UNIVERSITI PUTRA MALAYSIA

A HYBRID APPROACH OF HIDDEN MARKOV MODEL AND FUZZY LOGIC FOR ISOLATED HANDWRITTEN CHARACTERS RECOGNITION

AZIZAH SULIMAN

FSKTM 2011 15

A HYBRID APPROACH OF HIDDEN MARKOV MODEL AND FUZZY LOGIC FOR ISOLATED HANDWRITTEN CHARACTERS RECOGNITION

AZIZAH SULIMAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirement for the Degree of Doctor of Philosophy

August 2011

© COPYRIGHT UPM

Abstract of thesis presented to the Senate Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

A HYBRID APPROACH OF HIDDEN MARKOV MODEL AND FUZZY LOGIC FOR ISOLATED HANDWRITTEN CHARACTERS RECOGNITION

AZIZAH BINTI SULIMAN

August 2011

Chairman : Assoc. Prof. Md. Nasir bin Sulaiman, PhD

Faculty : Computer Science and Information Technology

Research in off-line handwriting recognition has been prevalent for many decades. After many years of intense research, it still remains an open problem. The challenging nature of handwritten characters and words recognition has attracted the attention of researchers from industry and academic circles. The commercial sector has shown significant interest in handwriting recognition research due to the large number of applications that exist.

In recent years, techniques for recognizing handwriting have become more sophisticated in dealing with real-world situation and to increase recognition rates. This thesis reviews all aspects of handwriting recognition research, from the recognition of handwritten numerals to handwritten words. The different methods employed by researchers are mentioned and the approaches adopted for the research are elaborated. The focus of this thesis

© COPYRIGHT UPM

would be the recognition of isolated handwritten characters, concentrating on the slightly more challenging group, lowercase English alphabets. The main aim of this research work is to present a hybrid approach of a syntactical method with a statistical method in a manner that will require less training data but still yield reasonable recognition rate and high reliability rate. The system will be designed with the use of Hidden Markov Model (HMM) as a linguistic variable quantifier for a Fuzzy rule based classifier. This hybrid method, as far as according to the result of the literature search is concerned, is among the first in the area of handwriting recognition.

The main advantage of this approach is a less training intensive method that does not rely on data abundance to achieve a good recognition result. The whole system that integrates the two approaches is tested against a standard database. A favorable outcome of the recognition results, has proven the approach is comparable to many other approaches as in the literature.

© COPYRIGHT UPM

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENDEKATAN HIBRID BAGI MODEL MARKOV TERSEMBUNYI DAN LOGIK KABUR UNTUK PENGECAMAN AKSARA TUNGGAL TULISAN TANGAN

AZIZAH BINTI SULIMAN

Ogos 2011

Pengerusi : Prof. Madya Md. Nasir bin Sulaiman, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Peyelidikan dalam pengecaman tulisan tangan adalah sangat meluas sejak beberapa dekad yang lepas. Walaupun dalam jangkamasa kajian yang agak lama , ia masih lagi kekal sebagai masalah yang masih belum dapat diatasi sepenuhnya. Sifat kajian yang penuh cabaran dalam pengecaman huruf dan perkataan yang bertulis tangan telah menarik perhatian ramai penyelidik daripada golongan industri dan akademik. Sektor komersil telah menunjukkan minat yang besar terhadap penyelidikan penulisan tangan disebabkan wujudnya permintaan yang banyak dalam bidang ini.

Dalam beberapa tahun ini, teknik untuk mengenalpasti gaya penulisan tangan menjadi semakin canggih untuk menangani situasi dalam dunia sebenar dan meningkatkan kadar pengecaman. Tesis ini mengkaji semua aspek dalam pengecaman penulisan tangan daripada pengecaman

© COPYRIGHT UPM

penulisan nombor sehingga kepada penulisan huruf. Pelbagai kaedah berlainan yang diguna-pakai oleh penyelidik telah disebut dan kaedah-kaedahnya telah diperincikan. Fokus tesis ini adalah pengecaman huruf kecil, yang dianggap lebih mencabar, yang tidak bersambung di dalam sistem abjad Inggeris. Sasaran utama dalam penyelidikan ini adalah untuk mempersembahkan penggunaan kaedah campuran dalam kaedah sintak dan kaedah perangkaaan yang memerlukan data yang kurang tetapi mencapai kadar pengecaman yang memuaskan dan kebolehpercayaan yang tinggi . Kaedah campuran yang dicadangkan adalah menggunakan model Hidden Markov (HMM) sebagai kuantifikasi kepada pembolehubah bahasa untuk sistem penggolong peraturan fuzzy. Kaedah campuran ini, menurut penyelidikan bahasa, adalah antara yang pertama dalam bidang pengecaman penulisan tangan.

Kelebihan utama kaedah ini adalah latihan intensif yang kurang dengan tidak bergantung kepada data yang banyak untuk mendapatkan keputusan pengecaman yang tepat. Keseluruhan sistem yang menyepadukan kedua-dua kaedah ini diuji dengan pengkalan data standad. Keputusan yang positif terhadap keputusan pengecaman telah membuktikan kaedah adalah sebanding dengan kaedah-kaedah lain di dalam literatur.

© COPYRIGHT UPM

ACKNOWLEDGEMENTS

Alhamdulillah, my praise to Allah the Almighty.

I would like to thank members of my supervisory committee, Prof. Madya Dr. Md. Nasir Sulaiman, Prof. Dr. Mohamed Othman and Prof. Madya Dr. Rahmita Wirza for their guidance in my work, and especially for their understanding and patience in listening to my excuses every time I missed a dateline. My respect goes to you.

My special gratitude and appreciation to my husband and my six children for their inspirations, encouragements, supports and beliefs that make these years of research work more enduring though sometimes more challenging. My love to you all.

My never ending gratitude to the management of Universiti Tenaga Nasional for financially supporting my work, without which this research would not have been possible. Also for their supports in approving my applications of study leave extensions. I hope to make it worth your while.

Last but certainly not least, my thanks to my family (of mother, brothers, sisters and in-laws) and friends, who are too many to be named, for just being there whenever they are needed. Hope to do the same for you.

Thank you to all of you, ever so much.

© COPYRIGHT UPM

I certify that an Examination Committee has met on 24th August 2011 to conduct the final examination of Azizah binti Suliman on her degree thesis entitled “A Hybrid Approach of Hidden Markov Model and Fuzzy Logic for Isolated Handwritten Characters Recognition” in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows :

Abdul Azim bin Abd. Ghani, PhD
Professor
Department of Information System
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Chairman)

Lili Nurliyana binti Abdullah, PhD
Associate Professor
Department of Multimedia
Department of Information System
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Internal Examiner 1)

Shamala a/p K. Subramaniam, PhD
Associate Professor
Department of Communication Technology and Networking
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Internal Examiner 2)

Anca L. Ralescu, PhD
Professor
Department of Computer Science
University of Cincinati
United States of America
(External Examiner)

……
SEOW HENG FONG, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 24 August 2011

© COPYRIGHT UPM

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows :

Md. Nasir Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Mohamad Othman, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Rahmita Wirza O.K. Rahmat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor
Dean
School of Graduate Studies
Universiti Putra Malaysia

Date : …….

© COPYRIGHT UPM

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or other institutions.

AZIZAH BINTI SULIMAN

Date: 24 August 2011

© COPYRIGHT UPM

TABLE OF CONTENTS

		Page
ABSTRACT		ii
ABSTRAK		iv
ACKNOWLEDGEMENTS		vi
APPROVAL		vii
DECLARATION		ix
LIST OF TABLES		xiii
LIST OF FIGURES		xiv
CHAPTER		
1	**INTRODUCTION**	1
	1.1 Problem Statement	3
	1.2 Research Objectives	4
	1.3 Research Scope	5
	1.4 Research Motivations	6
	1.5 Research Contributions	7
	1.6 Organisation of Thesis	8
2	**LITERATURE REVIEW**	**10**
	2.1 Handwritten Character Recognition System	11
	2.1.1. Off-Line Vs On-Line Handwriting Recognition	11
	2.1.2 Historical Perspective of OCR	12
	2.1.3 Terms Related to OCR	14
	2.2 Methodologies of OCR Systems	19
	2.2.1 Preprocessing	20
	2.2.2 Segmentation	25
	2.2.3 Feature Extraction	34
	2.2.4 Training and Recognition	39
	2.3 Common Approaches in Pattern Recognition	46
	2.3.1 Classifier and Supervised Pattern Recognition	46
	2.3.2 Knowledge-based Pattern Recognition	47
	2.3.3 Hybrid Pattern Recognition Systems	47

© COPYRIGHT UPM

2.4 Related Works 48
2.4.1 HMM in Character Recognition 49
2.4.2 Fuzzy Models for Character Recognition 50
2.5 Summary 52

3 RESEARCH METHODOLOGY 55
3.1 Methodology Employed 59
3.2 Hidden Markov Model 60
3.3 Fuzzy Logic Systems 65
3.3.1 Fuzzy Sets 68
3.3.1 Fuzzy Rules 73
3.4 Motivations of the chosen approach 78
3.5 Handwriting Databases 80
3.5.1 Performance Measurement 82
3.6 Summary 82

4 DESIGN AND PROPOSED TECHNIQUES 84
4.1 The Proposed System Structure 84
4.2 Pre-processing Phase 85
4.2.1 Binarization 87
4.2.2 Reference Line Estimation 89
4.2.3 Thinning 92
4.2.4 Checking for noise and dots 95
4.2.5 Chain Code 96
4.3 Feature Extraction Phase 101
4.3.1 Extracting Features from Chain-codes 102
4.3.2 Preparing Input for Linguistic Variables 108
4.4 Classification Phase 111
4.4.1 Development of the Fuzzy Inference System 113
4.4.2 Membership Function 114
4.4.3 Fuzzy Rule Based 115
4.4.4 Inference Process 117
4.5 Summary 126

© COPYRIGHT UPM

5	RESULTS AND DISCUSSIONS		127
	5.1	Testing Databases	128
	5.2	Results of Feature Extraction Phase	128
	5.3	Results of Classification Phase	133
	5.4	Analysis and Comparison of Results	135
	5.5	Analysis of Feature Extraction Results	139
	5.6	Summary	140
6	CONCLUSION AND FUTURE WORKS		142
	6.1	Conclusion	142
	6.2	Future Works	144
	REFERENCES		**146**
	APPENDICES		**156**
	BIODATA OF STUDENT		164

© COPYRIGHT UPM