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 PENGECAMAN 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 dintara 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 pengekstrakan 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 Mahmod, 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