Segmentation of Overlapping Sinhala Handwritten Characters

Kalhari Shanthika Ahangama Walawage

169339A

Faculty of Information Technology

University of Moratuwa

February 2019

Segmentation of Overlapping Sinhala Handwritten Characters

Kalhari Shanthika Ahangama Walawage

169339A

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for partial fulfillment of the requirements of Master of Science in Information Technology

February 2019

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student: K.S.A. Walawage

Signature of Student:

Date:

Supervised by

Name of Supervisor: Dr. L. Ranathunga

Signature of Supervisor:

Date:

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr L. Ranathunga, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for his support, patience, motivation, and immense knowledge. His guidance helped me throughout my MSc. research and writing of this thesis.

I'm grateful to my parents, Mr. A.W.B. Chandrasena and Ms. B.D.G. Liyanage for their enormous support given to me at various points in time during this research.

I thanks to my brother and sister for supporting me in any time of my research work.

Finally, I thanks to friends who motivated and supported me.

Abstract

Sinhala is the official and national language of Sri Lanka. Seventeen million people of Sri Lanka use Sinhala language to their day to day works. Most of the researches have been done to Sinhala printed character recognition with high accuracy. Nowadays, Sinhala handwritten character recognition is popular research in Sri Lanka. It is not like printed character segmentation; shape of the same type of handwritten character can be changed in different times. Therefore, characters will be overlapped or touched with each other. Handwritten character segmentation is more important to increase the accuracy of the character recognition. Currently there is lack of high accuracy finding to segment overlapping and touching Sinhala handwritten characters. The proposed methodology has six main sections. They are image acquisition, preprocessing, segmentation, classification, feature extraction and recognition. Collected image was loaded to the system and preprocessed it. Preprocess section is included noise removing, thresholding etc. After that, text line segmentation was done using horizontal projection profile. Mainly this research was introduced a connected pixel labeling method to segmentation of overlapping characters and peak and valley point identification method to segmentation of touching characters. According to tested result, connected pixel labelling method has 97% accuracy and peak and valley identification method has 72% accuracy.

Keywords - Sinhala, Overlapping, Touching, Segmentation, Connected Pixels Labeling, Peak and Valley Point Identification

Table of Contents

Declaration i
Acknowledgements ii
Abstract iii
List of Figures vii
List of Tables ix
Chapter 1 1
Introduction 1
1.1 Chapter Introduction 1
1.2 Background and Motivation 1
1.3 Aim and Objectives 2
1.3.1 Aim 2
1.3.2 Objectives 2
1.4 Solution 2
1.5 Chapter Summary 3
Chapter 2 4
Review of Literature 4
2.1 Chapter Introduction 4
2.2 Segmentation using Projection Profile 4
2.3 Segmentation using Pixel Cluster Detection Technique 5
2.4 Segmentation using Water Reservoir Based Approach 6
2.5 Segmentation using Self-Organizing Feature Maps 7
2.6 Classification using Statistical Classification Approach 7
2.7 Feature Extraction Technique 8
2.8 Character Recognition using Hidden Markov Models 8
2.9 Character Recognition using Artificial Neural Network 9
2.10 Chapter Summary 9
Chapter 3 10
Technology Adapted 10
3.1 Chapter Introduction 10
3.2 OpenCV library 10
3.3 Development Tool 10
3.4 Programming Language 11
3.5 Chapter Summary 11

Chapter 4 12
Analysis and Design 12
4.1 Chapter Introduction 12
4.2 Image Acquisition 12
4.3 Preprocessing 13
4.4 Segmentation 13
4.4.1 Line Segmentation 13
4.4.2 Character Segmentation 14
4.5 Classification 26
4.6 Feature Extraction 27
4.7 Recognition 28
4.8 Chapter Summary 28
Chapter 5 29
Implementation 29
5.1 Chapter Introduction 29
5.2 Preprocessing 29
5.3 Segmentation 29
5.3.1 Line Segmentation 29
5.3.2 Character Segmentation 30
5.4 Classification 37
5.5 Feature Extraction 37
5.6 Recognition 39
5.7 Chapter Summary 40
Chapter 6 41
Evaluation 41
6.1 Chapter Introduction 41
6.2 Result of Preprocessing 41
6.3 Result of Line Segmentation 41
6.4 Result of Overlapping Character Segmentation 41
6.5 Result of Touching Character Segmentation 42
6.6 Evaluation of Overlapping and Touching Character Segmentation 42
6.7 Result of Classification and Feature Extractions 44
6.8 Results of Character Recognition 44
6.9 Chapter Summary 45

Chapter 7 46
Conclusion and Future work 46
7.1 Chapter Introduction 46
7.2 Conclusion 46
7.3 Future Work 46
7.4 Chapter Summary 46
References 47

List of Figures

Figure 1.1: Character groups (a) overlapping, (b) touching 2
Figure 1.2: Straight line in between overlapping characters..... 3
Figure 4.1 Top Level Diagram of the Proposed System..... 12
Figure 4.2 Binarization (thresholding)..... 13
Figure 4.3: The horizontal projection profile of each line..... 13
Figure 4.4: Classification of Character Segmentation..... 14
Figure 4.5: A block that represents the pixels of part of the binary image..... 15
Figure 4.6: A matrix for the pixels of part of the binary image..... 15
Figure 4.7: The matrix that represents Step 1 of First Round..... 15
Figure 4.8: The matrix that represents Step 2 of First Round..... 16
Figure 4.9: The matrix with Some Changes 16
Figure 4.10: The matrix that represents Step 3 of First Round..... 17
Figure 4.11: The matrix with Some Changes 17
Figure 4.12: The matrix that represents Step 4 of First Round..... 17
Figure 4.13: The matrix with Some Changes 18
Figure 4.14: The matrix with Final Result of First Round 18
Figure 4.15: Result of after completed the first round for all the pixels of the image. 18
Figure 4.16: The matrix that represents the replaced number ‘1’ with red..... 19
Figure 4.17: The matrix that represents part of the pixel set of second round step 1 .. 19
completed image 19
Figure 4.18: Result of after completed the step 1 of second round 20
for all the pixels of the image 20
Figure 4.19: Some set of pixels are taken from the completed image of first round... 20
Figure 4.20: Some set of pixels are completed the step 1 of second round..... 20
Figure 4.21: Some set of pixels are completed the step 2 of second round..... 21
Figure 4.22: The output images those are applied Second Round Algorithm 21
until not any change 21
Figure 4.23: Some set of pixels are taken from a non-segmented character, ... 21
after completed second round 21
Figure 4.24: A new Array that represents after completed the third round 22
Figure 4.25: A set of pixels, after completed the fourth round..... 22
Figure 4.26: final output of segmented characters of entire image..... 22

Figure 4.27: A touching character image after overlapping character.....23
segmentation 23
Figure 4.28: Histogram of the first black pixel of each column 24
Figure 4.29: Histogram of the last black pixel of each column.....24
Figure 4.30: Three top and three bottom ranges.....24
Figure 4.31: Touching character segmentation (a) first step, (b) second step 25
Figure 4.32: Points that detected as touching point 25
Figure 4.33: Third step of Touching character segmentation.....25
Figure 4.34: Upper and Lower Baselines of a Segmented Text Line.....26
Figure 4.35: Upper and lower boundary lines of a character.....26
Figure 4.36: Three Classification Groups.....26
Figure 4.37: 8 horizontal and 8 vertical scrips for segmented character 27
Figure 4.38: 65-dimentional Feature Vector.....27
Figure 5.1: Code segment of thresholding.....29
Figure 5.2: Code segment of creating a black pixel line around image.....30
Figure 5.3: Code segment of Horizontal projection profile of binary image.....31
Figure 5.4: Code segment of Segmented line of given line.....31
Figure 5.5:Code segment to create a matrix 32
Figure 5.6: Procedure of First Round 33
Figure 5.7: Procedure of Second Round.....34
Figure 5.8: Procedure of Third Round.....35
Figure 5.9: Procedure of Touching Character Segmentation.....36
Figure 5.10: Algorithm of find the peaks of horizontal pixel densities.....37
Figure 5.11: Algorithm of find the upper baseline and lower baseline 38
Figure 5.12: Preliminary Classification Algorithm 38
Figure 5.13: Algorithm of create the 65-dimentional Feature Vector 39
Figure 5.14: Algorithm to trained data using SVM.....40
Figure 6.1: Image after applying the Preprocessing Techniques.....41
Figure 6.2: Segmented First Line.....41
Figure 6.3: Segmented Fourth character of the line.....42
Figure 6.4: Segmented Touched characters.....42
Figure 6.5: A character with its feature vector 44
Figure 6.6: Uploaded Sinhala Handwritten Character document 44
Figure 6.7: Processed Output.....45

List of Tables

Table 4.1: Classified Characters to Each Group.....27
Table 6.1: Results of Overlapping Character Segmentation.....43
Table 6.2: Results of Touching Character Segmentation 44