NOVEL TECHNIQUES FOR ENHANCING AUTOMATIC ARABIC HANDWRITING RECOGNITION

By

Hany Ahmed Sayed Mansour

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATIONS ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

NOVEL TECHNIQUES FOR ENHANCING AUTOMATIC ARABIC HANDWRITING RECOGNITION

By
Hany Ahmed Sayed Mansour

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATIONS ENGINEERING

Under the Supervision of

Prof. Mohsen A. Rashwan

Professor of Electronics and Communications Engineering,
Faculty of Engineering, Cairo University

Prof. Sherif Abdel Azeem Mohamed

Professor of Electronics and Communications Engineering,
American University in Cairo

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

NOVEL TECHNIQUES FOR ENHANCING AUTOMATIC ARABIC HANDWRITING RECOGNITION

By
Hany Ahmed Sayed Mansour

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATIONS ENGINEERING

Approved by the
Examining Committee

_____
Prof. Mohsen Abdul Raziq Rashwan
Electronics and Communications Engineering Department
Faculty of Engineering, Cairo University

_____
Prof. Sherif Abdel Azeem Mohamed
Electronics and Communications Engineering Department
American University in Cairo

_____
Prof. Sherif Mahdi Abdou
Faculty of Computers and Information, Cairo University

_____
Prof. Mohamed Waleed Talaat Fakhr
Faculty of Computer and Information Technology,
Arab academy for science and technology

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

Engineer's Name: Hany Ahmed Sayed Mansour
Date of Birth: 25/08/1988
Nationality: Egyptian
E-mail: hanyahmed@rdi-eg.com
Phone: 01144308495
Address: El-Sheikh Zayed

Registration Date: 1/10/2013.
Awarding Date: …./…./201
Degree: Master of Science
Department: Electronics and Communications Engineering

Supervisors:

Prof. Mohsen A. Rashwan
Prof. Sherif Abdel Azeem

Examiners:

Porf. Mohsen Abdul Raziq Rashwan
Porf. Sherif Abdel Azeem American University in Cairo
Prof. Sherif Mahdy Abdou
Prof. Mohamed Waleed Talaat Fakhr Arab academy for science and technology

Title of Thesis:

Novel techniques for enhancing automatic Arabic handwriting recognition

Key Words:
Arabic, Online handwriting, Offline handwriting, fusion

Summary:

In this thesis, we present a novel segmentation free Arabic handwriting recognition systems based on hidden Markov model (HMM). Three main contributions are introduced: online Arabic handwriting recognition system, offline Arabic handwriting recognition system and combining the both offline and online systems. Experimental results and Comparisons with state of the art techniques shows that our proposed techniques are robust, effective and competitive.

Acknowledgments

All the praises and thanks be to Allah, the Lord of the Worlds

I am using this opportunity to express my gratitude to everyone who supported me throughout this research work. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advices.

I would like to express my deepest appreciation to my supervisor Prof. Dr. Mohsen A. Rashwan. I would like to thank him for his great encouragement.
A very special thanks to prof. Dr. Sherif Abdel Azeem for his help and support. I would like to thank him for the great expertise which I gained in my life.
Many thanks for my colleagues for their help and support.
Grateful to my family for prayers for me, **especially my mother**.

Finally, there are no words can thank my wife.

Dedication

This dissertation and all of my academic achievements are dedicated to my beloved wife and my son.

Table of Contents

ACKNOWLEDGMENTS ..... **I**

DEDICATION ..... **II**

TABLE OF CONTENTS ..... **III**

LIST OF TABLES ..... **V**

LIST OF FIGURES ..... **VI**

NOMENCLATURE ..... **VIII**

ABSTRACT ..... **VIIIX**

CHAPTER 1 : INTRODUCTION **1**

CHAPTER 2 : BACKGROUND AND RELATED WORK **3**

2.1. ARABIC HANDWRITING PROPERTIES AND DIFFERENT RECOGNITION PROBLEMS . 4
2.2. PRE-PROCESSING TECHNIQUES 4
2.3. FEATURE EXTRACTION TECHNIQUES 5

CHAPTER 3 : HIDDEN MARKOV MODELS **7**

3.1. ISOLATED WORD RECOGNITION 7
3.2. OUTPUT PROBABILITY SPECIFICATION 9
3.3. PARAMETERS ESTIMATION 9
3.4. CONTINUOUS RECOGNITION 10

CHAPTER 4 :PROPOSED ONLINE ARABIC HANDWRITING RECOGNITION SYSTEM **11**

4.1. INTRODUCTION 11
4.2. SYSTEM DESCRIPTION 11
4.2.1. Pre-processing 12
4.2.1.1. Interpolation 13
4.2.1.2. Smoothing 14
4.2.1.3. Base-line extraction 14
4.2.1.4. Removing delayed strokes 14
4.2.2. Features extraction 15
4.2.2.1. Directional features 16
4.2.2.2. Aspect features 16
4.2.3. HMM models description 16
4.2.4. Word modeling and dictionary building 18
4.2.5. Classification and associating the delayed strokes 21

CHAPTER 5 : PROPOSED OFFLINE ARABIC HANDWRITING RECOGNITION SYSTEM **23**

5.1. INTRODUCTION 23
5.2. SYSTEM DESCRIPTION 23

5.2.1 PRE-PROCESSING 24
5.2.1.1 CHANGING THE THICKNESS OF THE WORD TO A PRE-DETERMINED NUMBER OF PIXELS 24
5.2.1.2 REDUCING THE DISTANCE BETWEEN PARTS OF WORDS (POWS) 25
5.2.2 HMM MODELS DESCRIPTION 26
5.2.3 WORD MODELING AND DICTIONARY BUILDING 27
5.2.4 FEATURES EXTRACTION 28
5.2.4.1 VERTICAL OVERLAPPING FRAMES 28
5.2.4.1.1 CONCAVITY FEATURES 28
5.2.4.1.2 GRADIENT FEATURES 31
5.2.4.2 SLANTED OVERLAPPING FRAMES 33
5.2.5 SYSTEM FUSION 36

CHAPTER 6 : COMBINING ONLINE AND OFFLINE ARABIC HANDWRITING RECOGNITION SYSTEMS 39

6.1. SYSTEM OVERVIEW 39
6.2. PRE-PROCESSING 39
6.3. DICTIONARY BUILDING 39
6.3.1. Online Dictionary building 39
6.3.2. Offline Dictionary building 40
6.4. FEATURE EXTRACTION 40
6.4.1. Online features extraction 40
6.4.2. Offline features extraction 40
6.4.2.1. F1- Number of foreground pixels: 40
6.4.2.2. F2- Gradient features: 41
6.5. FUSION 41

CHAPTER 7 : EXPERIMENTS AND RESULTS 43

7.1. PROPOSED ONLINE SYSTEM RESULTS 43
7.1.1. performance of the proposed recognition system 43
7.1.2. Effect of adding Aspect feature 43
7.1.3. Effect of HMM parameters 44
7.1.4. Main reasons of miss classification 44
7.2. PROPOSED OFFLINE SYSTEM RESULTS 46
7.2.1 PERFORMANCE OF THE PROPOSED RECOGNITION SYSTEM 46
7.2.2 EFFECT OF PRE-PROCESSING STAGE 48
7.2.3 EFFECT OF SEGMENTING THE IMAGE HORIZONTALLY WITH NON-UNIFORM SEGMENTS 49
7.2.4 EFFECT OF CONCAVITY FEATURES 50
7.2.5 EFFECT OF FUSION STAGE 50
7.2.6 EFFECT OF HMM PARAMETERS 51
7.3. RESULTS OF PROPOSED COMBINED SYSTEM 52

CONCLUSION AND FUTURE WORK 53

REFERENCES 54

List of Tables

Table 1.1: The Arabic Alphabet
Table 4.1: Features of ADAB database
Table 4.2: Grouping Arabic characters together in classes
Table 5.1: Data statistics for the IFN/ENIT database
Table 5.2: 10-candidate cities provided by each HMM classifier
Table 5.3: Majority and sum rules conducted on top candidates
Table 5.4: Final candidates
Table 6.1: Mapping the online candidates to their original names
Table 6.2: 10-Candidates words provided by each HMM
Table 6.3: Majority and sum rules conducted on given candidates
Table 7.1: Recognition rate of the proposed system
Table 7.2: Effect of using the Aspect features
Table 7.3: Influence of varying number of states per character on the whole system
Table 7.4: Influence of varying number of mixtures per character on the whole system
Table 7.5: Recognition results in percentage of correctly recognized images on test data sets f and s
Table 7.6: Comparison with other word recognition systems
Table 7.8: Effect of pre-processing stage
Table 7.9: Effect of normalizing the thickness of words
Table 7.10: Effect of varying the number of dilating pixels
Table 7.11: Effect of reducing the empty spaces between different POWs
Table 7.12: Comparison between two systems using uniform and non-uniform heights
Table 7.13: Effect of varying number of horizontal segments
Table 7.14: Comparison of different feature extraction techniques
Table 7.15: Performance of the proposed system using orientation angles
Table 7.16: Influence of varying orientation angle on the whole system
Table 7.17: Oriented words and results provided by individual classifiers
Table 7.18: Influence of varying number of states per character on the whole system
Table 7.19: Influence of varying number of mixtures per character on the whole system
Table 7.20: Recognition results of correctly recognized images
Table 7.21: Recognition rate of the proposed online and offline systems

List of Figures

Figure 3.1: The Markov Generation Model
Figure 3.2: Flow chart of training HMM Models
Figure 4.1: Architecture of online handwriting system
Figure 4.2: Example of solving the problem of inaccuracy of the digitalization process and removing the delayed strokes: (a) Original word, (b) Word after interpolating the missed points,(c) Word after removing the delayed strokes
Figure 4.3: Extracting the baseline using horizontal projection
Figure 4.4: The delayed stroke of the ADAB database
Figure 4.5: Example of rules used to extract the delayed strokes
Figure 4.6: Aspect feature for M=8
Figure 4.7: Grouping the characters in classes after removing the delayed strokes
Figure 4.8: Left-to-right states
Figure 4.9: Flow chart of the algorithm used to construct the dictionary of the unique set of words in the ADAB database with their delayed strokes
Figure 4.10: Dictionary of words with the delayed strokes (dict1).
Figure 4.11: Flow chart of the algorithm used to construct the dictionary of unique set of words in the ADAB database without their delayed strokes.
Figure 4.12: Dictionary of words without the delayed strokes (dict2).
Figure 4.13: Dictionary of words which confuse with other words after removing their delayed strokes (dict3).
Figure 4.14: Block diagram shows the two stages of the classification.
Figure 5.1: Decoding stage architecture.
Figure 5.2: The steps used in changing the thickness of the input word: a, b same original words whose thickness is different in the two figures, c after thinning, d after fixing the thickness.
Figure 5.3: (a) Original word, (b) the word after reducing the empty spaces.
Figure 5.4: Eight-state HMM, left to right.
Figure 5.5: Figure shows diacritic marks and delayed strokes.
Figure 5.6: Eight configurations used to extract the features.
Figure 5.7: Decomposing the input image into eight images.
Figure 5.8: Dividing the image into (a) three horizontal non-uniform segments, (b) four horizontal non-uniform segments, (c) five horizontal non-uniform segments.
Figure 5.9: Feature extraction using a sliding window with overlapping.
Figure 5.10: Eight non overlapping regions of 45 degrees.
Figure 5.11: Twenty-four features per frame by dividing the image into three non-uniform heights.
Figure 5.12: (a) Straight writing of a city name, (b) skewed city names, (c) city names with shifted and slanted delayed strokes, (d) city names with slanted characters.
Figure 5.13: Slanted characters in the written word.
Figure 5.14: Problems appeared after Skew correction.
Figure 5.15: Slanted bounding boxes and slanted sliding windows (frames), (a) positive angle orientation, (b) positive angle slanted cell, (c) negative angle orientation, (d) negative angle slanted cell.
Figure 6.1: Flow chart shows the method of combining online and offline systems.
Figure 6.2: (a) Converting the online trace to bitmap image, (b) the word after dilating the pixels.

Figure 7.1: Problem of moving the hand up while writing one character: (a) Original word, (b) word after removing the delayed strokes and the sub stroke of the isolated Ein.
Figure 7.2: (a) Large spaces between different POWs, (b) thick characters.
Figure 7.3: (a) Words that could not be recognized using fixed heights, (b) same words recognized using non-uniform heights.

Nomenclature

DNN	Deep Neural Network
DS	Delayed Strokes
HMM	Hidden Markov Model
MSA	Modern Standard Arabic
NN	Neural Network
PDA	Personal Digital Assistant
POW	Parts Of Word
SVM	Support Vector Machine

Abstract

In this thesis, we present a novel segmentation free Arabic handwriting recognition systems based on hidden Markov model (HMM). Three main contributions are introduced: online Arabic handwriting recognition system, offline Arabic handwriting recognition system and combining the both offline and online systems.

For offline handwriting system, we introduce a new technique for dividing the image into non-uniform horizontal segments to extract the features and a new technique for solving the problems of the skewing of characters by fusing multiple HMMs. The proposed system first pre-processes the input image by setting the thickness of the input word to three pixels and fixing the spacing between the different parts of the word. The input image is divided into constant number of non-uniform horizontal segments depending on the distribution of the foreground pixels. A set of robust features representing the gradient of the foreground pixels is extracted using sliding windows. The input image is decomposed into several images representing the vertical, horizontal, left diagonal and right diagonal edges in the image. A set of robust features representing the densities of the foreground pixels in the various edge images is extracted using sliding windows. The proposed system builds character HMM models and learns word HMM models using embedded training. Besides the vertical sliding window, two slanted sliding windows are used to extract the features. Three different HMMs are used: one for the vertical sliding window and two for the slanted windows. A fusion scheme is used to combine the three HMMs. The proposed system is very promising and competes with the other Arabic handwriting recognition systems reported in the literature.

For online handwriting recognition system*,* delayed strokes are removed from the online Arabic word to avoid the difficulty and the confusion caused by the delayed strokes in the recognition process. Dictionaries for all the words in the database have been constructed with and without the delayed strokes. Word matching in both dictionaries along with effective online features and careful choice of the HMM parameters have significantly improved the recognition rate of the proposed system.

For the combined system, the integration between online and offline approaches has proven to give a better performance. With the combination we could increase the system performance over the best individual recognizer.

Chapter 1 : Introduction

Arabic is a very rich language which can be recognized as one of the most important and difficult languages in the world. Arabic has influenced many different languages around the world and civilizations throughout its history. Some of these influenced languages are Urdu, Somali, Hindi and Bosnian. During the Middle Ages, Arabic was a vehicle of culture in the world, especially Europe. As a result, many European languages have also borrowed many words from it.

The modern written language (Modern Standard Arabic - MSA) has been derived from the language of the Quran (Classical Arabic) which widely taught in universities and schools.

Arabic has some features which is differ from any another language such as: cursive nature which means that the letters are joined together along a line, dots and other small marks that can change the meaning of a word, written right to left and the shapes of the letters differ depending on position in the word (start, middle, end, isolated), as shown in Table 1.1.

Table 1.1 The Arabic Alphabet

Character	Isolated	End	Middle	Start
	ا	ـا		
BAA	ب	ـب	ـبـ	بـ
TAA	ت	ـت	ـتـ	تـ
THAA	ث	ـث	ـثـ	ثـ
JEEM	ج	ـج	ـجـ	جـ
HAA	ح	ـح	ـحـ	حـ
KHAA	خ	ـخ	ـخـ	خـ
DAL	د	ـد		
THAL	ذ	ـذ		
RAA	ر	ـر		
ZAIN	ز	ـز		
SEEN	س	ـس	ـسـ	سـ
SHEEN	ش	ـش	ـشـ	شـ
S'AD	ص	ـص	ـصـ	صـ
D'AD	ض	ـض	ـضـ	ضـ
TTAA	ط	ـط	ـطـ	طـ
TTHAA	ظ	ـظ	ـظـ	ظـ
AIN	ع	ـع	ـعـ	عـ
GHAIN	غ	ـغ	ـغـ	غـ
FAA	ف	ـف	ـفـ	فـ
QAAF	ق	ـق	ـقـ	قـ
KAAF	ك	ـك	ـكـ	كـ
LAM	ل	ـل	ـلـ	لـ
MEEM	م	ـم	ـمـ	مـ
NOON	ن	ـن	ـنـ	نـ
HHAA	ه	ـه	ـهـ	هـ
WAW	و	ـو	ـو	و
YAA	ي	ـي	ـيـ	يـ

Handwriting recognition can be divided into two categories: online handwriting recognition and offline handwriting recognition.

For offline recognition, the writing is usually captured optically by a scanner or a digital camera. Offline handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within text-processing applications and computer.

For online recognition, a digitizer samples the handwriting to time-sequenced points representing the pin-tip position as it is being written on a digital instrument. Hence, the online handwriting signal contains additional time information which is not presented in the offline signal.

The elements of an online handwriting recognition interface typically include:

- A stylus or pen for the user to write with.
- A touch sensitive surface, which may be integrated with an output display.
- Software applications which detect the movements of the pen across the writing surface and translating the resulting strokes into digital text.

Due to the advantages of hidden Markov models (HMM), many researchers have used them for Arabic handwriting recognition [1–7]. Since these are stochastic models, they can cope with noise and variations in the handwriting, and also the observation sequence that corresponds to features of an input word can be of variable length, and most importantly, word HMMs can solve the problem of segmentation implicitly.

This thesis describes some algorithms used for Arabic online and offline handwriting recognition. Different pre-processing techniques have been introduced to solve well-known problems in Arabic. Different features extraction techniques have been introduced to cope with HMM classifier for both online and offline handwriting systems. Fusion between online and offline systems for online handwriting recognition has been introduced in this thesis to prove that different information are very useful to enhance any online system.

Outline of the thesis

The thesis is structured in six chapters, as follows:

Chapter 1: This Introduction chapter which provided a brief description of the whole work.
Chapter 2: Related works to online and offline handwriting recognition are proposed.
Chapter 3: Summary of Hidden Markov Model theory and its applications.
Chapter 4: Online Arabic handwriting recognition system is proposed.
Chapter 5: Offline Arabic handwriting recognition system is proposed.
Chapter 6: Combination of offline and online Arabic handwriting recognition systems is proposed.
Chapter 7: Experiments and results are described in details in this chapter.