



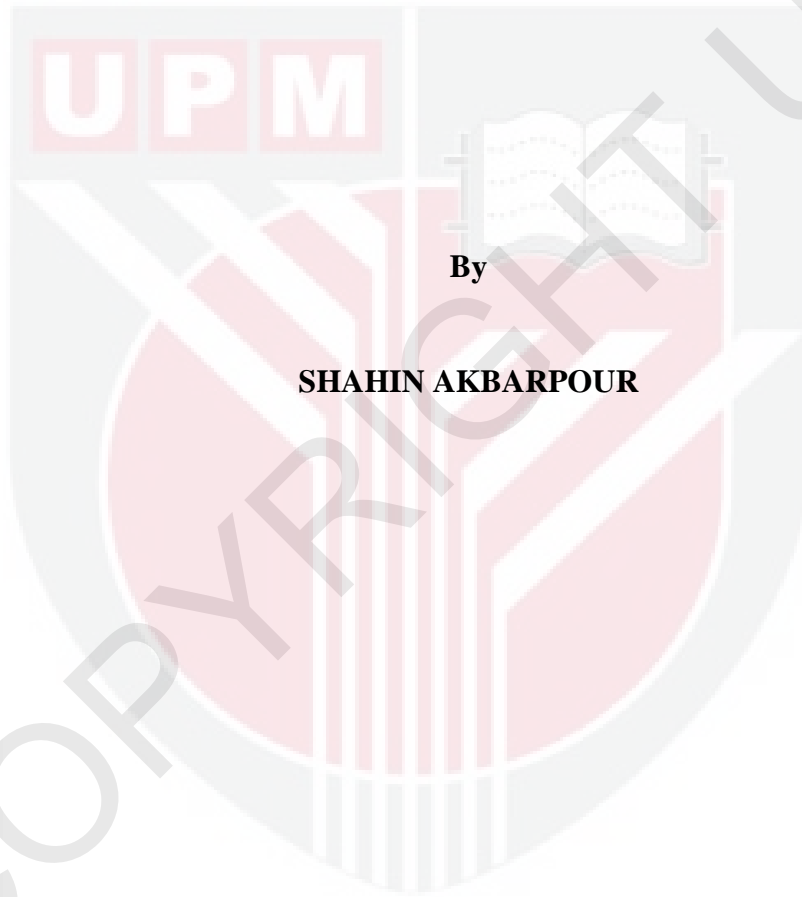
UNIVERSITI PUTRA MALAYSIA

**IMPROVED FEATURE EXTRACTION AND LEXICON REDUCTION
METHODS CLASSIFIED BY SUPPORT VECTOR MACHINE FOR FARSI
HANDWRITTEN WORD RECOGNITION SYSTEM**

SHAHIN AKBARPOUR

FSKTM 2011 21

**IMPROVED FEATURE EXTRACTION AND LEXICON REDUCTION
METHODS CLASSIFIED BY SUPPORT VECTOR MACHINE FOR FARSI
HANDWRITTEN WORD RECOGNITION SYSTEM**



By

SHAHIN AKBARPOUR

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

August 2011

Dedicated to:

Behnaz, my wife

Shayan, my Son

My Parents

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

**IMPROVED FEATURE EXTRACTION AND LEXICON REDUCTION
METHODS CLASSIFIED BY SUPPORT VECTOR MACHINE FOR FARSI
HANDWRITTEN WORD RECOGNITION SYSTEM**

By

SHAHIN AKBARPOUR

August 2011

Chair: Associate Professor Md. Nasir bin Sulaiman

Faculty: Computer Science and Information Technology

Automatic word recognition has proved an intensive research subject for many languages in the last decades, but it is still far from the final frontier for some languages. The word recognition is divided into two types: online and offline. The current research is focused on the offline handwritten word recognition (FHWR). An offline handwritten word recognition system includes many stages. All stages should be improved in order to enhance accuracy of the system. In addition, one of the most significant current discussions in enhancement of the accuracy of handwritten word recognition is reducing the lexicon size.

Many studies have been carried out so far, but FHWR has not been researched as thoroughly as Latin or Chinese handwritten systems. Several attempts have been made to address FHWR, most of which focusing on the image preprocessing and segmentation. It is also worth mentioning that some studies have already been done on

the feature extraction, classification and lexicon reduction methods. In the latest and the most successful prior studies, a feature extraction method, a lexicon reduction, and hidden Markov model (HMM) have been used. However, the recognition rate is not superior owing to the fact that the feature extraction method could not truly describe the Farsi word. Moreover, there exist some limitations in HMM, and several segmentation errors occurred in their lexicon reduction.

The current research is focused on solving the mentioned problems through improving the accuracy of recognition rate of FHWR by proposing a new feature extraction and lexicon reduction methods, and finding a suitable classification. In this regard, some special attributes of Farsi manuscripts such as the stroke directions, non-unique black pixels distribution on binary image of the word, the number of the sub-word(s) and dot(s) of the word will be considered. In addition, several classification methods will be tested in order to determine which one is the best for better accuracy of recognition rate other than HMM. We developed two word recognizer systems to cater for different applications based on different lexicon size. For small lexicons, the word recognizer system consists of a new feature extraction and a classifier, and for medium and large lexicons, the system includes a new feature extraction and lexicon reduction methods and a classifier.

For the performance evaluation of the proposed methods, we use four different Farsi handwritten datasets such as Farshids' Legal amount, 198-Cities, Iranshahr, and IFN-AUT, which contained 45, 198, 503, and 1080 class-words, respectively. In addition, for

comparison of the obtained results with the previous works, we need proper datasets used by prior researchers. AUT and IFN-AUT were applied previously. The AUT, which included 198 class-words, was not available, but a similar dataset, 198-Cities, was created by random selection of 198 class-words from Iranshahr dataset. In order to conduct more experiments based on different lexicon size, the proposed methods were run on Farshids' Legal amount and Iranshahr datasets as well.

Moreover, we re-implemented the existing word recognizer and lexicon reduction method so that we could test for comparison using the same dataset such as 198-Cities and IFN-AUT. It might be concluded that our methods, which consist of a new feature extraction and lexicon reduction methods and the classifier, perform better than the latest works.

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

**PENAMBAHBAIKAN KAEDAH PENGEKSTRAKAN CIRI DAN
PENGURANGAN LEKSIKON DIKLASIFIKASIKAN MENGGUNAKAN
MESIN VEKTOR SOKONGAN UNTUK SISTEM PENGECEMAN TULISAN
TANGAN PERKATAAN FARSI**

Oleh

SHAHIN AKBARPOUR

Ogos 2011

Pengerusi: Profesor Madya Md. Nasir bin Sulaiman

Fakulti: Sains Komputer dan Teknologi Maklumat

Pengecaman perkataan automatik merupakan topik penyelidikan intensif untuk bahasa-bahasa yang berbeza dalam dekad-dekad akhir ini, tetapi ia masih jauh dari sempadan akhir. Pengecaman perkataan dibahagikan kepada dua jenis: dalam talian dan luar talian, tetapi penyelidikan ini lebih tertumpu kepada pengecaman perkataan luar talian sistem bagi pengecaman tulisan tangan mempunyai banyak peringkat. Keseluruhan peringkat harus diperbaiki supaya dapat meningkatkan ketepatan sistem tersebut. Tambahan pula, salah satu daripada perbincangan semasa yang paling penting dalam peningkatan ketepatan pengecaman tulisan tangan adalah dengan mengurangkan saiz leksikon.

Banyak kajian telah dijalankan setakat ini, walaubagaimanapun pengecaman Farsi tulisan tangan luar talian (FHWR) tidak dikaji sebaik tulisan tangan Cina atau Latin. Beberapa percubaan telah dibuat kepada FHWR, kebanyakannya tertumpu kepada pra-

pemproses imej-imej dan segmentasi, walaupun terdapat beberapa kajian tentang penyarian sifat, pengelasan dan leksikon. Terbaru dan paling berjaya dalam kajian sebelum ini adalah menggunakan kaedah penyarian sifat, pengurangan leksikon, dan ‘*Hidden Markov model (HMM)*’. Bagaimanapun, kadar pengecaman tidaklah begitu baik, oleh kerana kaedah penyarian sifat tidak berupaya menghuraikan perkataan Farsi dengan sebaiknya. Tambahan pula, terdapat beberapa batasan dalam HMM. Sementara itu, beberapa kesilapan segmentasi telah berlaku dalam pengurangan leksikon mereka.

Penyelidikan ini tertumpu kepada penyelesaian masalah-masalah yang telah disebutkan tadi, dengan meningkatkan ketepatan kadar pengecaman FHWR. Mencadangkan pengekstrakan ciri baru dan kaedah-kaedah penurunan leksikon dan mencari satu pengelasan sesuai. Dalam perhatian ini, beberapa atribut istimewa manuskrip-manuskrip Farsi seperti ‘*stroke directions*’, ‘*non-unique black pixels distribution*’ pada imej binari, bilangan sub-perkataan dan dot akan dipertimbangkan. Sebagai tambahan, cara-cara pengelasan seperti ‘*back-propagation*’, ‘*k-Nearest-Neighbor classification (KNN)*’, dan ‘*support vector machine (SVM)*’ akan diuji supaya dapat mengenalpasti ketepatan dan pengecaman mana yang lebih baik selain daripada HMM. Kami membangunkan dua Sistem Pengecam Perkataan untuk menyediakan aplikasi yang berlainan dengan saiz leksikon yang berbeza: Untuk leksikon kecil, sistem pengecam perkataan mengandungi pengekstrakan ciri baru dan satu pengelas dan untuk leksikon sederhana dan leksikon besar, sistem itu mengandungi pengekstrakan ciri baru dan kaedah-kaedah penurunan leksikon dan satu pengelas.

Untuk penilaian prestasi bagi kaedah-kaedah yang dicadangkan, kami memerlukan set data tulisan tangan Farsi yang sesuai dan digunapakai oleh penyelidik sebelumnya. AUT AND IFN-AUT telah digunakan sebelum ini, ia mengandungi 198 dan 1080 leksikon. Bagaimanapun, AUT tidak boleh didapati. Satu set data serupa diwujudkan secara rawak, pemilihan 198 kelas perkataan diantara 503 kelas perkataan daripada set data Iranshahr dalam kajian ini. Tambahan pula, kami melaksanakan kembali pengecam perkataan dan kaedah pengurangan leksikon sedia ada supaya kami boleh menguji dan membandingkannya menggunakan set data yang sama. Secara keseluruhannya, keputusan menunjukkan bahawa kaedah-kaedah kami, yang mengandungi pengestrakan ciri baru dan kaedah-kaedah penurunan leksikon dan pengelas SVM adalah lebih baik daripada kerja-kerja sebelum ini.

ACKNOWLEDGEMENTS

My thanks to God for all things throughout my voyage of knowledge exploration.

First, I would like to express my sincere gratitude to my supervisor Associate Professor Dr. Md. Nasir bin Sulaiman for giving me an opportunity to start off this project. Through the course of my study, I have had the great fortune to get to know and interact with him. His comments and suggestions for further development as well as his assistance during writing this thesis are invaluable to me. His talent, diverse background, interest, teaching and research style has provided for me an exceptional opportunity to learn and made me become a better student.

I would like to express my sincere thanks and appreciation to the supervisory committee members Assoc. Prof. Dr. Norwati Mustapha and Assoc. Prof. Dr. Rahmita Wizra O.K. Rahmat for their guidance, valuable suggestions and advice throughout this work in making this a success.

My deepest appreciation to my wife Dr. Behnaz Hassanbaglou and my son Shayan, who have been very supportive and patiently waiting for me to complete my study. Finally, I owe my sincere thanks to my parents for their encouragement and affirmation, which made it possible for me to achieve this work.

For the others who have directly or indirectly helped me in the completion of my work especially Prof. Dr. Kabir in Tarbiat Modares University, MVRL research groups in Amirkabir University, and CENPARMI in Concordia University, I thank you all.

I certify that an Examination Committee has met on 00 / 00 / 2011 to conduct the final examination of Shahin Akbarpour on his Doctor of Philosophy thesis entitled "Improved Feature Extraction and Lexicon Reduction Methods Classified by Support Vector Machine for Farsi Handwritten Word Recognition System" in accordance with Universities and University College Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Rusli Hj Abdullah, PhD

Associate Professor and Deputy Dean
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ramlan b Mahmud, PhD

Associate Professor and Dean
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Abd. Rahman bin Ramli , PhD

Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Internal Examiner)

Maria Petrou, PhD

Professor
Department of Electrical and Electronic Engineering
Imperial College London
United Kingdom
(External Examiner)

NORITAH OMAR, PhD

Associate Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

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 bin Sulaiman, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Norwati Mustapha, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Rahmita Wizra O.K. Rahmat, PhD

Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

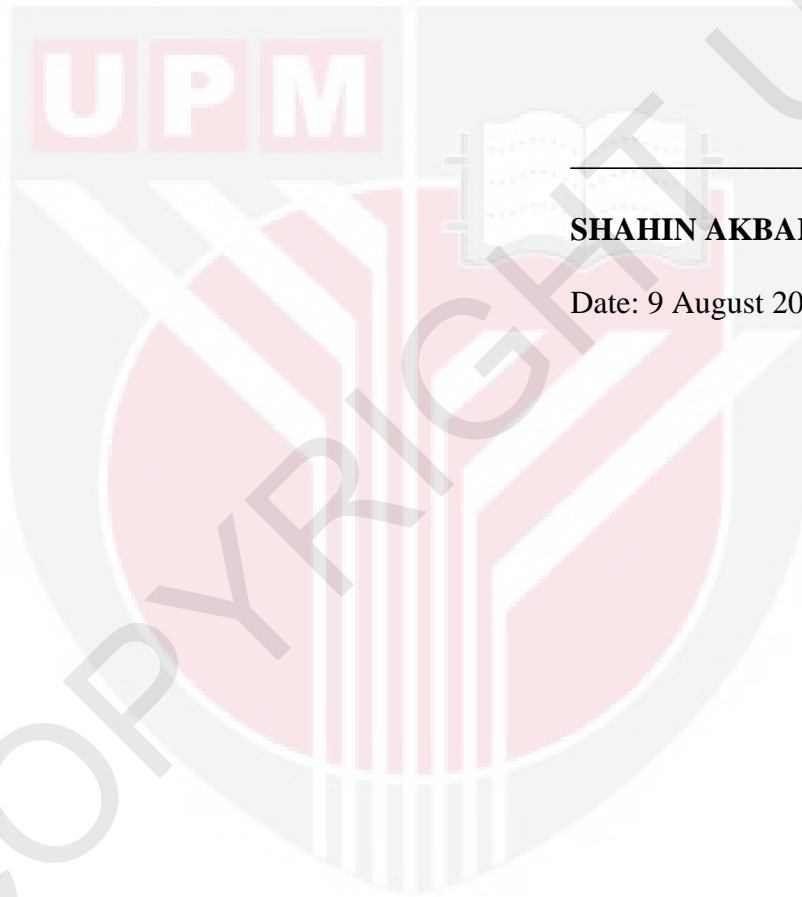
HASANAH MOHD. GHAZALI, PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

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 University Putra Malaysia or other institution.



SHAHIN AKBARPOUR

Date: 9 August 2011

TABLE OF CONTENTS

	Page
ABSTRACT	iii
ABSTRAK	vi
ACKNOWLEDGEMENTS	ix
APPROVAL	x
DECLARATION	xii
LIST OF TABLES	xvii
LIST OF FIGURES	xix
LIST OF ABBREVIATION	xxii
CHAPTER	
1 INTRODUCTION	1
1.1 Overview	1
1.2 Problem Statement	3
1.3 Research Objectives	5
1.4 Research Scope	6
1.5 Research Contribution	6
1.6 Organization of Thesis	7
2 BACKGROUND OF STUDIES	9
2.1 Introduction	9
2.2 Characteristics of Farsi Writing	10
2.3 Handwritten Word Recognition	13
2.4 Handwritten Word Recognition Stages	14
2.5 Pattern Recognition	27
2.5.1 Feature Extraction	27
2.5.2 Data Preprocessing	30
2.5.3 Classification	33
2.6 Different Classification methods	35
2.6.1 Support Vector Machine	35
2.6.2 <i>k</i> -Nearest-Neighbor Classifier	37
2.6.3 Hidden Markov Models	38
2.6.4 Other Classifiers	40
2.7 Lexicon	40
2.7.1 Large Lexicons	42
2.7.2 Lexicon Reduction	43
2.8 Summary	46
3 LITERATURE REVIEW	47
3.1 Introduction	47
3.2 Chronicle OCR	47

3.3	Handwritten Word Recognition Stages	49
3.3.1	Image Preprocessing	49
3.3.2	Image Segmentation	52
3.3.3	Feature Extraction	53
3.3.4	Classification	55
3.4	Lexicon Reduction	61
3.5	Offline Handwritten Standard Databases	62
3.5.1	Farsi Handwritten Words Databases	62
3.5.2	Farsi and Arabic Characters Datasets and Arabic words Datasets	64
3.5.3	Latin Datasets	65
3.6.	Support Vector Machines Existence Packages	65
3.7	Summary	66
4	RESEARCH METHODOLOGY	67
4.1	Introduction	67
4.2	Research Overview	67
4.3	Research Steps	69
4.3.1	Review of the Farsi Handwritten Recognition	69
4.3.2	System Design	69
4.3.3	System Implementation	71
4.3.4	Dataset Preparation	72
4.3.5	Performance Evaluation	73
4.4	Datasets	73
4.5	Experimental Setup	75
4.5.1	System Specification	75
4.5.2	Parameter Setup	75
4.6	Quality Metrics	76
4.7	Experimentation Phase of the Proposed Methods	77
4.7.1	The Proposed Word Recognizer with the New Feature Extraction	78
4.7.2	The Proposed Lexicon Reduction Method	79
4.7.3	The Proposed Word Recognizer with the New Feature Extraction and the Lexicon Reduction	80
4.8	Summary	82
5	METHODS FOR FARSI HANDWRITTEN WORD RECOGNITION	83
5.1	Introduction	83
5.2	The Proposed Feature Extraction Methods	83
5.2.1	Identical-size Image Partitioning	86
5.2.2	Identical-mass Image Partitioning	87
5.2.3	Non-contour-tracing Directional Decomposition	88
5.2.4	Contour-tracing Directional Decomposition	91
5.2.5	Four Novel Feature Extraction Methods for Farsi handwritten Words	92

5.3	The Proposed Word Recognizer with New Feature Extraction	94
5.3.1	Feature Extraction Phase	96
5.3.2	Classification Phase	97
5.4	The Proposed Lexicon Reduction Method	100
5.4.1	Image Segmentation	101
5.4.2	The New Lexicon Reduction Method	103
5.5	The Proposed Word Recognizer with the New Feature Extraction and Lexicon Reduction	103
5.5.1	Lexicon Reduction	105
5.5.2	Classification Phase	105
5.5.3	Improvement of the Word Recognizer	108
5.4	Summary	108
6	RESULTS AND DISCUSSIONS	109
6.1	Introduction	109
6.2	The Proposed Word Recognizer with the New Feature Extraction	109
6.2.1	ARR of the Proposed Word Recognize with the New FE	110
6.2.2	Results of the Word Recognizer on 198-Cities	111
6.2.3	Results of the Word Recognizer on Farshids' Legal amount	114
6.2.4	Results of the Word Recognizer on Iranshahr	117
6.2.5	Discussion of the Proposed Word Recognizer	119
6.2.6	Comparison the Word Recognizer with the Previous FHWR Systems	121
6.3	The Proposed Lexicon Reduction Method	123
6.3.1	Performance Metrics of the New LR Methods	124
6.3.2	The Results from 198-Cities	126
6.3.3	The Results from Iranshahr	126
6.3.4	The Results from IFN-AUT	126
6.3.5	Discussion of the Proposed Lexicon Reduction Methods	127
6.3.6	Comparison the Proposed LR with the Previous Methods	128
6.4	The Proposed Word Recognizer with the New Feature Extraction and Lexicon Reduction	129
6.4.1	ARR of the Word Recognizer with the FE and LR	130
6.4.2	Results of the Word Recognizer on 198-Cities	131
6.4.3	Results of the Word Recognizer on Farshids' Legal amount	132
6.4.4	Results of the Word Recognizer on Iranshahr	134
6.4.5	Discussion of the Word Recognizer	135
6.4.6	Comparison the Word Recognizer with the Previous Methods	137

6.5	Discussion of the Proposed Word Recognizers with and without the New LR	139
6.6	Summary	140
7	CONCLUSION AND FUTURE RESEARCH	141
7.1	Research Conclusion	141
7.2	Future Works	145
	REFERENCES	146
	APPENDIX A	154
	APPENDIX B	159
	APPENDIX C	166
	BIODATA OF STUDENT	168
	LIST OF PUBLICATIONS	169

