	Method of Character Identification	35

## **Chapter 5**

<b>Conclusions and discussion</b>	38
-----------------------------------	----

<b>References</b>	40
-------------------	----

<b>Appendix</b>	41
-----------------	----