UNIVERSITI PUTRA MALAYSIA

NEW DISTANCE MEASURES FOR ARABIC HANDWRITTEN TEXT RECOGNITION

MOHAMMAD SAID MANSUR EL-BASHIR

NEW DISTANCE MEASURES FOR ARABIC HANDWRITTEN TEXT RECOGNITION

MOHAMMAD SAID MANSUR EL-BASHIR

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2008

NEW DISTANCE MEASURES FOR ARABIC HANDWRITTEN TEXT RECOGNITION

By

MOHAMMAD SAID MANSUR EL-BASHIR

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

April 2008

بسم الله الرحمن الرحيم

"كما أرسلنا فيكم رسولا منكم يتلو عليكم ءايتنا ويزكيكم ويعلمكم الكتاب والحكمة ويعلمكم مالم تكونو تعلمون"

To my First Teachers: My Father and Mother

To my lovely sisters and brothers

Mohammad

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

NEW DISTANCE MEASURES FOR ARABIC HANDWRITTEN TEXT RECOGNITION

By

MOHAMMAD SAID MANSUR EL-BASHIR

April 2008

Chairman : Rahmita Wirza O.K. Rahmat, PhD

Faculty : Computer Science and Information Technology

In recent years, optical character recognition has attracted scientists and researchers. Latin, Chinese, Korean and Thai characters have been researched more thoroughly than Arabic characters. The research has concentrated firstly on printed and typeset characters until acceptable recognition accuracy has been achieved. Nowadays, most of the researches have gone towards handwritten character recognition.

Arabic text is cursive as characters in a sub-word are connected to each other. This makes the recognition process more complex and a segmentation procedure is required to separate the connected characters from each other before they can be recognized. Features extracted have to be chosen carefully since it has a very important role in the segmentation and recognition process. The recognition accuracy mostly depends on the classifier applied and the segmentation procedure. In this research work, a framework for recognizing the Arabic handwriting is presented. Two approaches have been proposed. The first approach has been designed to recognize the word as a whole to fit applications such as sorting postal mails and bank checks where the number of words or digits that need to be recognized is limited. The words may include country and city

names written on postal mails, or some reserved words or amounts used on bank checks. The second approach represents the general case where any type of documents or handwritten text can be recognized by this approach.

In both approaches, a preprocessing stage including image enhancement and normalization. The most significant features are extracted by implementing the Principal Components Analysis. A new segmentation-based approach is designed and implemented for the second approach to segment the text into characters, while no or simple segmentation procedure is performed in the first approach. The recognition step is performed by applying the nearest neighbor algorithm. Four different distance measures are used with the nearest neighbor, the first norm, second norm (Euclidean), and two new norms proposed called ENorm, EEuclidean. The two new norms proposed (ENorm, EEuclidean) are derived from the first and second norm respectively. The recognition accuracy is enhanced by using the two new norms proposed.

The approaches have been tested as well, and a number of experiments have been discussed more thoroughly. The first approach is experimented by four datasets, which are sub-words containing two characters, sub-words containing three characters, Latin letters and Hindi digits which are used with Arabic language nowadays. The recognition accuracy is the attribute used for measurement, and an 8-fold cross validation technique is used to test this attribute. The average recognition accuracy is 94.8% for the digits, 78% for the three-character sub-words, 77% for the two-character sub-words and 67% for Latin letters. The second approach has achieved recognition accuracy of 73% without detecting dots and 77% with dot detection.

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

PENGUKURAN JARAK BAHARU UNTUK PENGECAMAN TEKS ARAB BERTULISAN TANGAN

Oleh

MOHAMMAD SAID MANSUR EL-BASHIR

April 2008

Pengerusi : Rahmita Wirza O.K. Rahmat, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Dalam beberapa tahun kebelakangan ini, pengecaman aksara optik telah menarik minat para saintis dan penyelidik. Aksara Latin, Cina, Korea dan Thai telah dikaji dengan lebih mendalam berbanding aksara Arab. Penyelidikan lebih menumpukan kepada aksara cetakan dan set taip, sehinggalah penerimaan ketepatan pengecaman telah diperolehi. Kini, kebanyakan penyelidikan telah menjurus ke arah pengecaman aksara bertulisan tangan.

Teks bahasa Arab merupakan aksara kursif di dalam sub-perkataan yang berhubungan antara satu sama lain. Ini menyebabkan proses pengecaman semakin rumit dan prosedur segmentasi diperlukan untuk mengasingkan atau memisahkan karakter-karakter yang berhubungan antara satu sama lain sebelum dicam. Fitur yang diekstrak perlulah dipilih dengan teliti disebabkan peranan yang penting dalam proses segmentasi dan pengecaman. Ketepatan pengecaman bersandar kepada pengkelasan yang diaplikasikan dan prosedur pengecaman. Dalam kajian ini, satu rangka kerja untuk mengecam penulisan tangan bagi aksara Arab dipersembahkan. Dua pendekatan telah dicadangkan. Pendekatan pertama telah direkabentuk untuk kesesuaian aplikasi seperti pengisihan

surat yang dihantar dan juga cek bank, di mana jumlah perkataan atau nombor yang memerlukan pengecaman adalah terhad. Perkataan tersebut merangkumi nama negara dan bandar seperti yang tertulis di alamat surat, atau beberapa perkataan yang dikhaskan ataupun amaun yang digunakan untuk cek bank. Pendekatan kedua pula mewakili kes umum di mana apa jua bentuk dokumen atau teks bertulisan tangan boleh dicam menerusi pendekatan ini.

Dalam kedua-dua pendekatan, langkah pra-pemprosesan merangkumi penambahbaikan imej dan penormalan. Fitur yang paling signifikan diekstrak dengan mengimplementasi Analisis Komponen Utama. Pendekatan baharu berasaskan pensegmenan direkabentuk dan diimplementasi untuk pendekatan kedua bagi membahagikan teks kepada aksara, yang mana tidak ada atau pun hanya prosedur pensegmenan yang mudah sahaja dilakukan dalam pendekatan pertama. Langkah pengecaman dilaksanakan dengan mengaplikasikan algorithma jiran terdekat. Empat pengukuran jarak yang berbeza digunakan bersama jiran terdekat, norm pertama, norm (Euclidean) kedua, dan dua norm baharu yang dicadangkan dengan panggilan ENorm, EEuclidean. Dua norm baharu (ENorm, EEuclidean) yang dicadangkan diterbitkan daripada norm pertama dan kedua masing-masing. Kejituan pengecaman ditambahbaikkan dengan menggunakan dua norm baharu yang dicadangkan.

Pendekatan-pendekatan ini telah diuji, dan beberapa eksperimen telah dibincangkan dengan mendalam. Pendekatan pertama telah diuji dengan tiga set data, iaitu sub-perkataan yang mengandungi dua aksara dan tiga aksara serta digit Hindi yang digunakan dalam bahasa Arab kini. Ketepatan pengecaman merupakan atribut yang digunakan untuk pengukuran, dan teknik pengesahan silang 8-lipatan digunakan untuk

menguji atribut ini. Purata ketepatan pengecaman adalah 94.8% bagi digit, 78% bagi sub-perkataan tiga aksara dan 77% bagi sub-perkataan dua aksara. Pendekatan kedua pula mencapai ketepatan pengecaman dengan 73% tanpa pengesanan titik dan 77% dengan pengesanan titik.

ACKNOWLEDGEMENTS

In the name of *ALLAH*, the most merciful and most compassionate. Praise to *ALLAH S.W.T.* who granted me strength, courage, patience and inspirations to complete this research work.

This work would not have been possible without the nicest guidance from my research supervisor, Associate Professor Dr. Rahmita Wirza O.K. Rahmat. She inspires me about the right way of the research.

I would like to express my gratitude and thanks to the supervisory committee, Associate Professor Dr. Hjh. Fatimah Dato Ahmad for her valuable comments and fruitful discussions and Associate Professor Dr. Hj. Md. Nasir Sulaiman for his guidance and valuable suggestions.

My noblest father Dr. Said El-Bashir and my great mother Basimah are the reasons of my success. They are my first teachers who taught me the mystery of success and the greatness of science. Furthermore, they taught me that humbleness is the reason of getting more and more knowledge. I am indebted to them for all the stages left and remaining of my life.

To my lovely brothers and sisters Essam, Ala, Ahmad, Alaa, Huthaifa, Ayat and Tasnem for their patience and encouragement during my study. My dearest uncles Ibrahim El-Bashir and Ahmad Abdul Hadi deserve much respect for their honest encouragement.

Special appreciation to all my friends that help me to finish my study, mainly, Dr. Raed Khasawneh, Dr. Ayman Omar, Dr. Qasem Al-Radaideh, Dr. Zeyad Al-Zhour and Syaiba Balqish for their joy sharing during the period of my study in Malaysia and their encouragement.

A special thanks also for the higher council for science and technology in Jordan, Professor Mohammad Zaki Khedher and Dr. Gheith Abanadh for their cooperation in providing the database used in this thesis and for professor Khedher valuable comments and suggestions. Thanks also go to Dr. Somaya Alma'adeed for her cooperation in giving her database to have the same platform to compare with the previous research.

I certify that an Examination Committee has met on 17 April 2008 to conduct the final examination of Mohammad Said Mansur El-Bashir on his Doctor of Philosophy thesis entitles "New Distance Measures for Arabic Handwritten Text 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:

Hamidah Ibrahim, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ramlan Mahmod, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Abd. Rahman Ramli, PhD
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Internal Examiner)

Khairuddin Omar, PhD
Associate Professor
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
(External Examiner)

_____

HASANAH MOHD. GHAZALI, PhD
Professor / Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

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

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

Fatimah Dato Ahmad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

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

AINI IDERIS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 14 August 2008

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 at any other institution.

_____

MOHAMMAD SAID MANSUR EL-BASHIR

Date: 8 July 2008

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	After Death
ANN	Artificial Neural Network
AOCR	Arabic Optical Character Recognition
CC	Connected Components
CCLA	Connect Components Labelling Algorithm
COM	Center Of Mass
CPU	Central Processing Unit
CV	Cross Validation
EM	Expectation Maximization
HMMs	Hidden Markovian Models
ICDAR	International Conference on Document Analysis and Recognition
ILP	Inductive Logic Programming
MIT	Massachusetts Institute of Technology
NN	Neural Network
OCR	Optical Character Recognition
PCA	Principal Components Analysis
PCD	Principal Component Discrimination
PHMM	Planar Hidden Markov Model
REAM	Reconnaissance de l'Ecriture Arabe Manuscrite
RGB	Red Green Blue
STDA	Secondary Type Detection Algorithm
UOB	University Of Balamand
WWW	World Wide Web

CHAPTER 1

INTRODUCTION

1.1 Background

Automatic recognition of text has been found since the early days of computer invention. Optical character recognition (OCR) machines have been commercially available since early 1950s (Mori *et al.*, 1992). Initially the recognition process has been performed on isolated characters, but nowadays methods are used to recognize the entire documents. Before, the recognition research has been limited to recognizing machine printed characters, but nowadays the research comprises of handwritten text. Despite the age of the subject, it remains one of the most challenging and exciting areas of research (Srihari and Ball, 2007).

Handwritten character recognition is one of the challenging fields for research. It can be defined as the task of transforming text represented in the spatial form of graphical marks into its symbolic representation. Handwritten character recognition is applied in several types of fields, such as making digital copies of handwritten documents, sorting mail in a post office (Dzuba *et al.*, 1997), check processing and office automation (Dimauro *et al*, 2002).

Handwritten character recognition is categorized based on the method of acquiring data into two types: on-line and off-line (Khorsheed, 2003). In on-line, the symbols are recognized as they are drawn (Klassen and Heywood, 2002). The most common device used for acquiring data is the digital tablet with a stylus pen as data is captured in x and y coordinates as a function of time. In off-line, the recognition is performed after writing or printing is completed, as the recognition of text is in a form of an image. Off-line is considered as the most general case (Khorsheed, 2002). Data is acquired by the computer through an optical device such as a scanner or a camera. This thesis deals with off-line handwritten recognition.

Several language characters have been researched, such as Latin, Chinese, Japanese (Amin, 1997), Korean (Jin-Soo Lee *et al.*, 1999), Tamil (Suresh and Ganesan, 2005) and Thai (Pornchaikajornsak and Thammano, 2003). In this thesis, the concentration is on the recognition of Arabic characters.

This chapter elaborates on the research motivation, the problem statement, the objectives, the scope, the research methodology, contributions of the research and organization of the thesis.

1.2 Research Motivation

Character recognition is one of the important and significant fields of research, especially when considering that its goal is to simulate the human reading capabilities.

This will make it possible to enter text documents or manuscripts to the computer automatically, which can improve the interaction between man and machine in several applications such as sorting mail (Farah *et al.*, 2006), as it is able to read the address written on the envelope and organize it according to its location of destination. Processing checks in banks automatically is also one of the important applications as the number of checks that circulate daily is becoming enormous to process manually (Rafael and Amar, 2006). Several databases are originally available on papers and now converted to an electronic media, such as various government division application forms and transactions, products specifications, several types of manuals, various archives of different knowledge divisions, the existence of the World Wide Web and online services emphasize the necessity of having automatic text documents (Yaseen *et al.*, 2001).

Latin, Chinese, Japanese, Korean and Thai (Lorigo and Govindaraju, 2006) are researched more thoroughly, but recently some researches are conducted on Arabic characters, though not as much as other language scripts because of the cursiveness of Arabic language.

Arabic handwritten character recognition is a challenging task as Arabic is spoken by more than 230 million people (Ethnologue, 2000) as their native language, and used by over one billion as several religion related activities. Some researches have been done for recognizing handwritten Arabic characters, despite that still more researches are needed to achieve its ultimate goal which is the ability to read characters as good as the human being. Also, automatic reading of handwritten text will help in reducing the processing time, and a greater amount of work will be executed in a limited time.