

WRITER IDENTIFICATION OF
ARABIC HANDWRITTEN
DOCUMENTS

BY

SAMEH MOHAMMAD AWAIDA

A Dissertation Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In

COMPUTER SCIENCE AND ENGINEERING


KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This dissertation, written by *SAMEH MOHAMMAD AWAIDA* under the direction of his dissertation advisor and approved by his dissertation committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** in *COMPUTER SCIENCE AND ENGINEERING*.

Dissertation Committee:



Prof. SABRIA A. MAHMOUD (Chairman)



Prof. RADWAN ABDEL-AL(Member)



Prof. MOUSTAFA ELSHAFEI (Member)




Dr. MUHAMMAD AL-SUWAIYEL (Member)



Dr. WASFI AL-KHATIB (Member)



Dr. UMAR AL-TURKI
(Dean, College of Computer Sciences and Engineering)



Dr. SALAM ZUMMO
(Dean of Graduate Studies)

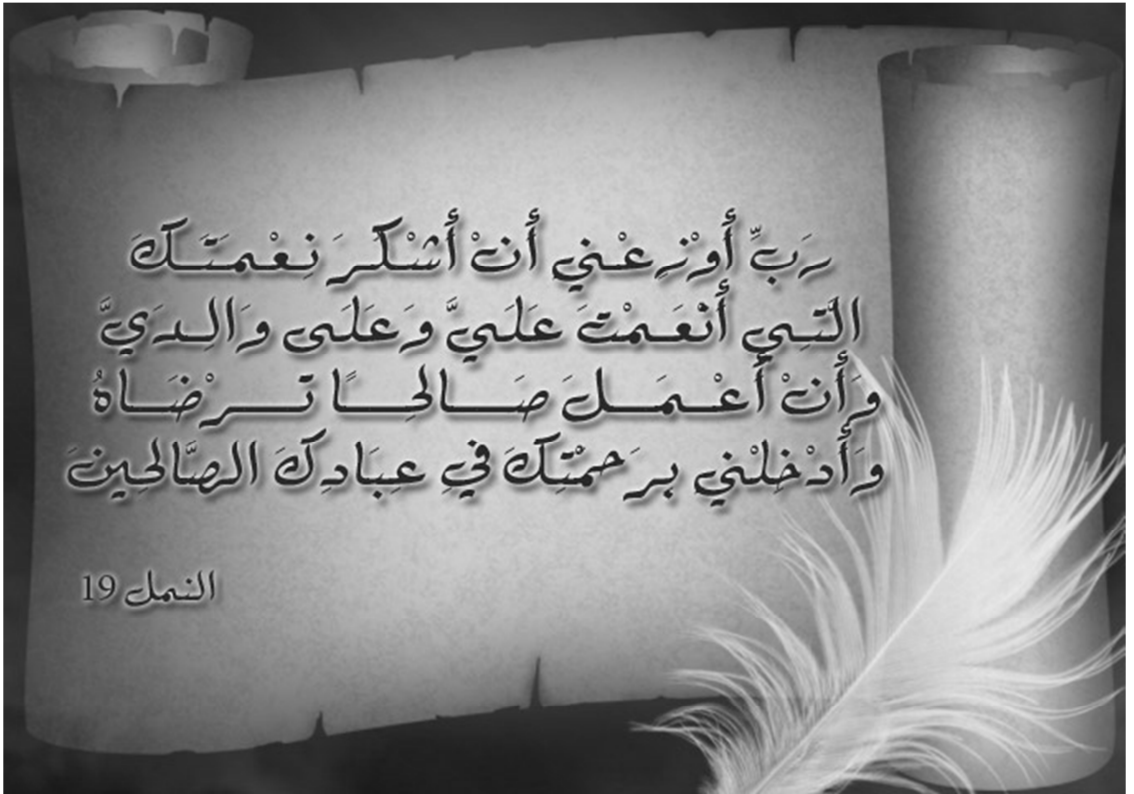
Date: 14/6/11



Dedicated to my Mother and Father.

Thank you both for instilling in me the fervor to continue my education and to live life fully while pursuing dreams, albeit I understand any amount of gratitude shown to you is woefully inadequate.

Words are insufficient to your contribution to my life.



ACKNOWLEDGEMENT

In the name of Allah, Most Kind, Most Merciful. All Praise be to Allah, the Lord of the worlds, who says in His Glorious Book, "My Lord! Increase me in knowledge". All praise is due to Allah for His limitless blessings. Peace and blessings of Allah be upon the noblest of the Prophets, our Prophet Mohammed who said "he who treads the path in search of knowledge, Allah would make that path easy, leading to Paradise for him". May peace be upon him, his family, his companions, and his followers.

I have been blessed to receive tremendous support from several members in my family. I am particularly grateful to my wife Bayan and my two amazing children, Abd Al-Rahman and Mohammad, who have kept exemplary patience while I completed my thesis. I am indeed blessed to have them in my life. I am forever grateful to the support I have received from my older brother Sami during my whole life. Your topmost patience and encouragement have fueled my passion to knowledge throughout these years.

I am indebted to my advisor, Prof. Sabri Mahmoud. Prof. Sabri has influenced not only my graduate studies, but my whole life. Prof. Sabri is a deeply committed researcher, teacher, advisor, and father-figure for all his students. The inspiration that he embedded in me, and the tirelessly working example that he has set, will always be remembered.

My sincere thanks and appreciation are to my thesis committee members Prof. Mustafa Elshafei, Prof. Radwan Abd Al-Aal, Dr. Mohamad Siuwayel and Dr. Wasfi Al-Khatib for their constructive comments and repeated encouragement on this work.

The concept of this dissertation was first suggested to me by Sh. Mashhoor Hasan Salaman, who I owe a debt of gratitude to. His never-ending support to my research and providing me with office space, time and effort is greatly appreciated. I would also like to thank two fellow graduate students who have supported and assisted me throughout my graduate years in KFUPM. They are Dr. Mohammad Tanvir Parvez and Mr. Yousef Elarian. Many ideas in this dissertation are the fruit of the invaluable discussions with them. Their collaboration in developing the new handwritten Arabic database was very important to the completion of this thesis.

I am grateful to Dr. Husni Al-Muhtaseb, who had his door open for me any time I needed help and advice. This work was completed with his great support, useful ideas, and long hours of discussions.

Thanks are also due to the many colleagues who have contributed ideas, support, encouragement and motivation. Dr. Ashraf Mahmoud, Dr. Adnan Gutub, Dr. Mohammad Suwaiyel, Dr. Ghaith Hammouri, Dr. Mohammad Alshayeb, Irfan Ahmad, Dr. Omar Al-Turki, Dr. Salam Zummo, Dr. Thomas Ploetz, Prof. Sargur Srihari, Prof. Gernot Fink, Ahmed Al-Massri, Dr. Andreas Schlapbach, Mahmoud Ankeer, and all other members and staff of King Fahd University of Petroleum and Minerals for all their help and support. Finally, I acknowledge that this research would not have been possible without the support provided by King Abdul-Aziz City for Science and technology (KACST) through the Science & Technology Unit at King Fahd University of Petroleum and Minerals (KFUPM) for funding this work through project number (A-T-18-95).

TABLE OF CONTENTS

ACKNOWLEDGEMENT	IV
TABLE OF CONTENTS	VI
LIST OF TABLES	IX
LIST OF FIGURES.....	X
PUBLICATIONS	XII
ABSTRACT	XIV
خلاصة	XV
CHAPTER 1 INTRODUCTION	1
1.1 WRITER IDENTIFICATION VS. TEXT RECOGNITION	1
1.2 TEXT DEPENDENT AND TEXT INDEPENDENT WRITER IDENTIFICATION	2
1.3 ONLINE AND OFFLINE WRITER IDENTIFICATION	2
1.4 BACKGROUND ON THE ARABIC LANGUAGE.....	3
1.5 PROBLEM STATEMENT.....	6
1.6 SIGNIFICANCE OF THE STUDY.....	6
1.7 CONTRIBUTIONS OF THE THESIS.....	9
1.8 THESIS OUTLINE	12
CHAPTER 2 DATABASE SURVEY AND DESIGN.....	13
2.1 INTRODUCTION	13
2.2 DATABASES FOR WRITER IDENTIFICATION AND VERIFICATION OF WESTERN SCRIPT.....	15
2.3 DATABASES USED IN WRITER IDENTIFICATION OF ARABIC TEXT	22
2.4 DATABASE DESIGN FOR NEW ARABIC HANDWRITTEN TEXT.....	27
2.4.1 <i>Motivation</i>	27
2.4.2 <i>Minimum Paragraph Form</i>	29
2.4.3 <i>Full Page Form</i>	29
2.4.4 <i>Ligatures Form</i>	35
2.4.5 <i>Common Words/Sentences Form</i>	36
2.4.6 <i>Summary of the Designed Form</i>	36
2.5 CONCLUSIONS.....	39
CHAPTER 3 LITERATURE REVIEW	40
3.1 INTRODUCTION	40
3.2 APPLICATIONS OF WRITER IDENTIFICATION AND VERIFICATION	42
3.3 FEATURE EXTRACTION APPROACHES.....	43
3.4 CLASSIFICATION APPROACHES	58
3.4.1 <i>Minimum Distance Classifiers</i>	58

3.4.2	<i>Statistical Classifiers</i>	61
3.4.3	<i>Other Classifiers</i>	62
3.5	WRITER IDENTIFICATION AND VERIFICATION OF ARABIC TEXT	62
3.6	CONCLUSIONS.....	67
CHAPTER 4 FEATURE DEVELOPMENT FOR WRITER IDENTIFICATION.....		69
4.1	INTRODUCTION	69
4.2	OVERLAPPED GRADIENT DISTRIBUTION FEATURES (OGDF)	71
4.3	GRADIENT DISTRIBUTION FEATURES (GDF)	75
4.4	WINDOWED GRADIENT DISTRIBUTION FEATURES (WGDF)	77
4.5	CONTOUR CHAIN CODE DISTRIBUTION FEATURES (C^3DF)	79
4.6	WINDOWED CONTOUR CHAIN CODE DISTRIBUTION FEATURES (WC^3DF)	81
4.7	MEDIUM SCALE FEATURES (MSF)	81
4.7.1	<i>Density Features (DF)</i>	83
4.7.2	<i>Horizontal and Vertical Run Features (HVRF)</i>	83
4.7.3	<i>Stroke Features (SF)</i>	83
4.7.4	<i>Concavity Features (CF)</i>	84
4.8	CONNECTED COMPONENT FEATURES (CCF)	84
4.9	CONCLUSIONS.....	89
CHAPTER 5 WRITER IDENTIFICATION USING ARABIC HANDWRITTEN DIGITS.....		91
5.1	INTRODUCTION	92
5.2	FEATURES	93
5.2.1	<i>Overlapped Gradient Distribution Features (OGDF)</i>	95
5.2.2	<i>Contour Chain Code Distribution Features (C^3DF)</i>	96
5.2.3	<i>Medium Scale Features (MSF)</i>	96
5.3	ARABIC DIGITS DATABASE	97
5.4	NEAREST NEIGHBOR (NN) AND NEAREST MEAN (NM) CLASSIFIER.....	99
5.5	FEATURE SELECTION.....	101
5.5.1	<i>Optimal Number of Grid Segments</i>	102
5.5.2	<i>Selecting Overlapped Gradient Distribution Features (OGDF) Sliding Window Size</i>	104
5.5.3	<i>Feature Combinations Selection</i>	108
5.6	EXPERIMENTAL RESULTS	111
5.7	CONCLUSIONS.....	116
CHAPTER 6 WRITER IDENTIFICATION OF ARABIC HANDWRITTEN TEXT		117
6.1	INTRODUCTION	118
6.2	FEATURES	120
6.2.1	<i>Connected Component Features (CCF)</i>	120
6.2.2	<i>Gradient Distribution Features (GDF)</i>	120
6.2.3	<i>Windowed Gradient Distribution Features (WGDF)</i>	121
6.2.4	<i>Contour Chain Code Distribution Features (C^3DF)</i>	121
6.2.5	<i>Windowed Contour Chain Code Distribution Features (WC^3DF)</i>	121

6.3	EXPERIMENTAL RESULTS	124
6.3.1	<i>Database Used</i>	124
6.3.2	<i>Classifier Selection</i>	125
6.3.3	<i>Feature Combination</i>	129
6.3.4	<i>Dimensionality Reduction</i>	131
6.3.5	<i>Writer Identification Accuracy vs. Number of Writers</i>	133
6.3.6	<i>Comparison with Other Published Results</i>	135
6.4	CONCLUSIONS.....	137
CHAPTER 7 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS		138
7.1	CONCLUSIONS.....	138
7.2	FUTURE DIRECTIONS	140
NOMENCLATURE		143
REFERENCES		145
VITA.....		157

LIST OF TABLES

TABLE 1-1: ARABIC LETTERS' FOUR DIFFERENT SHAPES.	5
TABLE 2-1: DATABASES USED IN WRITER IDENTIFICATION/VERIFICATION.	25
TABLE 2-2: HISTOGRAM COMPARISON BETWEEN CORPUS AND SELECTED 1000 PARAGRAPHS.	32
TABLE 2-3: ELEVEN COMMON WORDS/SENTENCES.	38
TABLE 2-4: SUMMARY OF THE DESIGNED FORMS.	38
TABLE 3-1: GSC FEATURE DEFINITIONS.	50
TABLE 3-2: WRITER IDENTIFICATION/VERIFICATION FEATURES AND CLASSIFIERS.	55
TABLE 4-1: OVERVIEW OF FEATURES USED IN WRITER IDENTIFICATION OF ARABIC HANDWRITTEN DIGITS AND TEXT.	90
TABLE 5-1: EXP1 RESULTS (ALL SAMPLES USING THE NN CLASSIFIER) FOR DIFFERENT SLIDING WINDOW SIZES.	106
TABLE 5-2: EXP2 RESULTS (ALL SAMPLES USING THE NM CLASSIFIER) FOR DIFFERENT SLIDING WINDOW SIZES.	106
TABLE 5-3: EXP3 RESULTS (GROUP OF DIGITS, 0 TO 9, ARE CONCATENATED TO FORM ONE FEATURE VECTOR) FOR DIFFERENT SLIDING WINDOW SIZES.	107
TABLE 5-4: EXP4 RESULTS (EACH DIGIT IN TESTING SET IS COMPARED AGAINST ALL DIGITS IN TRAINING AND THEN DISTANCES ARE ADDED FOR EACH GROUP OF DIGITS) FOR DIFFERENT SLIDING WINDOW SIZES.	107
TABLE 5-5: EXP1 RESULTS (ALL SAMPLES USING THE NN CLASSIFIER) FOR DIFFERENT FEATURE COMBINATIONS.	109
TABLE 5-6: EXP2 RESULTS (ALL SAMPLES USING THE NM CLASSIFIER) FOR DIFFERENT FEATURE COMBINATIONS.	109
TABLE 5-7: EXP3 RESULTS (GROUP OF DIGITS, 0 TO 9, ARE CONCATENATED TO FORM ONE FEATURE VECTOR) FOR DIFFERENT FEATURE COMBINATIONS.	110
TABLE 5-8: EXP4 RESULTS (EACH DIGIT IN TESTING IS COMPARED AGAINST ALL DIGITS IN TRAINING AND THEN DISTANCES ARE ADDED FOR EACH GROUP OF DIGITS) FOR DIFFERENT FEATURE COMBINATIONS.	110
TABLE 5-9: WRITER IDENTIFICATION AND DIGIT RECOGNITION ACCURACY FOR EACH DIGIT USING NN. ..	112
TABLE 5-10: WRITER IDENTIFICATION AND DIGIT RECOGNITION ACCURACY FOR EACH DIGIT USING NM.	113
TABLE 5-11: TEXT-DEPENDENT WRITER IDENTIFICATION.	115
TABLE 5-12: TEXT-INDEPENDENT WRITER IDENTIFICATION.	115
TABLE 6-1: SAMPLE FEATURE VECTOR FOR FIGURE 6-1 (A).	122
TABLE 6-2: DISTANCE MEASURES EQUATIONS.	127
TABLE 6-3: DISTANCE MEASURES RESULTS ON WRITER IDENTIFICATION FOR 100 WRITERS.	128
TABLE 6-4: FEATURE SELECTION RESULTS ON WRITER IDENTIFICATION FOR 100 WRITERS.	132
TABLE 6-5: COMPARISON OF WRITER IDENTIFICATION SYSTEMS.	136

LIST OF FIGURES

FIGURE 1-1: SAMPLE PROCESS FOR OFFLINE WRITER IDENTIFICATION.....	4
FIGURE 1-2: (A) OBLIGATORY LIGATURES (B) OPTIONAL LIGATURES.	8
FIGURE 2-1: CEDAR LETTER: A) SOURCE DOCUMENT, B) SCANNED SAMPLE (S. SRIHARI, S. H. CHA, ARORA, & LEE, 2002).	17
FIGURE 2-2: A SAMPLE IAM FILLED FORM (MARTI & BUNKE, 2002)	17
FIGURE 2-3: FIREMAKER PAGE 2: A) SOURCE DOCUMENT, B) SCANNED SAMPLE (SCHOMAKER & BULACU, 2004).	19
FIGURE 2-4: SAMPLES OF THE UNIPEN (A), TRIGRAPH (B), HIFCD2 (C), IRONOFF (D), AND RIMES (E) DBs.	21
FIGURE 2-5: AN EXAMPLE OF AN IFN/ENIT FILLED FORM (EL ABED & MÄRGNER, 2007).	24
FIGURE 2-6: FREE HANDWRITING SAMPLE FROM THE AHDC DATASET (AL-MA'ADEED, ELLIMAN, & HIGGINS, 2002).	24
FIGURE 2-7: GUI FOR IDENTIFYING, COUNTING AND VISUALIZING OF CHARACTER SHAPES IN TEXT.	30
FIGURE 2-8: MINIMUM PARAGRAPH SAMPLE IMAGE.	30
FIGURE 2-9: PARAGRAPH DATABASE SAMPLE IMAGE.	34
FIGURE 2-10: LIGATURES FORM SAMPLE IMAGE.	37
FIGURE 2-11: PART OF THE COMMON WORDS/SENTENCES FORM SAMPLE IMAGE.	37
FIGURE 3-1: TEXTURE AND ALLOGRAPH FEATURES (BULACU & SCHOMAKER, 2007A).	47
FIGURE 3-2: EXEMPLAR WORD IMAGE WITH 4 X 8 DIVISIONS USING GSC (B. ZHANG, 2003).	51
FIGURE 3-3: THE FOUR ADDITIONAL FARSI ISOLATED CHARACTERS.	66
FIGURE 4-1: (A) AND (B): TWO SAMPLE IMAGES FOR WRITER X, (C): SAMPLE IMAGE FOR WRITER Y.	72
FIGURE 4-2: X AND Y SOBEL OPERATOR MASKS.	72
FIGURE 4-3: FIRST AND SECOND GRADIENT FEATURE BINS.	74
FIGURE 4-4: A) HISTOGRAM FOR OGDF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	74
FIGURE 4-5: EXAMPLES OF GRADIENT ANGLES FOR FOUR REGIONS OF INTEREST.	76
FIGURE 4-6: A) HISTOGRAM FOR GDF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	76
FIGURE 4-7: A) HISTOGRAM FOR WGDF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	78
FIGURE 4-8: FREEMAN CHAIN CODES RELATIVE TO THE CENTER POINT.	80
FIGURE 4-9: A) HISTOGRAM FOR C^3DF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	80
FIGURE 4-10: A) HISTOGRAM FOR WC^3DF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	82
FIGURE 4-11: A) HISTOGRAM FOR MSF FEATURE VECTOR FOR THE SAME WRITER, B) FOR DIFFERENT WRITERS.	82
FIGURE 4-12: LETTER SEGMENTED IMAGE.	86

FIGURE 4-13: SAMPLES OF CONNECTED COMPONENTS OF ALEFS, CIRCLES, HALF CIRCLES, AND HORIZONTAL SEGMENTS, RESPECTIVELY.	86
FIGURE 4-14: A) SAMPLE HISTOGRAM FOR CCF FEATURE VECTOR FOR THE SAME WRITER (LEFT), B) FOR DIFFERENT WRITERS (RIGHT).....	88
FIGURE 5-1: ARABIC DIGIT 6 DIVIDED INTO 4 DIFFERENT DIVISIONS.	94
FIGURE 5-2: SAMPLES OF ADBASE.....	98
FIGURE 5-3: WRITER IDENTIFICATION ACCURACY AT DIFFERENT DIVISIONS.....	103
FIGURE 5-4: DIGIT '9' (9) DIVIDED INTO 5 X 5 DIVISIONS.	103
FIGURE 6-1: (A) PARAGRAPH DATABASE SAMPLE IMAGE (TOP). (B),(C) VERTICAL DIVISION OF SAMPLE IMAGE (BOTTOM).	123
FIGURE 6-2: WRITER IDENTIFICATION RECOGNITION ACCURACY FOR DIFFERENT FEATURES (100 WRITERS).	130
FIGURE 6-3: TOP - 1, TOP - 5, AND TOP - 10 ACCURACY % VS. THE AS DSA NUMBER OF WRITERS.	134

PUBLICATIONS

Following is the list of journal and conference papers, at the time of submitting this dissertation, based on the work in this dissertation and other research interests.

Journal Papers

- **S.M. Awaidah** and S.A. Mahmoud, “Writer Identification of Arabic Text using Statistical and Structural Features,” 2011, Submitted for review.
- **S.M. Awaidah** and S.A. Mahmoud, “State of the Art in Off-line Writer Identification of Handwritten Text and Survey of Writer Identification of Arabic text,” 2011, Submitted for review.
- Gutub, L. Ghouti, Y. Elarian, **S. Awaideh**, and A. Alvi, “Utilizing Diacritic Marks for Arabic Text Steganography,” Kuwait Journal of Science & Engineering (KJSE), vol. vol.37, 1B, 2010, pp. 89-110.
- **S.M. Awaidah** and S.A. Mahmoud, “A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models,” Signal Processing, vol. 89, Jun. 2009, pp. 1176-1184.
- S. Mahmoud and **S. Awaida**, “Recognition of Off-Line Handwritten Arabic (Indian) Numerals Using Multi-Scale Features and Support Vector Machines vs. Hidden Markov Models,” Arabian Journal for Science and Engineering (AJSE), vol. 34, 2009, pp. 429-444.

Conference Papers

- **S.M. Awaida** and S.A. Mahmoud, “Writer Identification of Arabic Handwritten Digits,” 1st International Workshop on Frontiers in Arabic Handwriting Recognition, Istanbul, Turkey: 2010.
- **S.M. Awaidah** and S.A. Mahmoud, “The Identifiability of Arabic Handwritten Digits,” First Scientific Conference of higher education students in KSA, Riyadh, KSA: 2010, p. 100.
- M.A. Awad, M.A. Logmani, **S. Awaida**, and W. Al-Khatib, “Empirical Evaluation of LSI-Based Indexing Techniques of Arabic Gigaword Documents,” H.R. Arabnia, R.R. Hashemi, and F.I. Moxley, eds., Las Vegas Nevada, USA: CSREA Press, 2009, pp. 674-680.
- Gutub, Y. Elarian, **S. Awaideh**, and A. Alvi, “Arabic Text Steganography Using Multiple Diacritics,” WoSPA 2008 – 5th IEEE International Workshop on Signal

Processing and its Applications, University of Sharjah, Sharjah, U.A.E: 2008, p. 18 – 20.

- M.A. Aabed, **S.M. Awaideh**, A.-R.M. Elshafei, and A.A. Gutub, “Arabic diacritics based steganography,” IEEE International Conference on Signal Processing and Communications, 2007. ICSPC 2007, 2007, p. 756–759.

Patents

- **S.M. Awaida**, and H. Al-Muhtaseb, “Method of Generating a Transliteration Font”, (Patent Pending), (2010), U.S. Patent and Trademark Office (USPTO).

ABSTRACT

NAME: SAMEH MOHAMMAD AWADA
TITLE OF STUDY: WRITER IDENTIFICATION OF ARABIC
HANDWRITTEN DOCUMENTS
MAJOR FIELD: COMPUTER SCIENCE AND ENGINEERING
DATE OF DEGREE: JUNE, 2011

The issues related to writer identification are currently at the heart of numerous concerns in our modern day's society. Writer identification for Arabic text is receiving a renewed attention. We anticipate that the developed research techniques and algorithms in this thesis to help in establishing this area of research.

Writer identification of off-line Arabic handwritten text and digits is addressed by utilizing the state of the art identification and verification techniques, features, and classifiers. The presented identifiability of handwritten digits provides quantitative measurements for the uniqueness of each digit. The research work has also shown that we can design a successful writer identification system from only the basic 10 Arabic digits given that other writers write the same digits.

A successful writer identification system for handwritten Arabic text is designed and developed. Since there is no available Arabic database for this application, this research includes building a database for handwritten Arabic text by 250 writers. Several types of structural and statistical features are extracted. Connected component features as well as Overlapped gradient distribution features, gradient distribution features, windowed gradient distribution features, contour chain code distribution features, windowed contour chain code distribution features, density features, horizontal and vertical run length features, stroke features, and concavity shape features are implemented.

The accuracy results of each feature type are compared using statistical significance. The effects of increasing the number of writers on the accuracy results are presented and analyzed. Experimental results of applying these features on Arabic text and digits are presented. Feature reduction and selection techniques (e.g. PCA, LDA, MDA, ...) are applied and shown to improve both computation time and accuracy results.

خلاصة

درجة الدكتوراة في الفلسفة

الاسم: سامح محمد عويضة
عنوان الرسالة: تحديد هوية صاحب الخط في الوثائق العربية المكتوبة بخط اليد
التخصص: علوم وهندسة الحاسب الآلي
تأريخ التخرج: حزيران 2011

أصبحت القضايا المتعلقة بالتعرف على هوية الشخص محط اهتمام في مجتمعنا العصري الحديث. إن مجال تحديد هوية صاحب الخط في الوثائق العربية يحظى باهتمام متجدد. من المتوقع أن تساعد تقنيات البحث والخوارزميات المتقدمة في هذه الأطروحة في تقدم مجال تحديد هوية صاحب الخط العربي من البحث العلمي.

تناول هذه الأطروحة الأساليب المختلفة لتحديد هوية صاحب الخط في الوثائق العربية المكتوبة بخط اليد من نصوص وأرقام من خلال الاستفادة من التقنيات والسمات وعناصر التصنيف الحديثة. وقد أظهرت الأبحاث المقيّمة لتمييز الأرقام المكتوبة بخط اليد قدرة كل رقم على تمييز صاحب الخط، وأظهرت الأبحاث أيضا إمكانية تصميم نظام لتحديد هوية صاحب الخط باستخدام الأرقام الأساسية العربية العشرة.

تم تصميم وتطوير نظام فعال لتحديد هوية صاحب خط الأرقام والنصوص العربية المكتوبة بخط اليد. وقد شمل هذا العمل بناء قاعدة بيانات تتضمن نصوصا مكتوبة بخط اليد باللغة العربية من قبل 250 كاتب اذ لا توجد أية قاعدة بيانات عربية متاحة لهذا الغرض. في هذا العمل تم استخراج عدة أنواع من السمات الهيكلية والإحصائية من نص الخط العربي، وهذه السمات هي: ميزات الاطارات المتصلة، وميزات توزيع التدرج المتراكب، وميزات توزيع التدرج المتراكب ذي النوافذ، وميزات توزيع الانحدار، وميزات سلاسل محيط الأشكال، وميزات سلاسل محيط الأشكال ذي النوافذ، وميزات قياس الكثافة، وميزات قياس طول المسار الأفقي والرأسي، وميزات قياس جرة القلم ، وميزات قياس تقعر الشكل.

تمت مقارنة نتائج دقة النظام باستخدام الدلالة الإحصائية لكل سمة على حدة على النص العربي والأرقام العربية كما تم عرض النتائج التجريبية الحاصلة من تطبيق هذه الميزات. وقد تم تحليل وعرض آثار زيادة عدد الكتاب على دقة النتائج، بالإضافة الى تطبيق وعرض تقنيات لتقليل واختيار السمات التفصيلية (مثل PCA ، LDA ، MDA ، ...) لتحسين الدقة وتقليل الوقت اللازم لحساب نتائج نظام تحديد هوية صاحب الخط.

CHAPTER 1

INTRODUCTION

Writer identification is the process of determining the author from a set of possible writers through samples of his/her handwriting (Schlapbach, 2007). Writer verification is the process of comparing questioned handwriting with samples of handwriting obtained from known sources for the purposes of determining authorship or non-authorship (R. R. Bradford & R. B. Bradford, 1992). Writer verification involves an accept/reject decision-making criteria whilst writer identification involves a one-to-many classification problem and hence is considered more challenging (Gibbons, Yoon, S.-H. Cha, & Tappert, 2005; Zaher & Abu-Rezq, 2010). In recent years, writer identification and verification has become a common application used in confirming the document authenticity in the financial sector as well as revealing the identity of suspected criminals, etc.

1.1 Writer Identification vs. Text Recognition

Although both handwritten text recognition and writer identification are considered to be parts of the pattern recognition field, text recognition differs from writer identification in that they seek to maximize opposite characteristics. The objective of writer identification is to maximize the inter-writer variations and recognize the uniqueness of each writer with little regard to the text content. The objective of text

recognition is to minimize inter-writer differences for the same text and identify the text content (S. Srihari, S. H. Cha, Arora, & Lee, 2002). Nevertheless, the two fields have used similar techniques in feature extraction and classification as will be shown in more details in Chapter 3.

1.2 Text Dependent and Text Independent Writer Identification

Writer identification can be divided into two categories; text-dependent and text-independent writer identification. Text-dependent writer identification systems require certain known text to be written, whereas text-independent writer identification systems can work on any given text. In this work, research involving text-dependent and text-independent writer identification of offline handwritten text is addressed.

1.3 Online and Offline Writer Identification

Databases can be for printed text or for handwritten text. Handwritten databases are usually divided into offline and online, where online databases contain data of pen trajectory and offline databases contain digital images of the writing. Writer identification systems that process online images are labeled online writer identification, whereas those that process offline images are considered offline writer identification. Since online data contains useful information for writer identification (i.e. time order and pen pressure) that is lost in offline databases; offline writer identification/verification is considered more challenging than online writer identification/verification (He & Tang, 2004). Figure 1-1 shows a sample block diagram of the process of offline writer identification for Arabic

documents. The sample process shows the scanning of the documents from different writers, preprocessing done to the scanned documents, feature extraction to underline the distinctive properties of the scanned images while at the same time reducing its dimensionality, and finally the classification step done to recognize the writer.

1.4 Background on the Arabic Language

Arabic language is spoken by more than 280 million people as a first language (Brown (ed.), 2006). In addition, the Arabic alphabet, i.e. script, is used in a wide variety of other languages that include Urdu, Persian, Malay, and Kurdish. The Arabic script is also used by 1.41-1.57 billion Muslims worldwide (Central Intelligence Agency, 2003), as it is the language of the Qur'an, the holy book of Islam. The Arabic script is written from right to left, cursively, and contains 28 basic letters.

Since the language is cursive, some letters can have up to four basic shapes. The shape of a character depends on the connectivity characteristics of the previous and subsequent characters. The four shapes are hereafter referred to as: the "Beginning Shape", the "Middle Shape", the "Ending Shape" and the "Isolated Shape". Some letters do not connect to subsequent letters; hence, they are only allowed to take one of two shapes, viz. the "Ending Shape" and the "Isolated Shape". The four possible Arabic letters shapes are shown in Table 1-1.

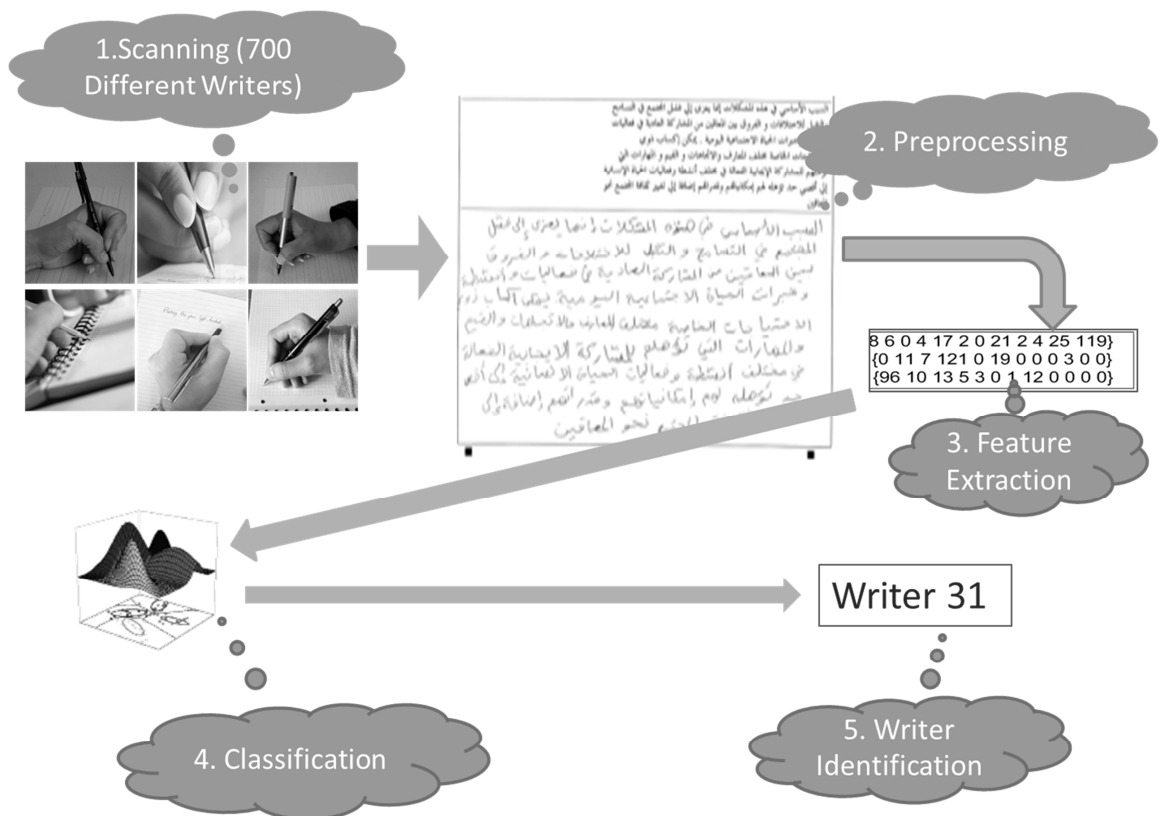


Figure 1-1: Sample process for offline writer identification.

Table 1-1: Arabic letters' four different shapes.

Name	Isolated	Beginning	Middle	Ending
Aleph	ا			ا
Ba	ب	بـ	بـ	بـ
Ta	ت	تـ	تـ	تـ
Tha	ث	ثـ	ثـ	ثـ
Jeem	ج	جـ	جـ	جـ
Ha	ح	حـ	حـ	حـ
Kha	خ	خـ	خـ	خـ
Dal	د			د
Thal	ذ			ذ
Ra	ر			ر
Zai	ز			ز
Seen	س	سـ	سـ	سـ
Sheen	ش	شـ	شـ	شـ
Sad	ص	صـ	صـ	صـ
Dad	ض	ضـ	ضـ	ضـ
Ta	ط	طـ	طـ	طـ
Tha	ظ	ظـ	ظـ	ظـ
Ain	ع	عـ	عـ	عـ
Ghain	غ	غـ	غـ	غـ
Fa	ف	فـ	فـ	فـ
Qaf	ق	قـ	قـ	قـ
Kaf	ك	كـ	كـ	كـ
Lam	ل	لـ	لـ	لـ
Meem	م	مـ	مـ	مـ
Noon	ن	نـ	نـ	نـ
Ha	هـ	هـ	هـ	هـ
Waw	و			و
Ya	ي	يـ	يـ	يـ

Ligatures may occur when certain characters appear in sequence. The characters combine in a way that is not the mere concatenation of the original character shapes. Often, they embrace vertical overlapping between the character shapes that form it. Only the family of “*Lam-Alef*” ligatures is mandatory in Arabic script. Figure 1-2 shows examples of the characters before and after forming ligatures, where Figure 1-2 (a) shows that the family of “*Lam-Alef*” only accepts mandatory ligatures and the mere concatenation of the two letters is considered incorrect.

1.5 Problem Statement

This thesis addresses the task of writer identification of Arabic handwritten text. The writer identifiability using Arabic digits and paragraphs is explored. Since there is no publicly available Arabic database for this work, this research includes building a database for handwritten Arabic text by many writers.

1.6 Significance of the Study

Automatic offline writer identification has enjoyed renewed interest over the last twenty years. One of the driving forces for this surge is the increasing need for writer identification techniques by forensic document examiners to identify criminals based on their handwriting (D. Zhang, Jain, Tapiador, & Sigüenza, 2004). Furthermore, threats of terrorist attacks have increased the use of writer identification and other biometric recognition techniques to identify the assailants (Schlapbach, 2007). In May 13, 1999, the

United States vs. Paul decided that Handwriting analysis qualifies as expert testimony and is therefore admissible (S. Srihari, S. H. Cha, Arora, & Lee, 2002).

Writer identification for Arabic text, which was not researched as thoroughly as Latin, Japanese, or Chinese, is receiving a renewed attention not only from Arabic speaking researchers but also from non-Arabic speaking researchers. The developed research techniques, algorithms, and publications in this field will help in establishing this area of research for the interested research community. In addition, it is expected that the developed state of the art research technology for writer identification of Arabic handwritten documents in this thesis may be used by other researchers to develop applications and related research contributions in the following fields:

1. Writer identification of Arabic and Islamic historical manuscripts.
2. Crime suspects identification in forensic sciences.
3. Forgery detection.
4. Bank Checks verification.

لا آ	لا أ
× لا آ	× لا أ
✓ لا آ	✓ لا أ

(a)

صدر	لم	بج
صدر	لم	بج
صدر	لـ	بج

(b)

Figure 1-2: (a) Obligatory ligatures (b) Optional ligatures.

1.7 Contributions of the Thesis

In this work, we addressed writer identification using off-line Arabic handwritten text by utilizing the state of the art identification and verification techniques, features, and classifiers for Arabic writing. The following are the contributions made in this thesis:

- 1- We conducted a literature survey of recent writer identification research for Arabic-like languages (Arabic, Urdu, Persian, etc.) as well as state of the art writer identification research for Latin and other languages. To the best of the author's knowledge, no surveys that specifically target writer identification and verification have been previously published.
- 2- Development of a new writer identification/verification database: Since there is no comprehensive database publicly available for use for this type of research, we built a database of Arabic handwritten text. The dataset design specifies four forms. These forms include the minimum Arabic paragraph that covers all character shapes, randomly selected paragraphs from a corpus, the ligatures form, and the common words/sentences form. Two hundred and fifty different writers have filled the forms. The collected data is used for training and testing the writer identification research. The database together with the associated research tools may also be used by Arabic computing researchers and students. For example, it can be used by the research

community in automatic recognition of Arabic handwriting, handwritten document analysis, Arabic handwritten text synthesis, and as a benchmark database.

3- Features design and selection: Initially different types of features that are used for Latin languages were tested. Subsequently, novel statistical and structural features that are suitable for Arabic writer identification research are designed. Classical features are modified for improved writer identification on Arabic text and digits, i.e. overlapped gradient distribution features, gradient distribution features, windowed gradient distribution features, contour chain code distribution features, windowed contour chain code distribution features, density features, horizontal and vertical run length features, stroke features, and concavity shape. Connected component features for Arabic handwritten text are original statistical and structural features that build on some of the main characteristics of the Arabic language. Several types of experiments are performed to choose the optimal number of features, best feature parameters, and the best combination of features.

4- The writer identifiability using Arabic digits is researched. To the best of the author's knowledge, no other research works have studied writer identification using Arabic digits. Measuring the discriminative power of digits offers insight into the most

discriminating digits for writer identification using Arabic handwritten documents. Analysis of the results indicates that writer identification systems using Arabic digits ‘٣’ (3), ‘٤’ (4), ‘٨’ (8), and ‘٩’ (9) are more identifiable than using other digits while writer identification systems using Arabic digit ‘٠’ (0) and ‘١’ (1) are the least identifiable. In addition, the research shows that combining the writer's digits increases the discriminability power of writer identification. Combining the features of all digits, Nearest Neighbor classifier provided the best accuracy in text-independent writer identification with top-1 result of 88.14%, top-5 result of 94.81%, and top-10 results of 96.48%. Results will help guide future researchers on what digits to use in their writer identification work. Digit recognition has also been applied in order to validate the effectiveness of the features in digit recognition as well as writer identification.

5- Text-independent writer identification using Arabic handwritten documents is addressed. A database of large number of writers (250 writers) is used in the analysis and experimentations. This is the first research effort for writer identification using Arabic handwritten paragraphs with such a large number of writers. Previous researchers either used less number of writers (< 100 writers) or handwritten words instead of full paragraphs. Effects of increasing the number of writers on the identification accuracy are

considered and analyzed. We also compare results from our developed system with those reported for other Latin and Arabic writer identification systems that used similar-sized databases. For all 250 writers, the system attained a top-1 result of 75.0%, top-5 result of 91.8%, and top-10 results of 95.4%.

1.8 Thesis Outline

This thesis is organized as follows: Chapter 2 provides an extensive review of databases used in writer identification and verification for Latin (and other western languages) and Arabic text. In addition, this chapter details the design of a natural and representative handwritten Arabic dataset that can be used for writer identification and other research fields. Chapter 3 presents the state of the art in writer identification and verification of handwritten text. In addition, a survey of writer identification and verification of Arabic handwritten text is also included. Feature design is described in Chapter 4. Arabic handwritten digits are analyzed for writer identifiability in Chapter 5. In addition to writer identification using digits, the chapter presents digit recognition; Chapter 6 addresses writer identification of Arabic handwritten text. Several types of structural and statistical features are extracted from Arabic handwriting text. A novel approach is used to extract structural features that build on some of the main characteristics of the Arabic language. Finally, conclusions and future directions are addressed in Chapter 7.

CHAPTER 2

DATABASE SURVEY AND DESIGN

This chapter provides a survey that focuses on Arabic and Latin handwritten databases. An extensive review of databases used in writer identification and verification for Latin (and other western languages) and Arabic text is presented. In addition, this chapter describes the design and implantation of a natural and representative handwritten Arabic database that can be easily collected and truth-grounded. The database consists of several forms to cover different aspects that might be needed by the users of the dataset. Four different forms that can be used in text-dependent and text-independent writer identification/verification are collected from more than 250 writers. The four forms consist of the minimum paragraph form that covers all character shapes of Arabic script, the full page form, the ligatures form, and the common words/sentences form.

2.1 Introduction

Databases for character recognition exist for printed text and handwritten text, where databases for writer identification consist of handwritten text. Handwritten databases are usually divided into offline and online, where online databases contain data of pen trajectory and offline databases contain digital images of the handwriting. Since online data contains useful information for writer identification (i.e. time order and pen pressure) that is lost in offline databases; offline writer identification/verification is

considered more challenging than online writer identification/verification (He & Tang, 2004). This chapter targets mainly offline handwritten databases.

Image datasets of handwriting samples, along with their ground truths, are essential for the development of techniques for document analysis and classification, writer identification, handwriting synthesis, etc. Researchers involved in Latin Optical Character Recognition (OCR) and writer identification systems enjoy the privilege of using standard datasets for developing their systems and comparing their results with published work using the same datasets (Al-Badr & S. Mahmoud, 1995).

Many researchers – e.g. (Al-Badr & S. Mahmoud, 1995; Al-Ohali, Cheriet, B, & Suen, 2003; Amin, 1998; Margner & Pechwitz, 2001) - consider the absence of standard datasets to be one of the main causes of Arabic text recognition systems lagging behind. The same claim can be made for Arabic writer identification systems. Al-Badr and Mahmoud (Al-Badr & S. Mahmoud, 1995) stated that the field crucially needs a standard dataset and performance evaluation tools. Clearly, it is necessary to build Arabic datasets to train, test and compare writer identification and recognition systems (Al-Ohali et al., 2003).

Designing and collecting an Arabic handwritten dataset encompasses several challenges. The processes of collecting and truth grounding writer information can be cumbersome and time consuming. A good design can reduce many subsequent efforts. The dataset size needs to be reasonably concise. Besides, the content should be natural in order to realistically depict writers' normal behavior. The data collected must represent and depict the characteristics of Arabic script in quantity and diversity. In addition, the gathered samples should preferably be representative of all Arabic speaking regions.

We have described in detail 12 Latin, Arabic, and other western databases and tabulated a total of 33 databases. In surveying writer identification and verification databases, we included all the papers we had access to in writer identification/verification. The databases described are either public datasets that are available to researchers or private datasets that had strong influence on the field of the writer identification/verification.

The chapter is organized as follows; Section 2 addresses the databases used for writer identification and verification of western scripts; Arabic handwritten databases used in writer identification and verification research are discussed in Section 3; Design of the Arabic database for writer identification/verification is presented in Section 4; and finally conclusions are stated in Section 5.

2.2 Databases for Writer Identification and Verification of Western Script

In this section the main databases used for writer identification and verification of handwritten Latin and other western scripts are described. The CEDAR letter was developed in the University of Buffalo (S. Cha & S. Srihari, 2000), and is considered one of the first large databases developed for writer identification and verification of handwritten Latin scripts. The CEDAR Letter, as shown in Figure 2-1, is concise (it has just 156 words) yet still each alphabet letter occurs in the beginning of a word as a capital and a small letter, and as a small letter in the middle and end of a word. In addition, the database also contains punctuations, numerals, and some letter and numeral combinations

(for example, ff, tt, oo, 00). The CEDAR letter was written by 1,000 individuals three times each (S. Cha & S. Srihari, 2000). Noticeably, (S. Srihari, S. H. Cha, Arora, & Lee, 2002) reported that the CEDAR letter was written by 1500 writers.

The IAM-database (Marti & Bunke, 2002) consists of handwritten English sentences that are based on the Lancaster-Oslo/Bergen (LOB) corpus (Johansson, Leech, & Goodluck, 1978). The corpus is a collection of texts that comprise about one million word instances. The database originally included 1,066 forms produced by approximately 400 different writers, and was later extended to include a total of 1539 forms produced by 657 different writers. The database consists of full English sentences. Figure 2-2 shows a sample filled form of the IAM database. Due to its public availability, flexible structure, and large number of writers involved, the IAM database has been commonly used for Latin writer identification/verification by a number of researchers, for example (A. Bensefia, T. Paquet, & L. Heutte, 2005; Ameer Bensefia, Thierry Paquet, & Laurent Heutte, 2005; Brink, Bulacu, & Schomaker, 2008; Bulacu, 2007; Bulacu & Schomaker, 2006; 2007a; Helli & Moghaddam, 2009; Schlapbach & Bunke, 2004a; 2004b; 2007; Schlapbach, Kilchherr, & Bunke, 2005; Schomaker & Bulacu, 2004; Siddiqi & Vincent, 2007; 2008; 2009). Researchers have used the IAM database alone (Brink, Bulacu, & Schomaker, 2008; Schlapbach & Bunke, 2007; Siddiqi & Vincent, 2008) or combined/compared it with other databases (Bulacu, 2007; Bulacu & Schomaker, 2006; 2007a; Schomaker & Bulacu, 2004; Siddiqi & Vincent, 2009).

From
 Jim Elder
 829 Loop Street, Apt 300
 Allentown, New York 14707

Nov 10, 1999

To
 Dr. Bob Grant
 602 Queensberry Parkway
 Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
 Jim

Nov 10, 1999

From
 Jim Elder
 829 Loop Street, Apt 300
 Allentown, N.Y. 14707

To
 Dr. Bob Grant
 602 Queensberry Parkway
 Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
 Jim

Figure 2-1: Cedar letter: a) source document, b) scanned sample (S. Srihari, S. H. Cha, Arora, & Lee, 2002).

Sentence Database M01-012

He slapped himself in the face and cuffed the sides of his head. Then by degrees the rotating objects slowed, and coming into focus took the form of the furnishings in Dan Brown's living room. He stood up unsteadily and looked about the room, trying to gather his wits. Outside the dusk was settling over Dow's Lake and the heights beyond were in silhouette, already a solid black.

He slapped himself in the face and cuffed the sides of his head. then by degrees the rotating objects slowed, and coming into focus took the form of the furnishings in Dan Brown's living room. He stood up unsteadily and looked about the room, trying to gather his wits. Outside the dusk was settling over Dow's Lake and the heights beyond were in silhouette, already a solid black.

Name _____

Figure 2-2: A sample IAM filled form (Marti & Bunke, 2002) .

The Firemaker dataset (Schomaker & Vuurpijl, 2000) consists of 1008 scanned pages of handwritten Dutch texts written by 252 students, four pages each. Page 1 contains a copied text in natural writing style; Page 2 contains copied upper-case text; Page 3 contains copied forged text (where writers write as to impersonate other people) while Page 4 contains a self-generated description of a cartoon image in free writing style. The text to be copied has been designed to cover a sufficient amount of different letters from the alphabet while remaining conveniently writable for the majority of writers. Figure 2-3 shows an example of Page 2.

Since the Firemaker database was not publicly available for some time, it has been mostly used by the researchers in the University of Groningen, for example (Brink, Bulacu, & Schomaker, 2008; Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007a; Bulacu, Schomaker, & Vuurpijl, 2003; Schomaker, Bulacu, & M. van Erp, 2003; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007) with one exception (Maaten & Postma, 2005). Lately, the Firemaker database has been publicly available (Int. Unipen Foundation, 2011). It should be noted that Schomaker et al. have combined parts of the Firemaker database with parts of the IAM database to make a western script database of 900 writers (Brink, Bulacu, & Schomaker, 2008; Bulacu, 2007; Bulacu & Schomaker, 2006; 2007a; Schomaker & Bulacu, 2004).

Other public Western handwritten databases used in writer identification/verification include the Unipen dataset, the Trigraph Slant Dataset, the HIFCD2 dataset, IRONOFF dataset, and the RIMES dataset. A brief description for each database follows next.

NADAT ZE IN NEW YORK, TOKYO, QUÉBEC, PARIJS, ZÜRICH
EN OSLO WAREN GEWEEST, VLOGEN ZE UIT DE USA TERUG
MET VLUCHT KL. 658 OM 12 UUR.

ZE KWAMEN AAN IN DUBLIN OM 7 UUR EN IN AMSTERDAM
OM 9:40 UUR 'S AVONDS. DE FIAT VAN BOB EN DE VW VAN
DAVID STONDEN IN R3 VAN HET PARKEERTERRAIN. HIER-
VOOR MOESTEN ZE HONDERD GULDEN (F 100,-) BETALEN.

NADAT ZE IN NEW YORK
QUÉBEC, PARIJS, ZÜRICH EN
GEWEEST, VLOGEN ZE UIT DE
MET VLUCHT KL. 658 OM

Figure 2-3: Firemaker page 2: a) Source document, b) Scanned sample (Schomaker & Bulacu, 2004).

The UNIPEN project (I. Guyon, Schomaker, Plamondon, Liberman, & Janet, 1994) described a format and methodology for creating a database for online handwritten text from several countries and languages, and has organized the collection of more than 5 million handwritten characters of more than 2200 writers. Offline images have been derived from the UNIPEN online database and used in writer identification (Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007a; Niels, Vuurpijl, & Schomaker, 2007; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007). The TriGraph Slant Dataset is a recent database that contains images for 47 writers of handwriting, produced under conditions of normal and disguised slant (Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010). The HIFCD2 database contains handwritten samples for the word ‘characteristic’ and its equivalent Greek word written 45 times for each writer, for a total of 50 total (Zois & Anastassopoulos, 2000). The IRESTE On/Off (IRONOFF) dual handwriting database (C. Viard-Gaudin, Lallican, Binter, & Knerr, 1999) contains French letters and words for 700 writers. It is dual in the sense that it contains both online data (pen trajectory) and offline data (digital images) for the same writing. The RIMES French database contains more than 5600 real mails written by 1300 writers completely annotated as well as secondary databases of isolated characters, handwritten words (300,000 snippets) and logos (Grosicki, Carré, Brodin, & Geoffrois, 2008). Figure 2-4 shows samples of the UNIPEN, TriGraph, HIFCD2, IRONOFF, and RIMES databases, respectively. As mentioned previously, all of these databases are available publicly for research purposes.

(a)	(b)
<p>Britain has a clear strategy for eradicating BSE. It consists mainly of banning the feed that causes the disease. The strategy has been in place since 1988 and has been highly successful. Cases of BSE have been fallen by 67% since their peak in 1992</p>	
(c)	(d)
(e)	
<p>Objet : Demande d'informations</p> <p>Monsieur, Monsieur,</p> <p>Ma femme et moi-même étions parents depuis peu, nous avons ressenti la nécessité d'assurer l'avenir financier de notre enfant pour le cas où nous arriverions à di's par suite -</p> <p>Je vous serais donc reconnaissant de bien vouloir nous transmettre une documentation concernant vos contrats d'assurance-vie.</p> <p>Si vous n'avez pas réponse,</p> <p>Cordialement,</p>	

Figure 2-4: Samples of the UNIPEN (a), TriGraph (b), HIFCD2 (c), IRONOFF (d), and RIMES (e) DBs.

2.3 Databases Used in Writer Identification of Arabic Text

The IfN/ENIT database (El Abed & Märgner, 2007) was created by the Institute of Communications Technology (IfN) at Technical University Braunschweig in Germany and the Ecole Nationale d'Inge'nieurs de Tunis (ENIT) in Tunisia. The database consists of 26,459 images of the of names of 937 cities and towns in Tunisia, written by 411 different writers. To this date, this database has been widely used by many researchers of Arabic handwritten text recognition (more than 100 research groups from more than 30 countries) and has appeared in several global competitions (Märgner & El Abed, 2007; 2009; 2010; 2011; Märgner, Pechwitz, & H.E. Abed, 2005). Due to its public availability, researchers have also used the IfN/ENIT database for writer identification of Arabic text (Abdi, Khemakhem, & Ben-Abdallah, 2009; Bulacu, Schomaker, & Brink, 2007; Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010; Chawki & Labiba, 2010; Lutf, Xinge You, & H. Li, 2010) although it is limited to city names and thus contains limited vocabulary. Figure 2-5 shows an example of a filled form of the IfN/ENIT database.

Al-Ma'adeed et al. presented the AHDB database (Al-Ma'adeed, Elliman, & Higgins, 2002), which contains Arabic words and texts written by one hundred writers. It also contains the most popular words in Arabic as well as sentences used in writing bank checks with Arabic words. Finally it contains free handwriting pages in a topic of interest to the writer. The form was designed in five pages. The first three pages were filled with ninety-six words, sixty-seven of which are handwritten words corresponding to textual words of numbers that can be used in handwritten check writing. The other twenty-nine

words are from the most popular words in Arabic writing. The fourth page is designed to contain three sentences of handwritten words representing numbers and quantities that can be written on bank checks. The fifth page is lined, and as shown in Figure 2-6 is designed to be completed by the writer in freehand on any subject of his choice. Further information, such as the availability of the dataset for public use, is not clear from the authors' published work. Al-Ma'adeed et al. used their database for Arabic writer identification in (Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008).

Ball et al. used a much smaller database for writer identification of Arabic handwritten text (G. R. Ball & Sargur N. Srihari, 2008) prepared from 10 different writers, each contributing 10 different full page documents in handwritten Arabic for a total of 100 documents.

Gazzah and Ben Amara (Gazzah & Ben Amara, 2006; 2007; 2008) designed their own Arabic letter database which contains 505 characters, 15 numerals and 6 punctuations. The choice of the letter contents was made to ensure the use of the various internal shapes of the letter within a sub-word (isolated, initial, middle and end). Handwriting samples of 60 persons were collected. Each person was required to copy the same letter three times: two samples were used for training and the other for the testing; a total of 180 A4 format sample pages.

Table 2-1 shows a summary of handwritten text databases used for writer identification and verification. It shows the databases used in writer identification & verification of handwritten text, the number of writers of each database, the language of the text, and published research work in which these databases are used.

code	description	code
6132	حساب ايراضة	حساب ايراضة 6132
2056	رزاد	رزاد 2056
2014	مستودع الرياض	مستودع الرياض 2014
4283	نقطة	نقطة 4283
2064	جبل الزمام	جبل الزمام 2064
1200	التعدين	التعدين 1200
7030	ماطر	ماطر 7030
1251	الخرايج	الخرايج 1251
3233	قطونة	قطونة 3233
2112	مبنى محمد زروق	مبنى محمد زروق 2112
110	المركبات	المركبات 110
2261	سيارة	سيارة 2261

Age:	< 20 <input type="checkbox"/>	21 - 30 <input checked="" type="checkbox"/>	31 - 40 <input type="checkbox"/>	> 40 <input type="checkbox"/>
Profession:	Etudiant/élève <input checked="" type="checkbox"/>	Enseignant <input type="checkbox"/>	Administratif <input type="checkbox"/>	Autre <input type="checkbox"/>
Num:	Moukt Nizar			
Ville:	Ariana			
Responsible:	Dawid St		Numero:	c71.

Figure 2-5: An example of an IFN/ENIT filled form (El Abed & Märgner, 2007).

يهدف البحث إلى دراسة الخواص الحرارية والضوئية والميكانيكية لمادة البوليمر والبحث عن تغير خواصها بفعل العوامل المؤثرة لكي نعرف مدى استجابتها للمؤثرات الخارجية بالتشيع والتعرض للجو الخارجي وتأثير ماء البحر. وذلك من أجل تحسين الاداء العملي. تستخدم هذه المادة في الصناعات كالعوازل الكهربائية ومحطات التغليف والطب وغيره. حيث تعود هذه الأهمية إلى سهولة التصنيع وانخفاض التكلفة وسهولة التشكيل.

Figure 2-6: Free handwriting sample from the AHDC dataset (Al-Ma'adeed, Elliman, & Higgins, 2002).

Table 2-1: Databases used in Writer Identification/Verification.

DB#	DB Name	DB Reference	Database Used in	Public	Language	Type	#Writers
DB01	na*	(Gazzah & Ben Amara, 2006)	(Gazzah & Ben Amara, 2006; 2007; 2008)	No	Arabic	Text	60
DB02	na*	(Al-Dmour & Zitar, 2007)	(Al-Dmour & Zitar, 2007)	No	Arabic	Text	20
DB03	IFN/ENT	(El Abed & Märgner, 2007)	(Abdi, Khemakhem, & Ben-Abdallah, 2009; Bulacu, Schomaker, & Brink, 2007; Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010; Chawki & Labiba, 2010; Lutf, Xinge You, & H. Li, 2010)	Yes	Arabic	Words	411
DB04	AHDB	(Al-Ma'adeed, Mohammed, & Al Kassis, 2008)	(Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008)	No	Arabic	Words /Phrases	100
DB05	na*	(S. Srihari & G. Ball, 2008)	(S. Srihari & G. Ball, 2008)	No	Arabic	Text	10
DB06	na*	(C.-L. Liu, Dai, & Y.-J. Liu, 1995)	(C.-L. Liu, Dai, & Y.-J. Liu, 1995)	No	Chinese	Characters	20
DB07	na*	(Zhu, T. Tan, & Y. Wang, 2000)	(Zhu, T. Tan, & Y. Wang, 2000)	No	Chinese	Text	17
DB08	na*	(Cong, Xiao-Gang, & Tian-Lu, 2002)	(Cong et al., 2002)	No	Chinese	Text	50
DB09	na*	(He & Tang, 2004)	(He & Tang, 2004)	No	Chinese	Text	50
DB10	na*	(He, Bin, Jianwei, Yuan Yan, & Xinge, 2005)	(He, Bin, Jianwei, Yuan Yan, & Xinge, 2005; He, Tang, & X. You, 2005)	No	Chinese	Text	10
DB11	HIT-MW	(Su, T. Zhang, & Guan, 2007)	(X. Li & Ding, 2009)	Yes	Chinese	Text	240
DB12	na*	(He, X. You, & Tang, 2008a)	(He, X. You, & Tang, 2008a; 2008b)	No	Chinese	Text	500
DB13	SET1 SET2	(X. Wang, Ding, & H. Liu, 2003)	(X. Wang, Ding, & H. Liu, 2003)	No	Chinese	Characters	25 626
DB14	Firemaker	(Schomaker & Vuurpijl, 2000)	(Brink, Bulacu, & Schomaker, 2008; Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007a; Bulacu, Schomaker, & Vuurpijl, 2003; Maaten & Postma, 2005; Schomaker & Bulacu, 2004; Schomaker, Bulacu, & M. van Erp, 2003; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007)	Yes	Dutch	Text	250
DB15	Unipen	(I. Guyon, Schomaker, Plamondon, Liberman, & Janet, 1994)	(Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007a; Niels, Vuurpijl, & Schomaker, 2007; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007)	Yes	Various	Text	215
DB16	IAM	(Marti & Bunke, 2002)	(Brink, Bulacu, & Schomaker, 2008; Bulacu, 2007; Bulacu & Schomaker, 2006; 2007a; Helli & Moghaddam, 2009; Schlapbach & Bunke, 2007; Schomaker & Bulacu, 2004; Siddiqi & Vincent, 2008; 2009)	Yes	English	Text	657
DB17	na*	(Bulacu & Schomaker, 2007b)	(Bulacu & Schomaker, 2007b)	No	Medieval English	Text	10
DB18	Trigraph	(Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010)	(Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010)	Yes	Dutch	Text	47
DB19	na*	(Hull, 1994)	(S. Srihari, 2000)	No	English	Words,	na*

						Digits	
DB20	Cedar Letter	(S. Cha & S. Srihari, 2000)	(S. Srihari & G. Ball, 2009; S. Srihari, Beal, Bandi, Shah, & Krishnamurthy, 2005; S. Srihari, S. H. Cha, Arora, & Lee, 2002; S. Srihari, Huang, Srinivasan, & Shah, 2007; Tomai & S. Srihari, 2004; B. Zhang, 2003; B. Zhang, S. Srihari, & Lee, 2003)	No	English	Text	1000
DB21	na*	(Matsuura & Qiao, 1989)	(Matsuura & Qiao, 1989)	No	English	Words	2
DB22	na*	(Said, Baker, & T. Tan, 1998)	(Said, Baker, & T. Tan, 1998)	No	English	Text	20
DB23	na*	(Leedham & Chachra, 2003)	(Leedham & Chachra, 2003)	No	English	Digits	15
DB24	HIFCD2	(Zois & Anastassopoulos, 2000)	(Zois & Anastassopoulos, 2000)	Yes	English & Greek	Words	50
DB25	IRONOFF	(C. Viard-Gaudin, Lallican, Binter, & Knerr, 1999)	(G. Tan, Christian Viard-Gaudin, & Kot, 2008)	Yes	French	Letters /Words	700
DB26	na*	(A. Bensefia, Nosary, T. Paquet, & L. Heutte, 2002)	(A. Bensefia et al., 2002; A. Bensefia, T. Paquet, & L. Heutte, 2003a)	No	French	Text	88
DB27	RIMES	(Grosicki et al., 2008)	(Siddiqi & Vincent, 2009)	Yes	French	Text	1300
DB28	na*	(Bar-Yosef, Beckman, Kedem, & Dinstein, 2007)	(Bar-Yosef et al., 2007)	No	Historical Hebrew	Characters	34
DB29	na*	(Mar & Thein, 2005)	(Mar & Thein, 2005)	No	Myanmar	Characters	20
DB30	na*	(Shahabi & Rahmati, 2007)	(Shahabi & Rahmati, 2006; 2007)	No	Persian	Text	40
DB31	PD100	(Helli & Moghaddam, 2008b)	(Helli & Moghaddam, 2008a; 2008b; 2009; 2010)	No	Persian	Text	100
DB32	na*	(Ram & Moghaddam, 2009b)	(Ram & Moghaddam, 2009a; 2009b)	No	Persian	Text	50
DB33	na*	(Ubul, Hamdulla, Aysa, Raxidin, & Mahmut, 2008)	(Ubul et al., 2008)	No	Uyghur	Text	23

na*: Information is not available.

2.4 Database Design for New Arabic Handwritten Text¹

The database is intended to be useful to several applications including writer identification and text recognition. Other applications include handwritten document analysis & classification, Arabic handwritten text synthesis, and as a benchmark database by the research community. For that goal, several forms are designed to address the different needs of such applications. Several design issues are considered when specifying each form of the datasets. Some of the important issues discussed below include: script completeness, natural frequency, and representativeness of writers.

Script completeness is what we assure for comprehensiveness in a form. If the character is the unit to be covered, then all character shapes must appear at least once in the form. By the natural frequency model, we refer to the frequencies with which each character shape appears in the form. Uniform distributions are sometimes useful for training. However, imitating real-world frequencies can be more accurate when testing and benchmarking are considered. Finally, the number of writers has to be large and contains different nationalities to be representative of the Arabic speaking community.

2.4.1 Motivation

The used databases for Arabic text writer identification do not match the representation and naturalness qualities of the databases available for Latin text. Available Arabic databases are individual efforts with inherent limitations in size and

¹ The database design and collection in this section is a collaborative work between Prof. Sabri Mahmoud, Dr. Mohammad Tanvir Parvez, Mr. Yousef Elarian, and myself.

comprehension. The IEF/ENIT consists of city names and researchers had to concatenate a number of city names to make an Arabic text. This is neither a natural nor a comprehensive representation of the Arabic text.

There is a need for an Arabic text database with a large number of writers for writer identification and verification. So far there is no Arabic text database that is publicly available for writer identification of Arabic text. This section of database design and collection is a first step to solve this problem.

Taking the previously mentioned limitations in Arabic handwritten database into considerations; four different forms were collected for each writer; the minimum paragraph form, the full page form, the ligatures form, and the common words/sentences form. A brief description of each form is as follows: In the first form, all character shapes of Arabic script are covered (minimum paragraph form), the natural distribution of Arabic character shapes from a large corpus is maintained in the second form (full page form), representative ligatures are grouped in different groups for the writers to make writing them practical in the third form (ligatures form), and common words/sentences in Arabic literature are collected in the fourth form (common words/sentences form). What follows is a more detailed description of each form.

Up till now, a total of 250 volunteers filled the forms. Except for an exterior box to write the paragraph inside, no constraints were employed in terms of pen usage, lined paragraphs, or writing certain number of words per line. The collected samples are scanned at 300 dpi, the scanned images are deskewed, then the handwritten paragraphs are extracted, and the images are binarized using a simple thresholding technique.

2.4.2 **Minimum Paragraph Form**

This form aims at covering all character shapes of Arabic script, by the smallest possible amount of text. One improvement we thought of is to obtain a paragraph that covers all character shapes, and hence can be more natural for volunteers to read and write. To help come up with such coherent and covering paragraph, we developed a GUI tool that counts and visualizes the shapes of characters in its input text area. The GUI is shown in Figure 2-7. A sample page of the form is shown in Figure 2-8 below. This form can be used for text-dependent writer identification.

2.4.3 **Full Page Form**

Arabic Gigaword corpus (Graff, 2007), which is electronically available, is an archive of newswire text data from Arabic news sources that have been collected over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. It contains 1994735 million documents with almost 577 million words. We selected the corpus as the source of texts that will be used in our full page form.

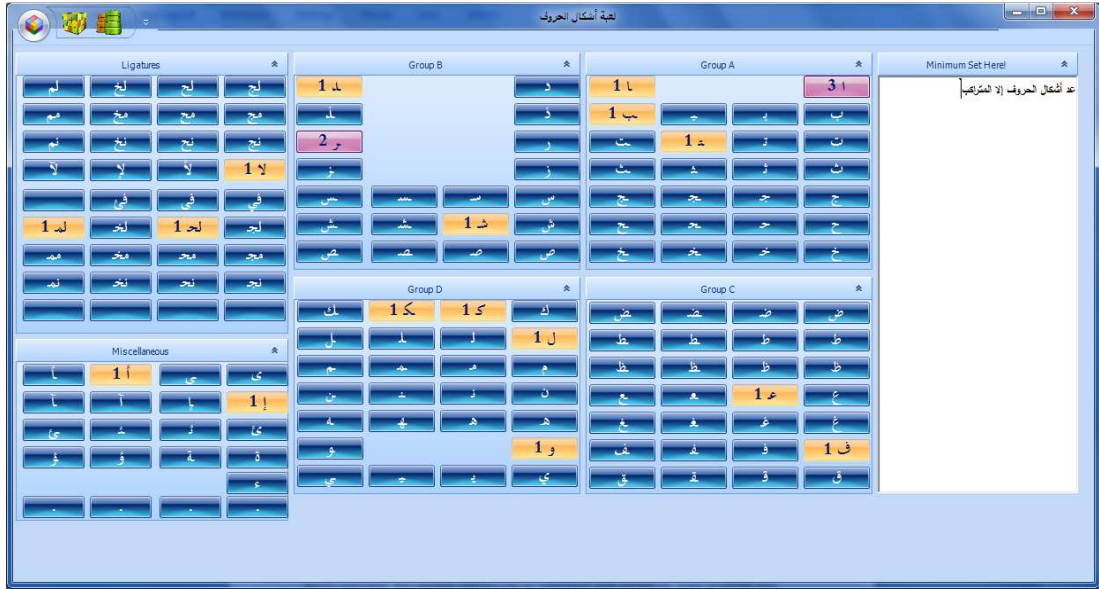


Figure 2-7: GUI for identifying, counting and visualizing of character shapes in text.

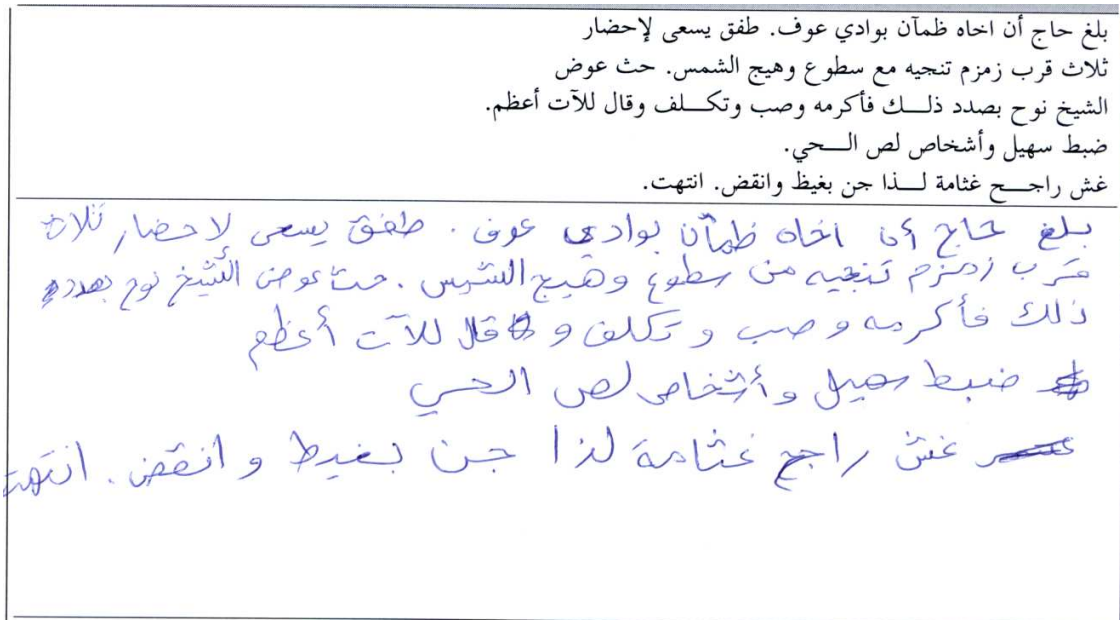


Figure 2-8: Minimum paragraph sample image.

In order to create a full-page dataset from the Gigaword corpus, we split the texts in the corpus into paragraphs by removing any tags or headers from the document. Next, we trim each paragraph from its end to contain only 50 words maximum and an average of 5 sentences long. We generate 1000 paragraphs randomly. These paragraphs were distributed in the volunteers' forms using the previously mentioned approach. The histogram and probabilities for each letter shape in the whole Gigaword corpus is estimated and used as a representation of Arabic text. The histogram of the extracted 1000 paragraphs is determined. If the histogram of the 1000 paragraphs doesn't match that of the corpus, paragraphs are replaced by more representative ones. This ensures that the extracted texts are representative of the Gigaword corpus in terms of the normalized frequencies of Arabic letter shapes. The used corpus contains 272,325,605 Arabic letters and the 1000 paragraphs contain 316,096 Arabic letters. Table 2-2 shows the percentages of each letter shape in both the corpus and the selected paragraphs. Average difference for all percentages is 0.1258%. A sample page of the form is shown in Figure 2-9. This form was used for text-independent writer identification explained in more detail in Chapter 5.

Table 2-2: Histogram comparison between corpus and selected 1000 paragraphs.

Letter Shape	Corpus	1000 Forms	Absolute Difference
ا	10.9664%	10.0194%	0.9469%
آ	7.0143%	7.1719%	0.1576%
ب	0.2346%	0.2091%	0.0255%
بـ	1.8611%	1.6337%	0.2274%
پ	1.2501%	1.3047%	0.0546%
پـ	0.2623%	0.3464%	0.0841%
ت	0.9566%	1.0978%	0.1411%
تـ	1.3595%	1.6397%	0.2802%
ث	2.3096%	2.7432%	0.4336%
ثـ	0.3771%	0.2597%	0.1173%
ج	0.0662%	0.0747%	0.0084%
جـ	0.2040%	0.1696%	0.0344%
چ	0.2528%	0.3018%	0.0491%
چـ	0.0728%	0.1202%	0.0474%
ح	0.0583%	0.0566%	0.0017%
حـ	0.7261%	0.6634%	0.0627%
خ	0.3865%	0.4606%	0.0741%
خـ	0.0245%	0.0620%	0.0375%
د	0.0937%	0.0914%	0.0023%
دـ	0.6968%	0.7260%	0.0293%
ذ	0.5758%	0.6141%	0.0383%
ذـ	0.0631%	0.0724%	0.0093%
ر	0.0072%	0.0022%	0.0050%
رـ	0.4007%	0.4673%	0.0666%
ز	0.1832%	0.2911%	0.1078%
زـ	0.0209%	0.0127%	0.0082%
س	1.2091%	1.0570%	0.1522%
سـ	2.1463%	2.2060%	0.0597%
ش	0.1674%	0.2006%	0.0332%
شـ	0.4130%	0.5881%	0.1751%
ص	1.9162%	1.7001%	0.2161%
صـ	3.4843%	2.9317%	0.5526%
ض	0.3415%	0.3385%	0.0030%
ضـ	0.4268%	0.3410%	0.0857%
ط	0.1055%	0.0914%	0.0141%
طـ	1.2160%	1.0946%	0.1214%
ظ	1.1899%	1.0361%	0.1538%
ظـ	0.3094%	0.3439%	0.0345%
ع	0.0196%	0.0028%	0.0168%
عـ	0.4749%	0.2850%	0.1899%
غ	0.6498%	0.7355%	0.0857%
غـ	0.0439%	0.0155%	0.0284%
ف	0.0343%	0.0528%	0.0186%
فـ	0.3752%	0.3433%	0.0319%
ق	0.6952%	0.5796%	0.1156%
قـ	0.0277%	0.0487%	0.0211%
ک	0.0749%	0.0807%	0.0058%
کـ	0.3303%	0.3268%	0.0035%
گ	0.2563%	0.2860%	0.0297%
گـ	0.0505%	0.0579%	0.0074%
ن	0.0316%	0.0380%	0.0063%
نـ	0.3408%	0.3186%	0.0222%
ی	0.6279%	0.6628%	0.0349%
یـ	0.0869%	0.0949%	0.0080%
ر	0.0056%	0.0082%	0.0026%
رـ	0.0359%	0.0449%	0.0090%
ز	0.1259%	0.2047%	0.0788%
زـ	0.0089%	0.0111%	0.0022%
س	0.1761%	0.1895%	0.0134%
سـ	1.4488%	1.6435%	0.1947%

ع	1.4127%	1.6704%	0.2577%
ع	0.3643%	0.4331%	0.0688%
غ	0.0032%	0.0051%	0.0019%
غ	0.1689%	0.1155%	0.0534%
غ	0.2130%	0.1392%	0.0738%
غ	0.0961%	0.0161%	0.0799%
ف	0.1638%	0.1879%	0.0241%
ف	0.5817%	0.6251%	0.0435%
ف	0.6335%	0.5916%	0.0419%
ف	0.1607%	0.1575%	0.0032%
ق	0.2026%	0.1848%	0.0178%
ق	1.0384%	0.9696%	0.0687%
ق	1.1660%	1.2480%	0.0820%
ق	0.1964%	0.2227%	0.0263%
ك	0.0918%	0.0807%	0.0111%
ك	0.9010%	0.8842%	0.0168%
ك	0.7469%	0.8010%	0.0541%
ك	0.1090%	0.1740%	0.0650%
ل	0.6296%	0.6077%	0.0219%
ل	5.6891%	5.8950%	0.2059%
ل	1.7054%	1.8852%	0.1798%
ل	0.6337%	0.7899%	0.1563%
م	0.6054%	0.4578%	0.1476%
م	2.6877%	2.5847%	0.1030%
م	1.3268%	1.5517%	0.2249%
م	0.4153%	0.5476%	0.1323%
ن	1.2059%	0.9696%	0.2363%
ن	1.3187%	1.0241%	0.2947%
ن	1.6599%	1.5660%	0.0939%
ن	1.5427%	1.4366%	0.1062%
ه	0.1559%	0.2360%	0.0801%
ه	0.5872%	0.7672%	0.1800%
ه	0.8587%	1.1149%	0.2562%
ه	0.4061%	0.5065%	0.1003%
و	3.0766%	3.4433%	0.3666%
و	3.1096%	2.5793%	0.5303%
ي	0.0822%	0.3420%	0.2598%
ي	2.3808%	2.0231%	0.3577%
ي	4.3610%	3.5499%	0.8111%
ي	0.2332%	0.9247%	0.6915%
ي	0.0035%	0.0019%	0.0016%
ن	0.5187%	0.4056%	0.1131%
ن	0.0874%	0.1193%	0.0318%
ي	0.0042%	0.0022%	0.0020%
ة	0.7716%	0.8384%	0.0668%
ة	3.0982%	3.3291%	0.2309%
ي	0.4153%	0.1588%	0.2564%
ي	1.5679%	0.7874%	0.7805%
ؤ	0.0114%	0.0146%	0.0031%
ؤ	0.1432%	0.1879%	0.0447%
ء	0.4047%	0.4116%	0.0069%
أ	0.4766%	1.0522%	0.5757%
أ	0.1118%	0.2588%	0.1470%
أ	0.1511%	0.4688%	0.3178%
أ	0.0106%	0.0383%	0.0277%
أ	0.0277%	0.0402%	0.0125%
أ	0.0042%	0.0044%	0.0002%

السبب الأساسي في هذه المشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات وأنشطة وخبرات الحياة الاجتماعية اليومية . يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين

السبب الأساسي في هذه المشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات وأنشطة وخبرات الحياة الاجتماعية اليومية . يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين

Figure 2-9: Paragraph database sample image.

2.4.4 Ligatures Form

Ligatures importance in Arabic writer identification and text recognition research has been mostly unexploited. Some ligature forms appear more frequently than some character shapes. Moreover, some ligatures are mandatory, as stated previously. For that reason, it is important to depict the behavior of writing these ligatures. Since most ligatures are optional and depend on the writer's style of writing, it is not enough to add these in the Minimum Paragraphs Form to assure their presence as some writers might abandon writing them as ligatures. Besides, since they are bigrams and trigrams, they are far more in count than single characters. For these reasons, a special ligature form with special specifications is added.

To find an appropriate list of ligatures for our goal, we computed the Cartesian product of character shapes. Ligatures are formed when a character is connectable to the next one. Hence, the first character shapes are restricted to be either Beginning or Middle shapes; and the next character shapes are restricted to be either of the Middle or the Ending shape. The resulting set of bigram ligatures contains 528 bigram ligatures from a total of 2622 connected bigrams (i.e. around 20% of the total). Since this number is not practical to implement, a subset was selected.

To form a ligature subset, we chose only the isolated ligatures. The reduction results in 184 ligatures (i.e. around 7% of the original connected bigrams). The number is still large to be filled by a single user. Hence, these were divided into groups of 27 ligatures for each user. In this grouping, we put one form of the ligatures that share a glyph (but differ in dotting and/or Hamzah) for the same writer. This way, we maximize

the contribution of each writer, in terms of basic glyphs. Figure 2-10 shows a subset of one of the ligature forms for a random writer. Since ligature forms are different for each group of writers, these forms can be used in both text-dependent (within each group) and text-independent (across different groups) writer identification/verification.

2.4.5 Common Words/Sentences Form

The common words/sentences form consists of five common Arabic sentences and six common words that are to be copied six times by each writer. These common words were selected by considering the Arabic traditional literature and by studying other researchers' work (Al-Ma'adeed, Mohammed, & Al Kassis, 2008; Buckwalter, 2002). Table 2-3 shows the eleven common words/sentences collected in the fourth form. Figure 2-11 shows part of the common words/sentences form. The fourth form can be used in text-dependent writer identification/verification. Since it contains common words/sentences that are found in the traditional Arabic literature, it should be helpful in the study of historical writer identification/verification.

2.4.6 Summary of the Designed Form

Table 2-4 gives a summary of the four designed forms. They clearly cover different stated needs, as can be seen especially from the last two columns. In summary, each writer writes the minimum set, writes one paragraph, one set of ligatures, and one set of common sentences/words. The database has been already collected from 250 writers so far.

لسي	سم	سر
لي	مع	سر
عج	طح	صي
عج	طح	صي
كج	كا	في
كج	كا	في

Figure 2-10: Ligatures form sample image.

قال	قال	قال	قال
قال	قال	قال	قال
مع	مع	مع	مع
مع	مع	مع	مع
الحمد لله	الحمد لله	الحمد لله	الحمد لله
الحمد لله	الحمد لله	الحمد لله	الحمد لله
بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم
بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم

Figure 2-11: Part of the common words/sentences form sample image.

Table 2-3: Eleven common words/sentences.

مع	قال	الذي	على	من	في
صلى الله عليه وسلم		الحمد لله		رضي الله عنه	
بسم الله الرحمن الرحيم			لا اله الا الله		

Table 2-4: Summary of the designed forms.

ID	Name	Covering Unit	Covered Unit	Coverage Model	Coverer
1	Minimum Text	Sentences/Paragraph	4 character shapes	Uniform	Each writer
2	Full Page	Paragraphs	4 character shapes	Real-World	Collectively
3	Ligatures	Set of ligatures	2 Ligatures shapes	Altered Real-World	Sets of writers
4	Common	Words and Sentences	Common words/sentences	Uniform	Each writer

2.5 Conclusions

In this chapter we presented the state of the art in the databases used in writer identification and verification of Arabic, Latin, and other western text. The previously published databases in the field of writer identification/verification were tabulated indicating the number of writers, samples, language, etc.

We think that the used data for Arabic text writer identification does not match the representation and naturalness qualities of the databases available for Latin text. We consider that our database design and collection is a first step to solve this problem.

We have designed and implemented an Arabic handwriting dataset for handwritten document analysis & classification, Arabic writer identification, and Arabic handwritten text synthesis. The database may also be used as a benchmark database by the research community. The dataset design specifies four forms and is expected to be easy to fill. These forms include the minimum Arabic paragraph that covers all character shapes, randomly selected paragraphs from a corpus, the ligatures form, and the common words/sentences form. So far, 250 different writers have filled the forms.

CHAPTER 3

LITERATURE REVIEW

In this chapter, we present the state of the art in writer identification and verification of handwritten text. In addition, an extensive survey of writer identification and verification of Arabic handwritten text is also included. Feature extraction techniques are addressed showing the different research groups' efforts as well as individual efforts. The different classification approaches used for writer identification and verification are presented and identification results of surveyed publications are tabulated for ease of reference. Examples of writer identification and verification of other languages are given. Conclusions relevant to writer identification of Arabic text are discussed and future directions stated.

3.1 Introduction

This chapter presents the state of the art in writer identification and verification using handwritten text with a special survey on writer identification and verification of Arabic handwritten text. For advances in the field prior to the year 1990, the reader is referred to (Plamondon & Lorette, 1989). Due to the similarity and close relationships of the used techniques, signature verification and handwriting recognition surveys such as (Plamondon, 1994; Plamondon & S.N. Srihari, 2000) usually discuss the state of the art

in writer identification and verification as well. The author is not aware of any survey specifically targeting writer identification and verification.

In this chapter, research involving text-dependent and text-independent writer identification of offline handwritten text is surveyed. We have included more than 100 accessible publications related to writer identification and verification. However, we cannot claim that we have addressed all published work in writer identification and verification of Latin or other languages. We tried our best to include the work of all the major research groups and individuals in the field. In surveying writer identification and verification of Arabic text, we included all the papers we could access and also incorporated research on Persian (Farsi) text because of its similarity to the Arabic script.

Quite interestingly, automatic offline writer identification has enjoyed renewed interest over the last twenty years, and more specifically during the past decade. One of the driving forces for this surge is the increasing need for writer identification techniques by forensic document examiners to identify criminals based on their handwriting (D. Zhang, Jain, Tapiador, & Sigüenza, 2004). Furthermore, threats of terrorist attacks have increased the use of writer identification and other biometric recognition techniques to identify the assailants (Schlapbach, 2007).

Although research in writer identification and verification is still dominated by the English language, research on other languages includes Chinese (Cong et al., 2002; He, Bin, Jianwei, Yuan Yan, & Xinge, 2005; He & Tang, 2004; He, Tang, & X. You, 2005; He, X. You, & Tang, 2008a; 2008b; X. Li & Ding, 2009; X. Li, X. Wang, & Ding, 2006; C.-L. Liu, Dai, & Y.-J. Liu, 1995; Su et al., 2007; X. Wang, Ding, & H. Liu, 2003; Zhu, T. Tan, & Y. Wang, 2000). Dutch (Brink, Niels, van Batenburg, van Den Heuvel, &

Schomaker, 2010; Maaten & Postma, 2005; Schomaker & Bulacu, 2004); Greek (Zois & Anastassopoulos, 2000), French (A. Bensefia et al., 2002; A. Bensefia, T. Paquet, & L. Heutte, 2003b; 2003a; 2004; Ameer Bensefia, Thierry Paquet, & Laurent Heutte, 2005; Siddiqi & Vincent, 2009), Japanese (Yoshimura, 1988), Uyghur (Ubul et al., 2008), Myanmar (Mar & Thein, 2005), Arabic (Abdi, Khemakhem, & Ben-Abdallah, 2009; Al-Dmour & Zitar, 2007; Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008; Bulacu, Schomaker, & Brink, 2007; Gazzah & Ben Amara, 2006; 2007; 2008; S. Srihari & G. Ball, 2008), Persian (Helli & Moghaddam, 2008a; 2008b; 2009; 2010; Ram & Moghaddam, 2009a; 2009b; Shahabi & Rahmati, 2006; 2007), as well as historical manuscripts and inscriptions in different ancient languages (Bar-Yosef et al., 2007; A. Bensefia, T. Paquet, & L. Heutte, 2003a; Bulacu & Schomaker, 2007b; Panagopoulos, Papaodysseus, Rousopoulos, Dafi, & Tracy, 2009; Schomaker, Katrin Franke, & Bulacu, 2007). Leedham and Chachra also implemented writer identification using Latin handwritten numerals (Leedham & Chachra, 2003).

3.2 Applications of Writer Identification and Verification

One of the main applications of writer identification and verification is its use in forensic sciences (Katrin Franke & Koppen, 2001; Katrin Franke et al., 2003; Niels, Vuurpijl, & Schomaker, 2007; S. Srihari, S. H. Cha, Arora, & Lee, 2002; D. Zhang, Jain, Tapiador, & Sigüenza, 2004). Identification of a person based on an arbitrary handwritten sample is a useful application. Writer identification allows for determining the suspects in conjunction with the inherent characteristic of a crime, e.g. the case of threat letters. This

is different than other biometric methods, where the relation between the evidence material and the details of an offense can be quite remote (Schomaker & Bulacu, 2004).

In addition to forensic applications of writer identification and verification, several other applications exist, including:

- Ink type recognition (Katrin Franke, Bünnemeyer, & Sy, 2002).
- Script and language identification (Hochberg, Bowers, Cannon, & Kelly, 1999).
- Forgery detection (Leedham & Chachra, 2003).
- Writer identification on medieval and historical documents (Bar-Yosef et al., 2007; A. Bensefia, T. Paquet, & L. Heutte, 2003a; Bulacu & Schomaker, 2007b; Panagopoulos, Papaodysseus, Rousopoulos, Dafi, & Tracy, 2009; Schomaker, Katrin Franke, & Bulacu, 2007).
- Writer identification on handwritten musical scores (Fornes, Lladós, Sanchez, & Bunke, 2008).
- Personalized handwriting text recognizers (Rodríguez-Serrano, Perronnin, Sánchez, & Lladós, 2010).

3.3 Feature Extraction Approaches

In this subsection we consider feature extraction research in the field of writer identification and verification using Latin and western texts. Feature extraction techniques and classification approached for Arabic and Persian texts will be covered in

section 3.5. The feature extraction approach has a crucial effect on the accuracy of any writer identification system. Feature extraction is used to underline the distinctive properties of an object under consideration while at the same time reducing its dimensionality. Researchers used different types of features for writer identification. Some of these features are also used in automatic handwritten text recognition. This section presents the types of features that have been used in writer identification and verification. Features used by groups of researchers in writer identification and verification will be presented in conjunction followed by other researchers' work. Categorizing features by research groups allows the reader to see the combination of features in their appropriate scope. It also indicates how these features were developed over time and the different applications and data used with these features.

Bensefia (A. Bensefia et al., 2002; A. Bensefia, T. Paquet, & L. Heutte, 2003b; Aneur Bensefia, Thierry Paquet, & Laurent Heutte, 2005) used graphemes generated by segmenting handwritten text to identify writers. Graphemes are commonly defined as the written representation of phonemes (Rey, Ziegler, & Jacobs, 2000). These graphemes are then clustered using sequential clustering algorithm. Clustering is repeated and graphemes that fall in the same clusters in these repeated clustering are kept in these clusters. Graphemes that change clusters are kept in separate clusters. First-level graphemes, bi-grams and tri-grams are used. Bi-grams and tri-grams of graphemes are connected and features extracted. This technique is applied to two datasets containing different number of writers; a self-built database of 88 writers and 150 writers of the IAM database (Marti & Bunke, 2002). Recognition rates reported on their own database were 93%, 95.45%, and 80% using first-level graphemes, bi-grams, and tri-grams respectively.

Schomaker et al. (Brink, Bulacu, & Schomaker, 2008; Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010; Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007a; 2007b; Bulacu, Schomaker, & Brink, 2007; Bulacu, Schomaker, & Vuurpijl, 2003; Katrin Franke et al., 2003; Niels, Vuurpijl, & Schomaker, 2007; Schomaker & Bulacu, 2004; Schomaker, Bulacu, & M. van Erp, 2003; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007) used two level analysis for feature extraction; texture level and character-shape (allograph) level. At the texture level, they used contour-direction Probability Distribution Function (PDF) ($p(\phi)$), where ϕ is the contour direction as shown in Figure 3-1 (a)), contour-hinge PDF ($p(\phi_1, \phi_2)$), where ϕ_1, ϕ_2 are the angles of the two sides of the hinge as shown in Figure 3-1 (b)), direction co-occurrence ($p(\phi_1, \phi_3)$), where ϕ_1, ϕ_3 are the angles with the horizontal- and vertical-run, as shown in Figure 3-1 (c)), the probability distribution of the white run lengths PDFs, and autocorrelation in horizontal scan. The contour-direction PDF features are assumed to capture orientation and curvature information, the contour-hinge PDF to capture the curvature of the contour, and the direction co-occurrence to measure the roundness of the written characters.

At the allograph level, graphemes were used. These features were initially applied to uppercase letters with success (Schomaker & Bulacu, 2004), and were later applied to cursive text. The graphemes were extracted as connected components. For each connected component, its contour was computed using Moore's algorithm (Gonzalez & Woods, 2007). Inner contours were discarded. The PDF of these connected components (graphemes) was computed using a common codebook obtained by clustering the graphemes of the data. Figure 3-1 (d) shows an illustration of the used graphemes. K-

means and Kohonen self-organization feature maps (Kohonen, 1989) were used to generate the code book.

In their research work, Shomaker et al. addressed both text-dependent and text-independent approaches for writer identification. They have concluded that text-dependent approaches achieve high performance even with small amounts of data. However, this has limited applicability due to the need for specific text and human intervention (Bulacu & Schomaker, 2007a). It is worth adding that having a successful text-independent writer identification system can operate on dependent-texts without any major modifications to the system, and not vice versa.

Schlapbach et al. (Schlapbach, 2007; Schlapbach & Bunke, 2004a; 2004b; 2006; 2007; Schlapbach, Kilchherr, & Bunke, 2005) used features that are normally used for text recognition. In one of their research works (Schlapbach & Bunke, 2004a), they used Hidden Markov Models (HMM) for writer identification and verification by recognizing a text line. Using a number of HMMs, they determined the identity of the writer by choosing the HMM of the writer that provided the best confidence measure of the recognized text line. As each HMM was trained with the data of one writer, the HMM that produced higher confidence measure for the text line identified the writer.

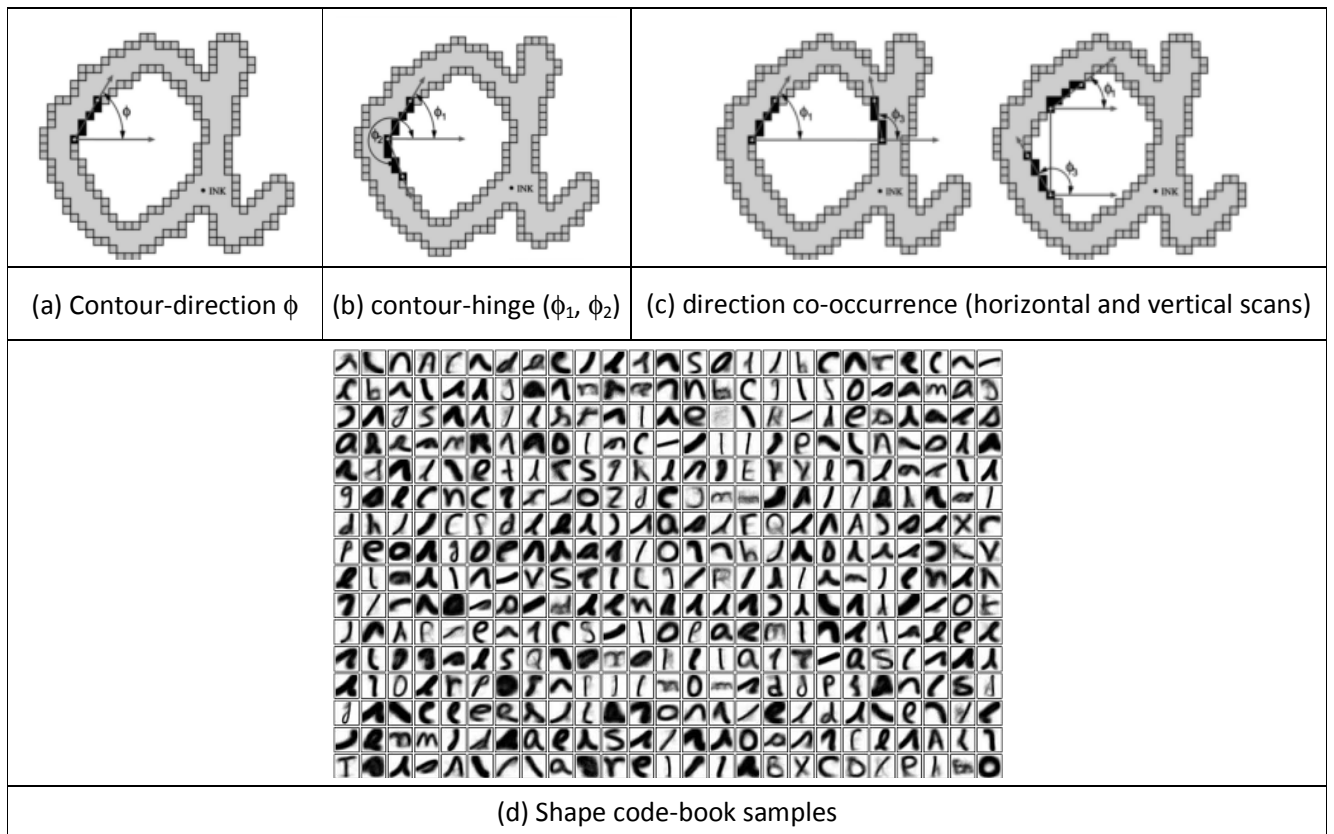


Figure 3-1: Texture and allograph features (Bulacu & Schomaker, 2007a).

For feature extraction, Schlapbach et al. used a sliding window which is common practice with HMM classifiers. A window of one pixel wide is shifted from left to right over the line of text. At each position, nine geometrical features are extracted; three global features and six local features. Global features represent the number of black pixels in the window, the center of gravity and the second order moment of the black pixels. The remaining six local features are the position and contour direction of the upper and lower-most pixels, the number of black-to-white transitions in the window, and the fraction of pixels between the upper and lower-most black pixels.

Srihari et al. (K. D., 2007; S. Srihari, 2000; S. Srihari & G. Ball, 2008; 2009; S. Srihari, Beal, Bandi, Shah, & Krishnamurthy, 2005; S. Srihari, S. H. Cha, Arora, & Lee, 2002; S. Srihari, Huang, Srinivasan, & Shah, 2007; Tomai & S. Srihari, 2004; B. Zhang, 2003; B. Zhang, S. Srihari, & Lee, 2003) used statistical features that are extracted at different levels of resolution. At the macro level, thirteen global features are extracted, viz. measures of pen pressure (entropy of gray values, gray-value threshold, number of black pixels); measures of writing movement (number of interior contours, number of exterior curves); measures of stroke formation (number of vertical, horizontal, positive, and negative strokes); average line height and average slant per line; stroke width, and average word gap.

At the micro level, Gradient, Structural and Concavity features (GSC) are extracted. First, the image is divided $n \times m$ grids with equal number of foreground pixels for each of n rows, and equal number of foreground pixels for each of m columns. Then for each grid cell, the GSC features column vector is extracted. The gradient features are computed by convolving two 3×3 Sobel operators with the binary image. These

operators approximate the x and y derivatives in the image at a pixel position. The vector addition of the operators' output is used to compute the gradient of the image. Although the gradient is a vector with magnitude and direction, only the direction is used in the computation of a feature vector which is stored in a gradient feature map. A histogram of gradient directions is taken at each pixel of the region, where each histogram value corresponds to the count of each gradient direction in the region.

The structural features capture certain patterns embedded in the gradient direction map. These patterns are “mini-strokes” of the image. A set of 12 rules are applied to each pixel. These rules operate on the eight nearest neighbours of the pixel. Each rule examines a particular pattern of the neighbouring pixels for allowed gradient ranges. For example, rule S_1 states that if neighbour (N_0) and neighbour (N_4) of a pixel both have a gradient range of $61^\circ \sim 150^\circ$, then the rule is satisfied and its corresponding value in the feature vector is incremented by 1. The concavity features are the coarsest of the GSC set. They can be broken down into three subclasses of features: segment density, large strokes, and concavity shape. The full list of rules for the GSC features is shown in Table 3-1. Figure 3-2 shows an example of the GSC features vector for the word “Medical” for 4×8 grid divisions.

Table 3-1: GSC feature definitions.

Gradient		Structural				Concavity	
ID	Angle	ID	Description	Neighbour 1 (Range)	Neighbour 2 (Range)	ID	Description
G ₁	1° ~ 30°	S ₁	Horizontal line (a)	N ₀ (61° ~ 150°)	N ₄ (61° ~ 150°)	CD	Pixel density
G ₂	31° ~ 60°	S ₂	Horizontal line (b)	N ₀ (241° ~ 330°)	N ₄ (241° ~ 330°)	CHRL	Horizontal run length
G ₃	61° ~ 90°	S ₃	Vertical line (a)	N ₂ (151° ~ 240°)	N ₆ (151° ~ 240°)	CVRL	Vertical run length
G ₄	91° ~ 120°	S ₄	Vertical line (b)	N ₂ (-29° ~ 60°)	N ₆ (-29° ~ 60°)	CCH	Hole concavity
G ₅	121° ~ 150°	S ₅	Diagonal rising (a)	N ₅ (121° ~ 210°)	N ₁ (121° ~ 210°)	CCU	Upward concavity
G ₆	151° ~ 180°	S ₆	Diagonal rising (b)	N ₅ (-59° ~ 30°)	N ₁ (-59° ~ 30°)	CCD	Downward concavity
G ₇	181° ~ 210°	S ₇	Diagonal falling (a)	N ₃ (31° ~ 120°)	N ₇ (31° ~ 120°)	CCR	Right concavity
G ₈	211° ~ 240°	S ₈	Diagonal falling (b)	N ₃ (211° ~ 300°)	N ₇ (211° ~ 300°)	CCL	Left concavity
G ₉	241° ~ 270°	S ₉	Comer (a)	N ₂ (151° ~ 240°)	N ₀ (241° ~ 330°)		
G ₁₀	271° ~ 300°	S ₁₀	Comer (b)	N ₆ (151° ~ 240°)	N ₀ (61° ~ 150°)		
G ₁₁	301° ~ 330°	S ₁₁	Comer (c)	N ₄ (241° ~ 330°)	N ₂ (-29° ~ 60°)		
G ₁₂	331° ~ 360°	S ₁₂	Comer (d)	N ₆ (-29° ~ 60°)	N ₄ (61° ~ 150°)		

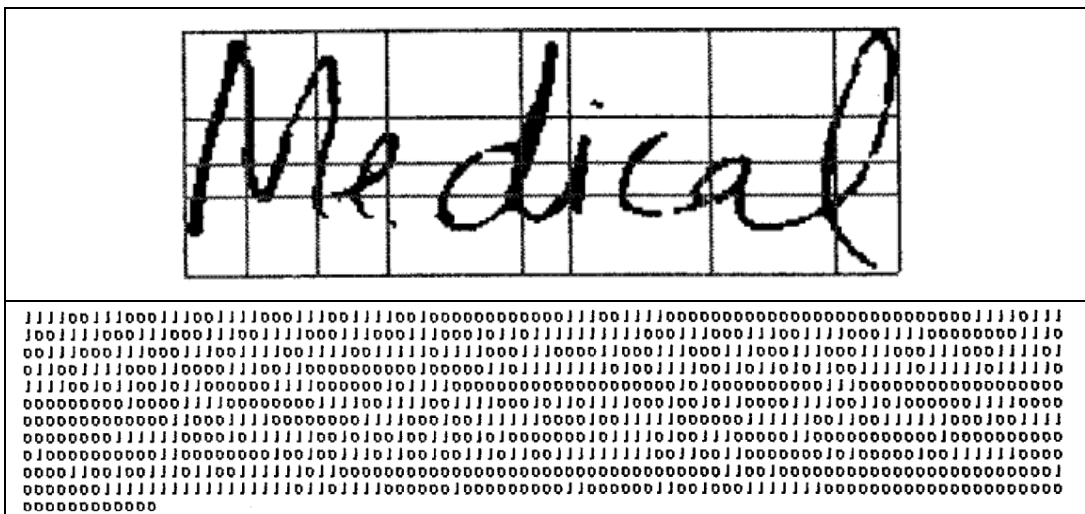


Figure 3-2: Exemplar word image with 4 x 8 divisions using GSC (B. Zhang, 2003).

Siddiqi et al. (Siddiqi & Vincent, 2007; 2008; 2009) divided each image into a large number of small sub-images using a window, and clustered these sub-images. They used these clusters as features. They also extracted the histograms of the chain code, the first and second order differential chain codes, and the histogram of the curvature indices at each point of the contour of handwriting. Leedham et al. (Leedham & Chachra, 2003) used a combination of local and global features. These included pixel density, fixed point distance and angular measure, center of gravity, gradient features, height to width ratio, number of end-points, number of junctions, number of loops, and degree of slant.

Ram et al. (Ram & Moghaddam, 2009a; 2009b) used gradient features, grapheme features, connected components contours, area features, and a collection of local features. Said et al. (Said, Baker, & T. Tan, 1998) used grey scale co-occurrence matrices. Franke et al. (Katrin Franke et al., 2002) used co-occurrence features like energy, correlation, inverse difference moment, and entropy. Bar-Yosef (Bar-Yosef et al., 2007) used the ratio between the area of each dominant background set and the convex hull, and the aspect ratio of the enclosing ellipse. Mar et al. (Mar & Thein, 2005) used mean and standard deviation of Region Of Interests (ROIs). Cha et al. (S. H. Cha, 2001) used sliding windows to extract both local and global features. Wang et al. (X. Wang, Ding, & H. Liu, 2003) used distribution of directional elements (gradient). Liu et al. (C.-L. Liu, Dai, & Y.-J. Liu, 1995) used features derived from 2nd and 3rd order moments. Zois et al. (Zois & Anastassopoulos, 2000) used erosion and dilation function on the horizontal projection.

Researchers have also used image transformations as features. For example, Gabor filters were used in (Al-Dmour & Zitar, 2007; Cong et al., 2002; He & Tang,

2004; Helli & Moghaddam, 2008a; 2008b; 2009; 2010; C.-L. Liu, Dai, & Y.-J. Liu, 1995; Said, Baker, & T. Tan, 1998; Shahabi & Rahmati, 2006; 2007; Siddiqi & Vincent, 2008; Ubul et al., 2008; Zhu, T. Tan, & Y. Wang, 2000), wavelet transforms in (Gazzah & Ben Amara, 2006; 2007; 2008; He, Bin, Jianwei, Yuan Yan, & Xinge, 2005; He, X. You, & Tang, 2008a; 2008b), and contourlet transformations in (He, Tang, & X. You, 2005).

It is worth noting that some of the same successful feature extraction techniques have been used by different research groups. For example, taking the histogram of the pixel angle was originally applied for writer identification by both Srihari (S. Srihari, 2000) and Schomaker et al. (Schomaker, Bulacu, & M. van Erp, 2003), and since then was used by their own research groups as shown previously and by other researchers (Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008; Leedham & Chachra, 2003; X. Li & Ding, 2009; Ram & Moghaddam, 2009a; 2009b; X. Wang, Ding, & H. Liu, 2003). Measuring slant (at least at the pixel level) using gradient distributions has been widely assumed to be an important writer-specific feature, although there have been experimental results that question the effect of slant on writer identification/verification (Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010). Using parts of letters (graphemes) was originally applied by Benesefia et al. (A. Benesefia et al., 2002) in 2002, and since then has been implemented by several other researchers as well (Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Bulacu, 2007; Bulacu & Schomaker, 2006; Leedham & Chachra, 2003; Ram & Moghaddam, 2009a; Schomaker & Bulacu, 2004; Schomaker, Bulacu, & K. Franke, 2004).

Table 3-2 details the published work of writer identification and verification including used features, classifiers and best reported top-1 accuracy results. The top-1 accuracy results in writer identification are obtained by summing the number of hits for the most probable writer for each test sample. Correspondingly, the top-5 and top-10 accuracy results are the number of hits using the top-5 and top-10 most probable writer for each test sample. Some researchers tried their writer identification system on multiple databases, and hence more than one accuracy result is reported per publication. For more information about the used databases, readers are referred to Table 2-1.

Table 3-2: Writer identification/verification features and classifiers.

Citation	Features	Classifier	DB	#Wr	Info	Dep./Ind.	Top-1%
(Gazzah & Ben Amara, 2006)	Entropy as global features, Wavelet transforms, and a set of structural features.	Neural networks.	DB01	60	3 docs/wr	Dep.	94.73%
(Gazzah & Ben Amara, 2007)	See (Gazzah & Ben Amara 2006).	Neural networks.	DB01	60	3 docs/wr	Dep.	95.68%
(Gazzah & Ben Amara, 2008)	See (Gazzah & Ben Amara 2006).	SVM, neural networks.	DB01	60	3 docs/wr	Dep.	94.00%
(Al-Dmour & Zitar, 2007)	Gabor filters.	Weighted Euclidean, SVM, LDC.	DB02	20	na*	na*	90.00%
(Abdi, Khemakhem, & Ben-Abdallah, 2009)	Length, height/width ratio, and curvature of strokes.	Euclidean, Square, Manhattan, X^2 , Chebechev, Hamming, Minkowski, and Mahalanobis distance.	DB03	40	> 100 words/wr	Ind.	92.50%
(Lutf, Xinge You, & H. Li, 2010)	Diacritics local histograms	Chi-squared (X^2)	DB03	287	> 200 diacritics	Ind	97.56%
(Chawki & Labiba, 2010)	Grey Level Co-occurrence Matrices	Euclidean	DB03	130	5 docs/wr	Ind.	82.62%
(Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010)	Fractals and multi-fractals	k-Nearest Neighbor	DB03	50	24 city names, 12 times each	Dep.	90.00%
(Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008)	Edge-hinge features. Grapheme features.	Euclidean distance.	DB04	10	2 docs/wr	Dep.	90.00%
(Al-Ma'adeed, Mohammed, & Al Kassis, 2008)	Edge-direction distribution. Moment Invariants, Area, length, Height, Length from Baseline to Upper Edge, Baseline to the Lower Edge.	Euclidean distance.	DB04	100	20 docs/wr	Dep.	93.80%
(C.-L. Liu, Dai, & Y.-J. Liu, 1995)	Gabor filters. Features from 2nd and 3rd order moments.	Manhattan distance.	DB06	20	7 docs/wr	Ind.	100.0%
(Zhu, T. Tan, & Y. Wang, 2000)	Gabor filters.	Weighted Euclidean.	DB07	17	1 doc/wr	Ind.	95.70%
(Cong et al., 2002)	Gabor filters.	Euclidean distance.	DB08	50	110 scripts	Ind.	97.60%
(He & Tang, 2004)	Gabor filters.	Weighted Euclidean.	DB09	50	2 docs/wr	Both	90.00%
(He, Tang, & X. You, 2005)	Contourlet transforms.	Kullback-Leibler Distance	DB10	10	2 docs/wr	Ind.	90.00%
(He, Bin, Jianwei, Yuan Yan, & Xinge, 2005)	Wavelet transforms.	Kullback-Leibler Distance	DB10	10	2 docs/wr	Ind.	80.00%
(He, X. You, & Tang, 2008a)	Wavelet transforms.	Hidden Markov Tree model.	DB12	500	2 docs/wr	Ind.	36.40%
(He, X. You, & Tang, 2008b)	Wavelet transforms.	Kullback-Leibler distance.	DB12	500	2 docs/wr	Ind.	39.20%
(X. Li & Ding, 2009)	Histogram of contour-hinge.	Weighted Euclidean, and modified X^2 distance measure	DB11	240	1 doc/wr	Ind.	95.00%
(X. Wang, Ding, & H. Liu, 2003)	Distribution of directional elements (gradient).	Euclidean distance.	DB13	25 626	16*34 char/wr 20 char/wr	Dep. Dep.	96.12% 82.16%
(Bulacu, Schomaker, & Vuurpijl, 2003)	Edge-direction distribution, edge-hinge distribution, run-length distributions, autocorrelation, and entropy.	Euclidean distance.	DB14	250	2 docs/wr	Ind.	75.00%

(Schomaker, Bulacu, & M. van Erp, 2003)	See (Bulacu, Schomaker, & Vuurpijl, 2003).	X^2 , Hamming, Minkowski, Bhattacharyya, and Hausdorff distance.	DB14	251	2 docs/wr	Ind.	88.00%
(Schomaker & Bulacu, 2004)	See (Schomaker, Bulacu, & M. van Erp, 2003). Grapheme emission PDFs.	X^2 and Hamming distance.	DB14	150	1 doc/wr	Dep.	87.00%
(Schomaker, Bulacu, & K. Franke, 2004)	See (Schomaker & Bulacu, 2004).	Euclidean distance.	DB14	150	1 doc/wr	Dep.	97.00%
(Bulacu & Schomaker, 2005)	See (Schomaker & Bulacu, 2004).	Euclidean distance.	DB14 DB14 DB15	250 250 150	2 docs/wr 1 doc/wr 2 docs/wr	Ind. Dep. Ind.	78.10% 64.90% 76.30%
(Bulacu & Schomaker, 2006)	See (Schomaker & Bulacu, 2004).	X^2 and Hamming distance.	DB14, 15,16	900	2 docs/wr	Ind.	87.00%
(Bulacu, 2007)	See (Schomaker & Bulacu, 2004).	X^2 and Hamming distance.	DB14, 15,16	900	2 docs/wr	Ind.	87.00%
(Bulacu & Schomaker, 2007b)	See (Schomaker & Bulacu, 2004).	X^2 and Hamming distance.	DB17	10	2 regions/wr	Ind.	89.00%
(Bulacu & Schomaker, 2007a)	See (Schomaker & Bulacu, 2004).	X^2 and Hamming distance.	DB14, 15,16	900	2 docs/wr	Ind.	87.00%
(Schomaker, Katrin Franke, & Bulacu, 2007)	See (Schomaker & Bulacu, 2004). Writer information: handedness, sex, age, and style.	X^2 distance.	DB14	150	1 doc/wr	Dep.	80.00%
(Bulacu, Schomaker, & Brink, 2007)	See (Schomaker & Bulacu, 2004).	X^2 and Hamming distance.	DB03	350	5 docs/wr	Ind.	88.00%
(Brink, Bulacu, & Schomaker, 2008)	See (Schomaker & Bulacu, 2004).	na*	DB14, 16	498	2 docs/wr	Ind.	Varies
(Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010)	See (Schomaker & Bulacu, 2004).	X^2 distance.	DB18	47	4 docs/wr	Dep.	97.00-100%
(Schlapbach & Bunke, 2004a)	Sliding window.	Hidden Markov Models (HMM).	DB16	50	5 docs/wr	Ind.	94.23%
(Schlapbach & Bunke, 2004b)	See (Schlapbach & Bunke 2004a).	Hidden Markov Models (HMM).	DB16	100	5 docs/wr	Ind.	96.56%
(Schlapbach, Kilchherr, & Bunke, 2005)	100 simple features: slant, skew angle, fractal features...	Euclidean distance	DB16	50	5 docs/wr	Ind.	98.36%
(Schlapbach & Bunke, 2006)	See (Schlapbach & Bunke 2004a).	HMM, Gaussian Mixture Models.	DB16	100	5 docs/wr	Ind.	98.46%
(Schlapbach, 2007)	See (Schlapbach & Bunke 2004a).	HMM, Gaussian Mixture Models.	DB16	100	5 docs/wr	Ind.	97.03%
(Schlapbach & Bunke, 2007)	See (Schlapbach & Bunke 2004a).	HMM, Gaussian Mixture Models.	DB16	100	5 docs/wr	Ind.	97.03%
(S. Srihari et al. 2002)	Gradient, structural, and concavity histograms. Eleven macro features.	Euclidean distance. Correlation measure.	DB20	1500	3 docs/wr	Dep.	98.00%
(B. Zhang, 2003)	See (S. Srihari et al. 2002).	Euclidean distance. Correlation measure.	DB20	1000	3 docs/wr	Dep.	98.06%
(Tomai & S. Srihari, 2004)	See (S. Srihari et al. 2002).	Manhattan and Correlation measure.	DB20	1000	3 docs/wr	Dep.	99.00%
(S. Srihari, Huang, Srinivasan, & Shah, 2007)	See (S. Srihari et al. 2002).	Manhattan and Correlation measure.	DB20	1000	3 docs/wr	Dep.	96.10%
(S. Srihari & G. Ball, 2008)	See (S. Srihari et al. 2002).	Manhattan and Correlation measure. Log-likelihood ratio.	DB05	10	10 docs/wr	Ind.	99.30%

(S. Srihari & G. Ball, 2009)	See (S. Srihari et al. 2002).	Log-likelihood ratio.	DB20	1000	3 docs/wr	Dep.	na*
(Matsuura & Qiao, 1989)	Impulse response of image	Euclidean distance.	DB21	2	5 words/wr	Dep.	100.0%
(Said, Baker, & T. Tan, 1998)	Gabor filters. Grey Scale Co-occurrence Matrices.	Weighted Euclidean.	DB22	20	25 block/wr	Ind.	95.30%
(Leedham & Chachra, 2003)	Pixel density, fixed point distance and angular measure, center of gravity, gradient features, connected components contours, and a collection of local features.	Hamming distance.	DB23	15	10 strings/wr	Ind.	100.0%
(Zois & Anastassopoulos, 2000)	Erosion and dilation function.	Linear Bayes classifier. Neural networks.	DB24	50	90 words/wr	Dep.	> 95.0%
(G. Tan et al., 2008)	x and y co-ordinates, the directions of x and y co-ordinates, the curvatures of x and y co-ordinates and the Pen-up or Pen-down information.	Fuzzy classifiers.	DB25	120	Characters, online	na*	98.30%
(A. Bensefia et al., 2002)	Grapheme clustering.	Correlation similarity measure.	DB26	88	1 doc/wr	Dep.	97.70%
(A. Bensefia, T. Paquet, & L. Heutte, 2003a)	Grapheme clustering.	Cosine similarity measure.	DB26	88	1 doc/wr	Dep.	97.70%
(Siddiqi & Vincent, 2007)	Modified sliding window.	Bayesian classifier.	DB16	50	2 docs/wr	Ind.	94.00%
(Siddiqi & Vincent, 2008)	Gabor filters.	Mahalanobis distance.	DB16	100	2 docs/wr	Ind.	92.00%
(Siddiqi & Vincent, 2009)	Chain code histograms.	Euclidean, X ² , Hamming, and Bhattacharyya distance.	DB16 DB27	650 225	2 docs/wr 2 docs/wr	Ind. Ind.	86.00% 79.00%
(Bar-Yosef et al., 2007)	The ratio between the area of the background and the convex hull. The aspect ratio of the enclosing ellipse. Concavity features. Ellipse aspect ratio. Moment features.	Euclidean distance and Linear Bayes classifier.	DB28	34	20 characters/wr	Dep.	100.0%
(Mar & Thein, 2005)	Mean and standard deviation of ROIs	Weighted Euclidean.	DB29	20	2 docs/wr	Dep.	97.50%
(Shahabi & Rahmati, 2006)	Gabor filters.	Weighted Euclidean. X ² distance.	DB30	25	4 blocks/wr	Dep.	88.00%
(Shahabi & Rahmati, 2007)	Gabor filters.	Euclidean and X ² distance.	DB30	40	3 docs/wr	Dep.	82.50%
(Helli & Moghaddam, 2008b)	Gabor filters.	Longest Common Subsequence.	DB31	100	5 docs/wr	Ind.	95.00%
(Helli & Moghaddam, 2008a)	Gabor filters.	Weighted Euclidian distance.	DB31	70	5 docs/wr	Ind.	77.00%
(Helli & Moghaddam, 2009)	Gabor filters.	Longest Common Subsequence.	DB31 DB16	100 30	5 docs/wr 7 docs/wr	Ind. Ind.	89.00% 94.40%
(Helli & Moghaddam, 2010)	Gabor filters.	Graph similarity.	DB31	100	5 docs/wr	Ind.	98.00%
(Ram & Moghaddam, 2009b)	Gradient features.	Neural networks.	DB32	50	5 docs/wr	Ind.	94.00%
(Ram & Moghaddam, 2009a)	Grapheme features. Gradient features. Used area features.	Fuzzy classifiers.	DB32	50	5 docs/wr	Ind.	90.00%
(Ubul et al., 2008)	Gabor filters.	Euclidean distance, weighted Euclidean, and SVM.	DB33	23	2 docs/wr	Dep.	88.00%

na*: Information is not available.

3.4 Classification Approaches

The research of writer identification and verification used different classifier approaches. Friedman et al. (Friedman & Kandel, 1999) categorize classifier types into five kinds; minimum distance classifiers, statistical classifiers, neural networks, fuzzy classifiers, and syntactic classifiers. Using this categorization, this section addresses the classifier types used in writer identification and verification.

3.4.1 Minimum Distance Classifiers

Minimum distance classifiers classify a new pattern by measuring its distance from the test sample to the training patterns and choosing the K-nearest classes to which the nearest neighbors belong (Friedman & Kandel, 1999). Various distance measures have been attempted; with the Euclidean distance measure remaining the most commonly used distance measure for writer identification and verification. Researchers who used the Euclidean distance measure include (Abdi, Khemakhem, & Ben-Abdallah, 2009; Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008; Bar-Yosef et al., 2007; Bulacu & Schomaker, 2005; Bulacu, Schomaker, & Vuurpijl, 2003; Cong et al., 2002; Matsuura & Qiao, 1989; Siddiqi & Vincent, 2009; S. Srihari, S. H. Cha, Arora, & Lee, 2002; Ubul et al., 2008; X. Wang, Ding, & H. Liu, 2003; B. Zhang, 2003). By adding weights to each feature value, researchers also used the weighted Euclidean distance measure (Al-Dmour & Zitar, 2007; He & Tang, 2004; X. Li & Ding, 2009; Mar & Thein, 2005; Said, Baker, & T. Tan, 1998; Shahabi & Rahmati, 2006; 2007; Ubul et al., 2008; Zhu, T. Tan, & Y. Wang, 2000).

Other used distance measures for writer identification/verification include:

- Square Euclidean distance (Abdi, Khemakhem, & Ben-Abdallah, 2009).
- Manhattan - a.k.a city block - distance measure (Abdi, Khemakhem, & Ben-Abdallah, 2009; K. D., 2007; C.-L. Liu, Dai, & Y.-J. Liu, 1995; S. Srihari & G. Ball, 2008; 2009; S. Srihari, Huang, Srinivasan, & Shah, 2007; Tomai & S. Srihari, 2004).
- Chi-squared (X^2) distance measure (Abdi, Khemakhem, & Ben-Abdallah, 2009; Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010; Bulacu, 2007; Bulacu, van Koert, Schomaker, & van Der Zant, 2007; Bulacu & Schomaker, 2006; 2007a; 2007b; Bulacu, Schomaker, & Brink, 2007; Schomaker & Bulacu, 2004; Schomaker, Bulacu, & M. van Erp, 2003; Shahabi & Rahmati, 2006; 2007; Siddiqi & Vincent, 2009).
- Li et al. (X. Li & Ding, 2009) used a modified version of the X^2 distance measure.
- Chebechev distance measure (Abdi, Khemakhem, & Ben-Abdallah, 2009).
- Hamming distance measure (Abdi, Khemakhem, & Ben-Abdallah, 2009; Bulacu, 2007; Bulacu, van Koert, Schomaker, & van Der Zant, 2007; Bulacu & Schomaker, 2006; 2007a; 2007b; Bulacu, Schomaker, & Brink, 2007; Leedham & Chachra, 2003; Schomaker & Bulacu, 2004; Schomaker, Bulacu, & M. van Erp, 2003; Schomaker, Bulacu, & K. Franke, 2004; Siddiqi & Vincent, 2009).

- Minkowski (Abdi, Khemakhem, & Ben-Abdallah, 2009; Schomaker, Bulacu, & M. van Erp, 2003; Tomai & S. Srihari, 2004), the Mahalanobis distance measure (Abdi, Khemakhem, & Ben-Abdallah, 2009; Siddiqi & Vincent, 2008).
- Correlation measure (A. Bensefia et al., 2002; A. Bensefia, T. Paquet, & L. Heutte, 2003a; K. D., 2007; S. Srihari & G. Ball, 2008; 2009; S. Srihari, Beal, Bandi, Shah, & Krishnamurthy, 2005; S. Srihari, Huang, Srinivasan, & Shah, 2007; Tomai & S. Srihari, 2004; B. Zhang, 2003; B. Zhang, S. Srihari, & Lee, 2003).
- Bhattachalyya distance (Schomaker, Bulacu, & M. van Erp, 2003; Siddiqi & Vincent, 2009).
- The Hausdorff distance (Schomaker, Bulacu, & M. van Erp, 2003).
- The Longest Common Subsequence (LCS) algorithm (Helli & Moghaddam, 2008b; 2009).

Since the performance of distance measures heavily rely on the nature of the used features, it is often hard to conclude the best distance measure for writer identification/verification. Nevertheless, many researchers have reported that the X^2 distance measure reported highest accuracy for their features when compared with other distance measures (Brink, Bulacu, & Schomaker, 2008; Brink, Niels, van Batenburg, van Den Heuvel, & Schomaker, 2010; Bulacu, 2007; Bulacu & Schomaker, 2005; 2006; 2007b; Bulacu, Schomaker, & Brink, 2007; Bulacu, Schomaker, & Vuurpijl, 2003; Katrin Franke et al., 2003; Niels, Vuurpijl, & Schomaker, 2007; Schomaker & Bulacu,

2004; Schomaker, Bulacu, & M. van Erp, 2003; Schomaker, Bulacu, & K. Franke, 2004; Schomaker, Katrin Franke, & Bulacu, 2007). In addition, Srihari et al. used binary feature vectors for writer identification and verification, and hence rely on (dis)similarity computation for classification (K. D., 2007; S. Srihari, 2000; S. Srihari & G. Ball, 2008; 2009; S. Srihari, Beal, Bandi, Shah, & Krishnamurthy, 2005; S. Srihari, S. H. Cha, Arora, & Lee, 2002; S. Srihari, Huang, Srinivasan, & Shah, 2007; Tomai & S. Srihari, 2004; B. Zhang, 2003; B. Zhang, S. Srihari, & Lee, 2003). They conducted various experiments to select the best performing (dis)similarity measure and concluded that the correlation distance measure provided the best results (B. Zhang, 2003).

3.4.2 Statistical Classifiers

Minimum distance classifiers are based on the assumption that training samples form distinct clusters. However, this is not usually the case. Training samples of various classes overlap, and in this case a statistical approach is more appropriate assuming that the samples come from statistical distribution (Friedman & Kandel, 1999). Examples of statistical classifiers used in writer identification and verification include Linear Bayes classifier (Bar-Yosef et al., 2007; Zois & Anastassopoulos, 2000), Support Vector Machines (SVM) (Al-Dmour & Zitar, 2007; Katrin Franke et al., 2002; Gazzah & Ben Amara, 2008; Ubul et al., 2008), Hidden Markov Models (HMM) (Schlapbach & Bunke, 2004b; 2007), Hidden Markov Tree (HMT) model (He, X. You, & Tang, 2008a), Gaussian Mixture Models (GMM) (Schlapbach, 2007), Kull back Leibler distance (KLD) between two PDFs (He, X. You, & Tang, 2008b), Cumulative Distribution Functions of the log-likelihood ratio (LLR) of the same and different writers (S. Srihari & G. Ball,

2008; S. Srihari, Beal, Bandi, Shah, & Krishnamurthy, 2005), and the linear discriminant classifier (LDC) (Al-Dmour & Zitar, 2007).

3.4.3 Other Classifiers

Researchers have also used neural networks (Gazzah & Ben Amara, 2006; 2007; 2008; Ram & Moghaddam, 2009b; Zois & Anastassopoulos, 2000), fuzzy classifiers (Ram & Moghaddam, 2009a; G. Tan et al., 2008). Structural classifiers are used less frequently and with less significant accuracy results (Helli & Moghaddam, 2010).

3.5 Writer Identification and Verification of Arabic Text

In this section we present a survey of research of writer identification and verification of Arabic text. It is to be noted that most of the efforts of writer identification and verification of Arabic text are based on the techniques that were used for English text. Most of the features and classifiers were previously used for writer identification of English text. Since Persian (Farsi) text is similar to Arabic, research of writer identification and verification of Persian text will also be presented.

Researchers in (Gazzah & Ben Amara, 2006) used a combination of global and structural features (Average line height, Spaces between sub-words, inclination of the ascender, height and the width of each diacritic dot) along with a multilayer perceptron (MLP) classifier. They reported an accuracy of 94.73% for 60 writers. (Gazzah & Ben Amara, 2007) used a 2D discrete Wavelet Transforms for feature extraction along with the MLP classifier with a reported accuracy of 95.68% on the same database. In their

latest reported work Support Vector Machines (SVM) classifier was used where they showed that MLP provided slightly better results than SVM (Gazzah & Ben Amara, 2008).

Bulacu et al. (Bulacu, Schomaker, & Brink, 2007) used the IFN/ENIT dataset (Pechwitz, Maddouri, Märgner, Ellouze, & Amiri, 2002) which is limited to Arabic town and city names. For tests involving 350 writers, they reported a best accuracy of 88%. They concluded that the identification and verification results obtained on Arabic text cannot be numerically compared with previous results for Western script because the experimental datasets are different (in terms of the amount of ink contained in the samples among others). They also indicated that the results obtained on Arabic text are generally lower than the ones obtained on Western script. Abdi et. al. (Abdi, Khemakhem, & Ben-Abdallah, 2009) used the IFN/ENIT dataset (Pechwitz et al., 2002), but with only 40 writers. Using statistical features (the length, height/width ratio, and the curvature of the strokes to calculate various probability distribution function (PDF) feature vectors) along with Euclidean, Manhattan, and Mahalanobis distance measures and the Borda count ranking algorithm, they reported a top-1 accuracy of 92.5%.

Chawki and Labiba (Chawki & Labiba, 2010) used 650 handwriting documents collected from 130 different Arabic writers from the IFN/ENIT dataset. Using Grey Level Co-occurrence Matrices (GLCM) as features and the Euclidean distance measure for classification, they reported a top-1 accuracy of 82.62%. Lutf et al. (Lutf, Xinge You, & H. Li, 2010) used the Chi-squared (χ^2) distance measure for classification, and histograms for extracted diacritics as features. Using diacritics samples from 287 writers in the IFN/ENIT database, they first reported writer identification top-1 accuracy of

51.22%, and later by calculating only the white and black pixels on the edge of the diacritics, they reported an almost perfect top-1 accuracy of 97.56%.

Chaabouni et al. (Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010) also used Tunisian city names for their dataset with 50 writers where each writer wrote 24 city names repeated 12 times. Using k-Nearest Neighbor for classifier and the “box-counting” method to calculate the images multi-fractal dimension, they reported 90% top-5 accuracy for some words.

Al-Dmour and Zitar (Al-Dmour & Zitar, 2007) presented a technique for feature extraction based on hybrid spectral-statistical measures (SSMs) of texture. Correct identification of 90% was reported using Arabic handwriting samples from 20 different writers. Al-Ma’adeed et al. (Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008) used edge-based statistical features for writer identification using Arabic handwritten words. They used their own generated database as described in Chapter 2. Some of the phrases are reported to score Top-10 result of more than 90% accuracy, whereas shorter words scored around 50% accuracy for 100 writers. Srihari and Ball (S. Srihari & G. Ball, 2008) used a dataset of 10 different writers, each contributing 10 different full page documents in handwritten Arabic for a total of 100 pages. Using macro- and micro-features along with likelihood ratio computation, they reported 86% accuracy.

Persian, a.k.a Farsi, handwriting is very similar to Arabic in terms of strokes and structure. Therefore, a Persian writer identification system can also be used for identification of Arabic text. Farsi character set comprises all of the 28 Arabic characters plus four additional ones, shown in Figure 3-3. Similar to Arabic, Persian writer

identification and verification has been increasingly popular lately. Shahabi et al. (Shahabi & Rahmati, 2006; 2007) used features based on Gabor filters for feature extraction, and different distance measures (Euclidean, Weighted Euclidean, and X^2 distance) for classifiers. Their latest work reported a top-1 accuracy of 82.50% for 40 writers. Ram et al. (Ram & Moghaddam, 2009a; 2009b) used gradient and grapheme features and tested them on a database of 50 writers (5 pages/writer) and reported top-1 accuracy of 94.0%.



Figure 3-3: The four additional Farsi isolated characters.

Helli et al. (Helli & Moghaddam, 2008a; 2008b; 2009; 2010) used modified Gabor filters and tried different classification techniques for identification. They used a database of 100 writers, 5 pages per writers. The volunteer was free to write anything in the pages, and hence their approach was text independent. They reported a top-1 accuracy of 98% for all 100 writers. Quite interestingly, they tried their system on the IAM database (Marti & Bunke, 2002) for 30 writers (7 pages/writer) and reported a top-1 accuracy of 94.4%. Since the databases are different, their results cannot be compared. Therefore no conclusion can be drawn based on Latin/Farsi text although the general understanding is that Latin text gives better identification rates. We think that the used data for Arabic text writer identification does not match in representation and naturalness the databases of Latin text as Latin databases usually contain more samples and writers than Arabic databases, and writing in Latin databases is usually less restricted than their Arabic counterparts.

3.6 Conclusions

In this chapter we presented the state of the art in writer identification and verification of Latin and western texts, the feature extraction approaches, and the classifier approaches. The state of the art was grouped by addressing the research work publications of different research groups due to similarities in used features and classifiers. This grouping helps in showing each group's own improvement over time and their performance compared to other groups. Published research work was tabulated indicating for every publication used features, classifiers, databases used, number of

writers, best identification rates achieved, and the year of publication. This makes it easier to compare the research work reported results from different researchers.

The chapter presented a survey of writer identification and verification of Arabic text. Comparing the work on Arabic text with Latin indicates that limited number of researchers is involved in writer identification of Arabic text. In addition, comparing features and classification approaches indicates that most of the work on Arabic text is based on features and classifiers used for English. In the coming chapter, the author will design features that better exploits some of the characteristics of the Arabic text.

It is clear that the published work related to Arabic text has lower accuracy than Latin. We cannot conclude that Arabic text is less identifiable than Latin text although the general understanding is that Latin text gives better identification rates. We think that the data used for Arabic text writer identification does not match in representation and naturalness that of the databases of Latin text. The used databases are individual efforts with inherent limitations in size and comprehensive. The IEF/ENIT which consists of city names in which researchers had to concatenate a number of city names to make an Arabic text. This is neither a good representation of Arabic text nor comprehensive.

CHAPTER 4

Feature Development for Writer Identification

In this chapter, several types of structural and statistical features are extracted from Arabic handwriting digits and text. Some features are classical features that are better modified for Arabic text and digits, like overlapped gradient distribution features, gradient distribution features, windowed gradient distribution features, contour chain code distribution features, windowed contour chain code distribution features, density features, horizontal and vertical run length features, stroke features, and concavity shape. Connected component features for Arabic handwritten text are original statistical and structural features that build on some of the main characteristics of the Arabic language. This chapter provides a detailed description of these features.

4.1 Introduction

The type of features used has a crucial effect on the accuracy of writer identification systems. Features are used to underline the distinctive properties of an object under consideration while at the same time reducing the size of the data. Forensic experts in writer identification and verification have long used a combination of qualitative and quantitative features while manually examining handwriting samples

(Huber & Headrick, 1999). Likewise, automatic writer identification and verification researchers like Schomaker et al. (Schomaker, Bulacu, & M. van Erp, 2003) have often concluded that one feature type will never suffice in automatic writer identification and that only a combination of feature types will yield reliable results in practice. Therefore, multiple features that pertain to both pixel and medium scale levels are extracted in this thesis.

Several types of features are extracted from the sample image (viz. connected component features, gradient distribution features, contour chain code distribution features, density features, horizontal and vertical run lengths features, stroke features, and concavity features) from the sample image. The gradient distribution features, density features, horizontal and vertical run length features, stroke features, and concavity features are classically known (Favata & Srikantan, 1996; Schomaker & Bulacu, 2004), and have been successfully used by the author in previous digit recognition application (Awaidah & S. A. Mahmoud, 2009; Sabri Mahmoud & Awaida, 2009). These features have been further tweaked and improved for the task of writer identification.

The overlapped gradient distribution features have been fine-tuned to better suit Arabic digits as shown in chapter 5, and an updated version of the gradient distribution and windowed gradient distribution features have been designed for handwritten Arabic text. The connected component features use a novel approach to extract a combination of structural and statistical features that build on some of the characteristics of the Arabic language.

In order to visualize the similarity for intra-writer feature vectors and the differences for inter-writer feature vectors, histograms for each feature type for the same

and different writers will be presented. These histograms along with the sum total of their absolute difference can indicate some of the differentiation powers of the implemented features. More analytical results on the discriminability power of the features will be demonstrated in Chapters 5 and 6. Using the text database developed (see Chapter 2), we chose three images from collected samples, two images for writer X and one image for writer Y as shown in Figure 4-1. Some features values are scaled up or down for a better histogram view. For writer identification, these feature vectors are normalized to have zero mean and unit variance before being presented to the classifier. The following sections provide a detailed description of these features.

4.2 Overlapped Gradient Distribution Features (OGDF)

The overlapped gradient distribution features are computed by convolving the x and y Sobel operators, shown in Figure 4-2, with the binary image. These operators approximate the x and y derivatives of the image at a pixel position, while giving the center points more importance with a weight of 2. The gradient of a center pixel is computed as a function of its eight nearest neighbors. The operators coefficients sum to 0 in an area of constant black level, indicating that the gradient degree of this area is zero (Gonzalez & Woods, 2007). The vector addition of the operators output are used to approximate the x and y gradients of the image. The gradient angle of each pixel is calculated by computing the inverse tangent of $(\text{Gradient}_y/\text{Gradient}_x)$. Subsequently, the histogram of the gradient angles is calculated and stored as the feature vector.

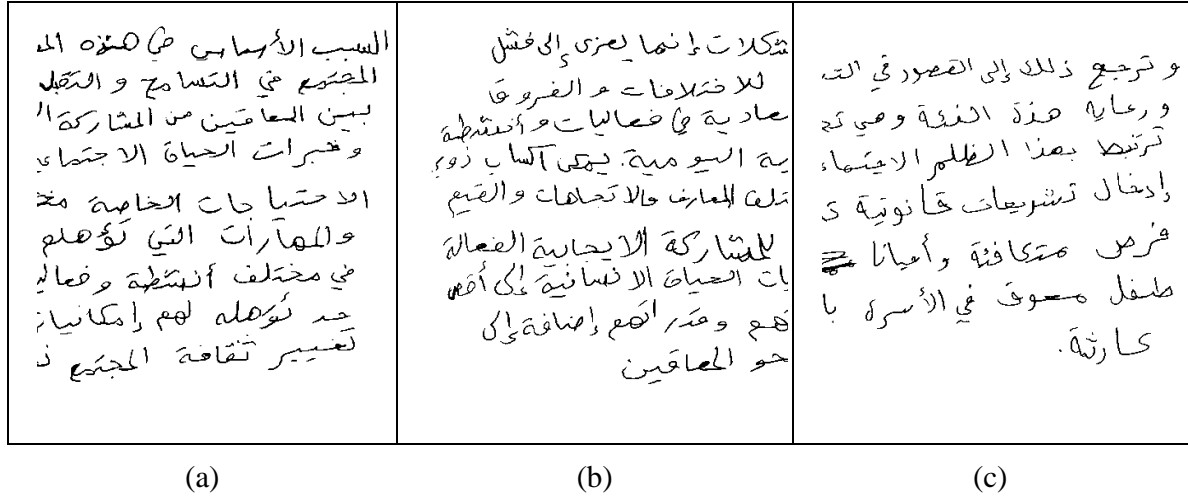


Figure 4-1: (a) and (b): Two sample images for writer X, (c): Sample image for writer Y.

-1	-2	-1	-1	0	1
0	0	0	-2	0	2
1	2	1	-1	0	1

Figure 4-2: X and Y Sobel operator masks.

The gradient angle (direction) can range from 0 to 2π radians. A sliding window of half a quadrant (45 degrees) is used to estimate the histogram of gradient directions of the pixels in the window. Each histogram value corresponds to the count of each gradient direction in the sliding window. The sliding window overlaps with the previous window by $1/3$ of the window range (i.e. 15 degrees). Starting at angle 0 , the first half quadrant window extends from 0 to 45 degrees; the second quadrant extends from 30 to 75 degrees (because of the overlap) and so on. Figure 4-3 shows an illustration of the Cartesian space with the first and second half quadrant windows highlighted. A total of 12 features are extracted for each image. Figure 4-4 (a) shows two sample histograms for the OGDF feature vectors for the same writer (first and second samples for Writer X), and Figure 4-4 (b) shows two sample histograms for the OGDF feature vectors for different writers (first sample for Writer X and first sample for Writer Y). The sum total of absolute differences for Figure 4-4 (a) is 39925.2 and for Figure 4-4 (b) is 144192.2.

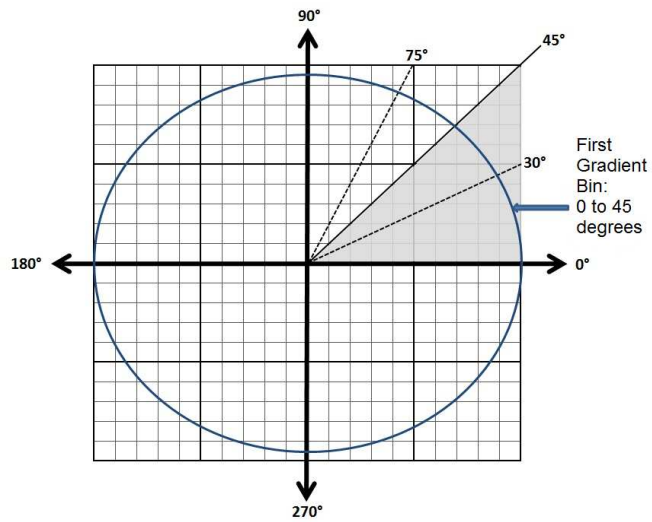


Figure 4-3: First and second gradient feature bins.

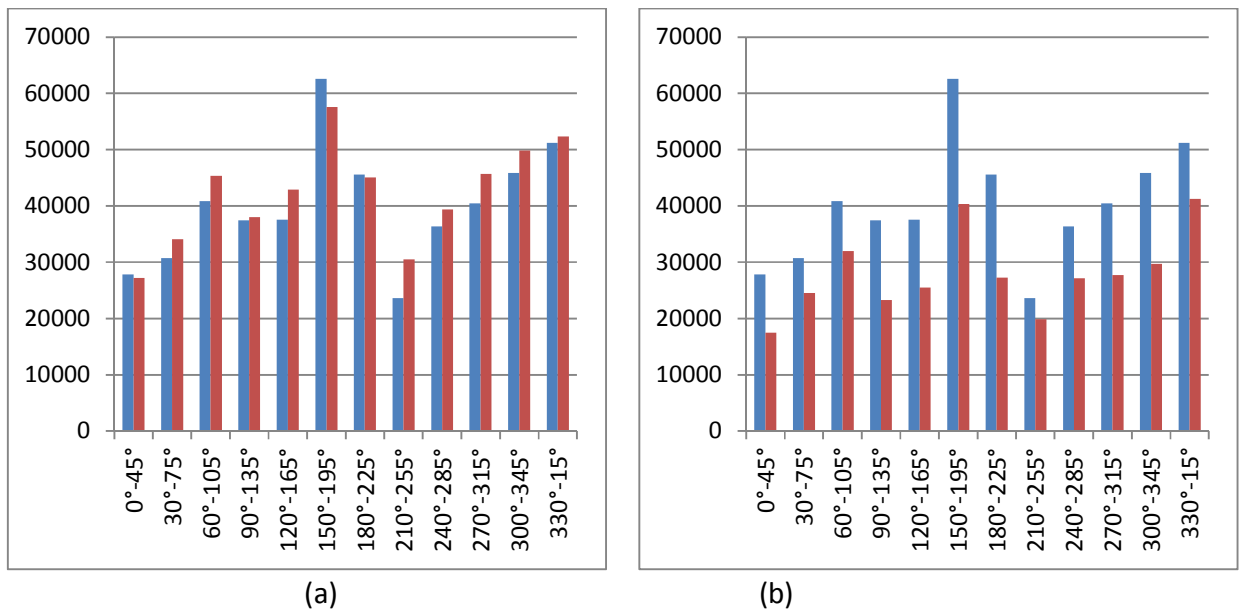


Figure 4-4: a) Histogram for OGDF feature vector for the same writer, b) for different writers.

4.3 Gradient Distribution Features (GDF)

Researchers have successfully implemented histograms of gradient angles in writer identification and in other pattern recognition fields as well (Dalal & Triggs, 2005; Lowe, 1999; S. Srihari, S. H. Cha, Arora, & Lee, 2002). However, researchers often reduced the number of original gradient angles bin (360 bins) into fewer histogram bins by taking the histogram of the range of angles into each bin producing an undesired loss of information. For the GDF features, we utilize the knowledge of the domain to devise a more representative solution.

A novel technique is implemented to retain the information stored in all gradient angles bin. Since our input images are binary images, the output of the $\text{Gradient}_y/\text{Gradient}_x$ can only be one of 9 values; -4, -3, -2, -1, 0, 1, 2, 3, 4. Moreover, the output of the inverse tangent is one of 24 angles. Specifically, 0.0, 18.4, 26.6, 45.0, 63.4, 71.6, 90.0, 108.4, 116.6, 135.0, 153.4, 161.6, 180.0, 198.4, 206.6, 225.0, 243.4, 251.6, 270.0, 288.4, 296.6, 315.0, 333.4, and 341.6 degrees. Since the histogram of zero degrees contain little information and has been shown experimentally that it doesn't improve results, we only calculate the histogram of the remaining 23 angles into 23 different and distinct bins. Figure 4-5 shows a sample picture along with the gradient angles corresponding to four sample regions of interest shown for further illustration. A total of 23 features are extracted for each image. Figure 4-6 (a, b) shows sample histograms for the GDF feature vectors for the same and different writers, respectively. The sum total of absolute differences for Figure 4-6 (a) is 8782 and for Figure 4-6 (b) is 24751.

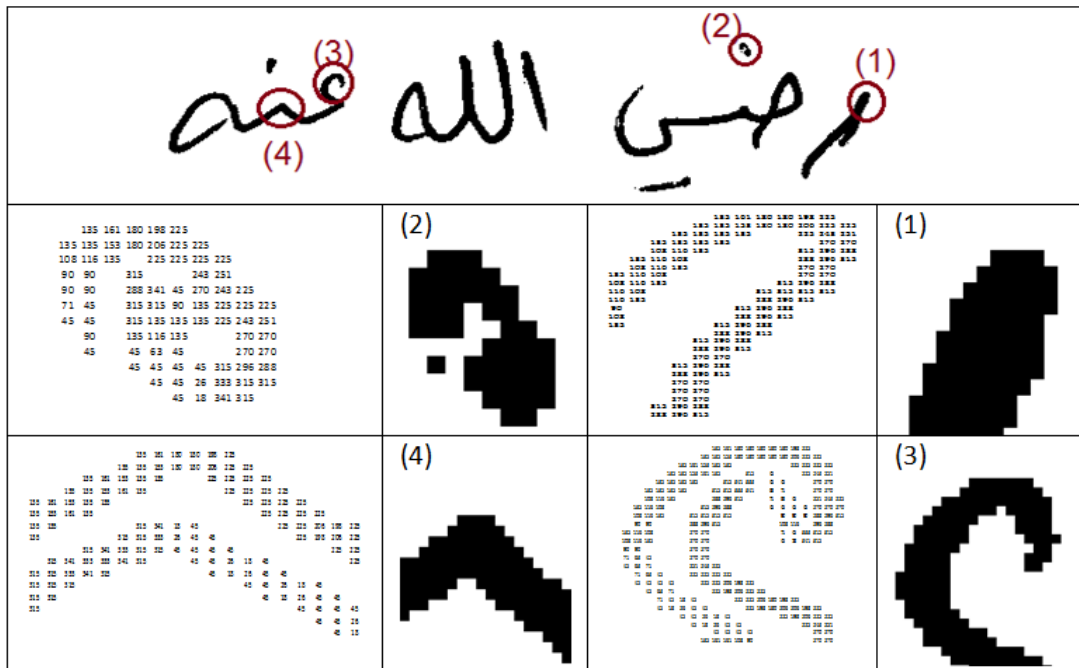


Figure 4-5: Examples of gradient angles for four regions of interest.

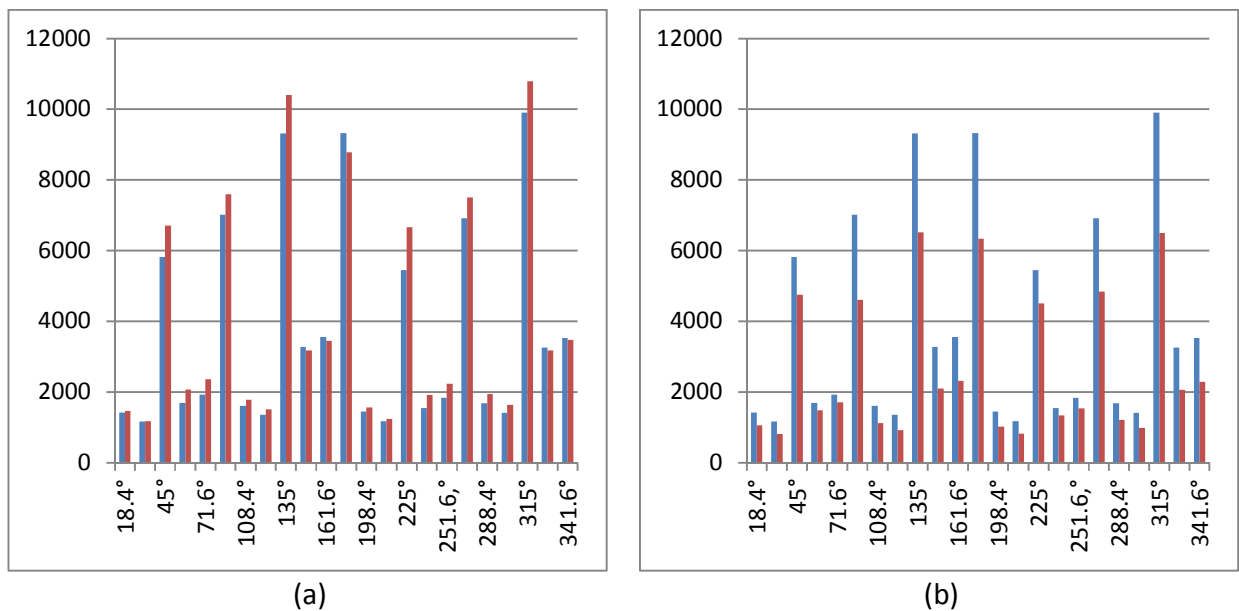
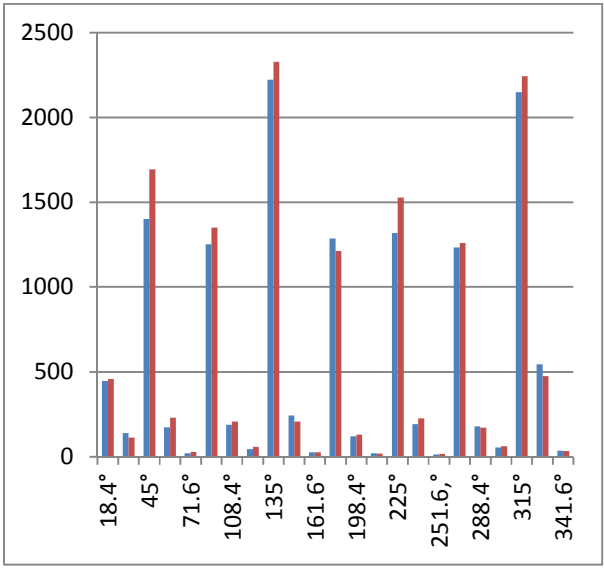


Figure 4-6: a) Histogram for GDF feature vector for the same writer, b) for different writers.

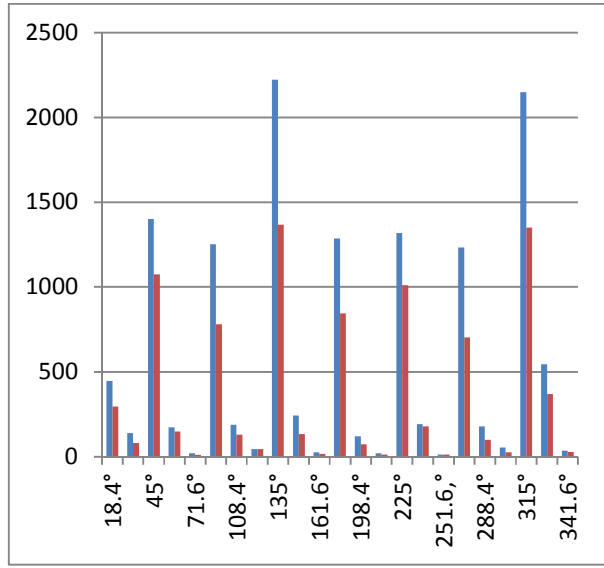
4.4 Windowed Gradient Distribution Features (WGDF)

The gradient distribution features work on the local scale by measuring the gradient angle of each pixel, and accumulating the histogram for each angle. To observe the writer identification features on a higher scale, we employ the sliding window technique on the gradient distribution features explained previously. New features can be extracted using a 3 by 3 sliding window on the gradient angles' image.

After applying the sliding window on the GDF feature map, each pixel will have one of 23 gradient values. Different functions were tried on those 9 values for each pixel; i.e. the maximum value, minimum value, average value, and the most frequent value just to name a few. The most frequent value (angle) of each sliding window attained the best experimental results, and hence is chosen for the windowed gradient distribution features. A total of 23 features are extracted for each image, corresponding to the histogram of the most frequent angle in each window. Figure 4-7 (a, b) below shows sample histograms for the GDF feature vectors for the same and different writers, respectively. The sum total of absolute differences for Figure 4-7 (a) is 1211 and for Figure 4-7 (b) is 4522.



(a)



(b)

Figure 4-7: a) Histogram for WGDF feature vector for the same writer, b) for different writers.

4.5 Contour Chain Code Distribution Features (C³DF)

Contour based features had been successfully implemented in writer identification systems (He, Tang, & X. You, 2005; Panagopoulos, Papaodysseus, Rousopoulos, Dafi, & Tracy, 2009; Schomaker, Bulacu, & K. Franke, 2004; Siddiqi & Vincent, 2009). The contour of the image is extracted and encoded using Freeman chain codes (Freeman, 1974) shown in Figure 4-8. For example, if the current pixel is in the center, and the next pixel on the contour is above it, then the chain code is 2. We implemented a contour tracing algorithm for the whole image.

The chain codes estimate the coarse external angle between every two consecutive points on the contour. After tracing along all pixels belonging to the foreground image (black pixels) and calculating Freeman chain codes for each pixel, the histogram of each chain code is evaluated and used as features. A total of 8 features are extracted for each image (number of possible chain codes). Figure 4-9 (a, b) below shows sample histograms for the C³DF feature vectors for the same and different writers, respectively. C³DF capture the specific style for each writer's curvature writing and exploit inter-writer curvature differences. The sum total of absolute differences for Figure 4-9 (a) is 3078 and for Figure 4-9 (b) is 10137.

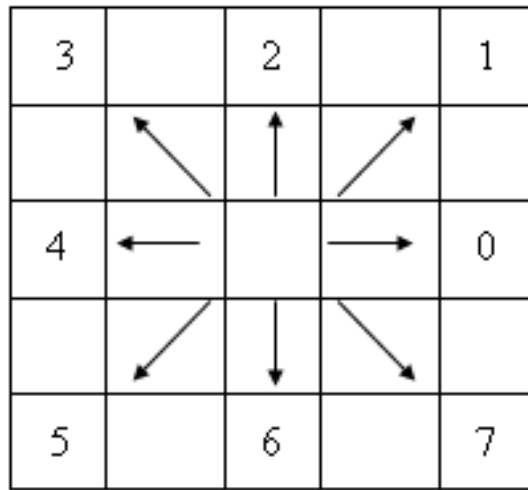


Figure 4-8: Freeman chain codes relative to the center point.

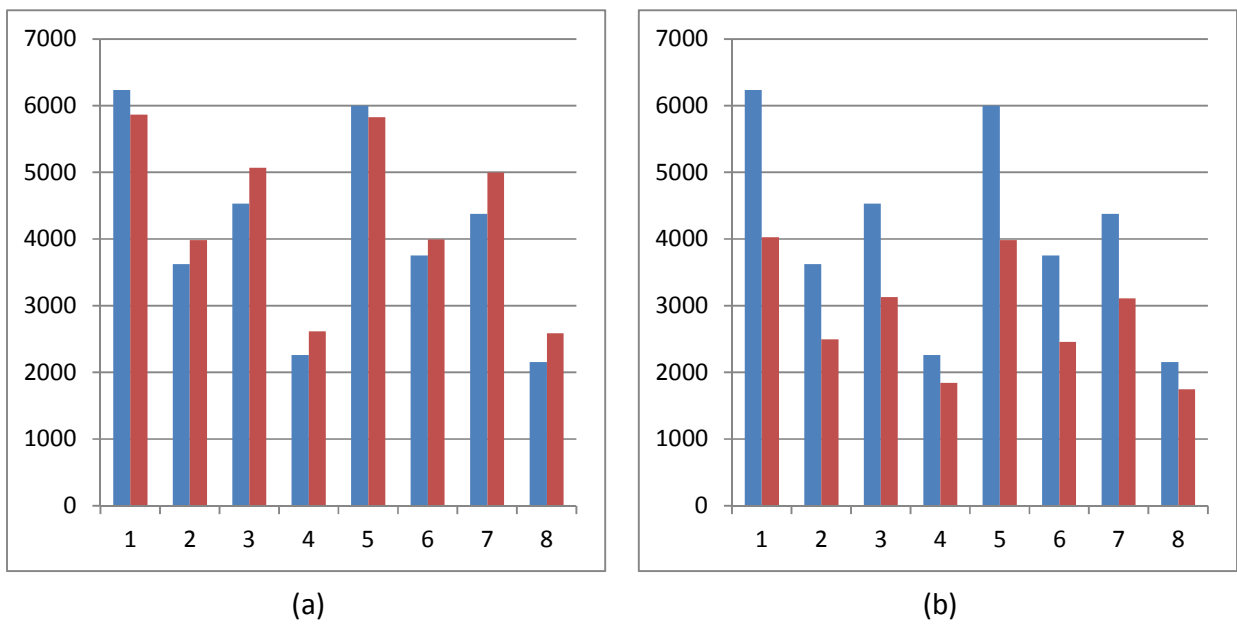


Figure 4-9: a) Histogram for C^3DF feature vector for the same writer, b) for different writers.

4.6 Windowed Contour Chain Code Distribution Features

(WC³DF)

The contour chain code distribution features work on the local scale by measuring the chain codes of each two adjacent points on the contour, and accumulating the histogram for each code. To observe the writer identification features on a higher scale, we employ the sliding window technique on the C³DF explained previously. Using a 3 by 3 sliding window, the most frequent chain code of each sliding window is chosen. A total of 8 features are extracted for each image. Figure 4-10 (a, b) below shows sample histograms for the WC³DF feature vectors for the same and different writers, respectively. The sum total of absolute differences for Figure 4-10 (a) is 1148 and for Figure 4-10 (b) is 3560.

4.7 Medium Scale Features (MSF)

Medium Scale Features (MSF), which represent statistical features on the medium-scale level of the image, can be broken down into four sub-classes of features: density features, horizontal and vertical run features, stroke features, and concavity shape features. The total contributions of these features are 12 values for each image. These features are all concatenated together and labeled MSF due to their relatively small size. Figure 4-11 shows sample histograms for all the MSF feature vectors for the same and different writers, respectively. The sum total of absolute differences for Figure 4-11 (a) is 19670.71 and for Figure 4-11 (b) is 29882.91.

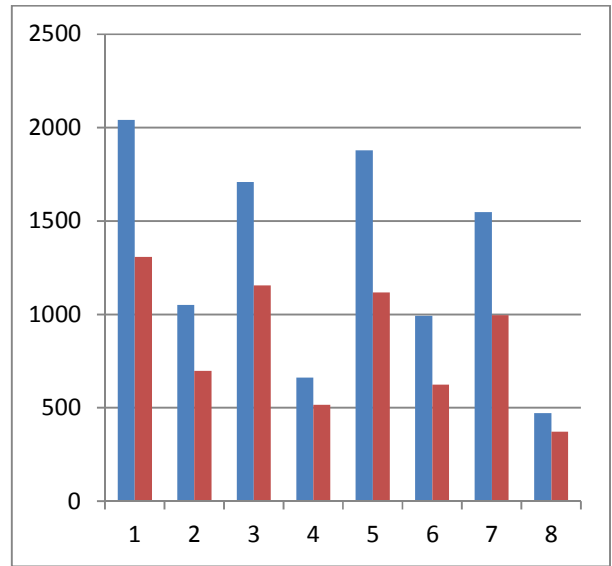
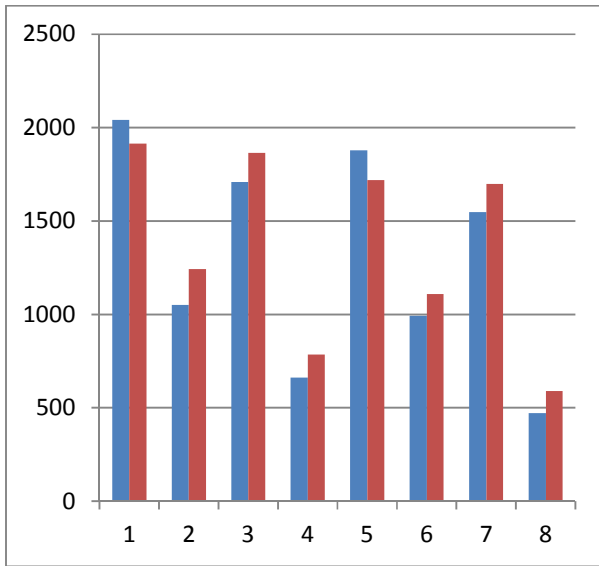


Figure 4-10: a) Histogram for WC³DF feature vector for the same writer, b) for different writers.

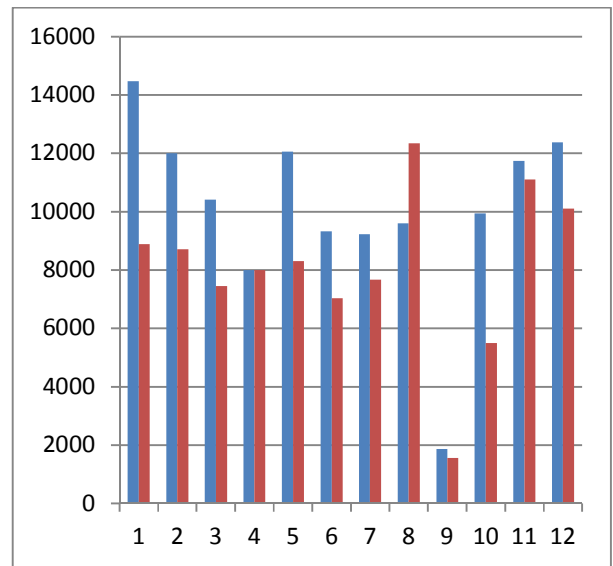
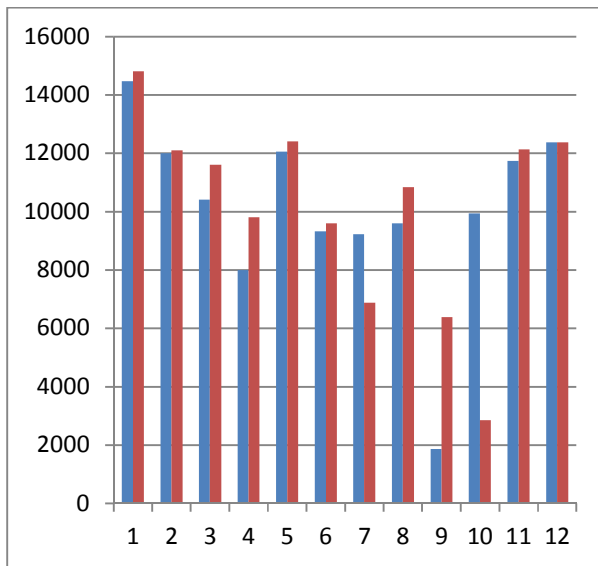


Figure 4-11: a) Histogram for MSF feature vector for the same writer, b) for different writers.

4.7.1 **Density Features (DF)**

The average density of the black pixels in each image is calculated and used as a feature by counting the number of foreground image pixels. One feature value is extracted for each image.

4.7.2 **Horizontal and Vertical Run Features (HVRF)**

The horizontal and vertical run lengths in each image are accumulated by adding the count of black horizontal and vertical lines that constitute a run of more than 2 pixels. This horizontal and vertical runs contribute 2 features.

4.7.3 **Stroke Features (SF)**

These features attempt to capture large horizontal and vertical strokes in the image. These features estimate the number of horizontal, vertical, left- and right-diagonal strokes. Run lengths of horizontal, vertical, left- and right-diagonal foreground pixels across the image are first computed. From this information, the presence of strokes is determined by storing the maximum horizontal, vertical, left- and right-diagonal run length in each region. The stroke features contribute 4 values. It should be noted that the count of horizontal, vertical, left- and right-diagonal runs above a threshold can also be used as another high-level feature type.

4.7.4 Concavity Features (CF)

These features are computed by convolving the image with a star like operator. This operator shoots rays in eight directions and determines what each ray hits. A ray can hit an image pixel or the edge of the image. A table is built for the termination status of the rays emitted from each white pixel of the image. The class of each pixel is determined by applying rules to the termination status patterns of the pixel.

Upward/downward, left/right pointing concavities are detected along with holes. For example, if rays emanating from a background pixel hit an image pixel in all direction, then we consider this background pixel to be a hole, and so on. The rules are relaxed to allow nearly enclosed holes (broken holes) to be detected as holes. This gives a bit more robustness to noisy images. A total of 5 features (viz. hole concavity, upward concavity, downward concavity, right concavity, and left concavity) are extracted for each image.

4.8 Connected Component Features (CCF)²

Despite the fact that automatic writer identification and verification is based on earlier work in forensic science, researches have seldom employed the forensic domain knowledge to design novel features. Furthermore, by mostly relying on statistical measures, researchers have not exploited the semantic knowledge to design features that

² Many thanks to Mohammad Parvez for providing his Letter/Sub-letter segmentation code used in this section.

are language specific. In this work, a novel approach is used to extract structural features that build on some of the main characteristics of the Arabic language.

The feature extraction technique starts by using a segmentation algorithm that divides a paragraph image into letters/sub-letters. The paragraph image is first split into lines, and then each line is split into connected component blobs. These blobs are segmented by inserting white pixels between consecutive letters/sub-letters. Finally, the resulting parts are grouped back together into a full paragraph image but with cuts on each letter/sub-letter as illustrated in Figure 4-12.

Next, connected components' ellipses that have the same second-moments as the connected components are estimated. Soft structural rules are constructed to extract common letters and shapes in the Arabic language. For example, The letter Aleph (ا), the most frequently used letter in Arabic, is extracted by filtering all shapes that have a height/width ratio larger than 4 with an ellipse's angle between -20 and 20 degrees. Other soft rules are created to extract the horizontal segments (that can be found in double dots and parts of many common letters), the circles (found in letters hah (ح), meem (م), waw (و), and the half circles. Figure 4-13 shows sample images of the extracted Alefs, circles, half circles, and horizontal segments with their superimposed ellipses.



Figure 4-12: Letter segmented image.

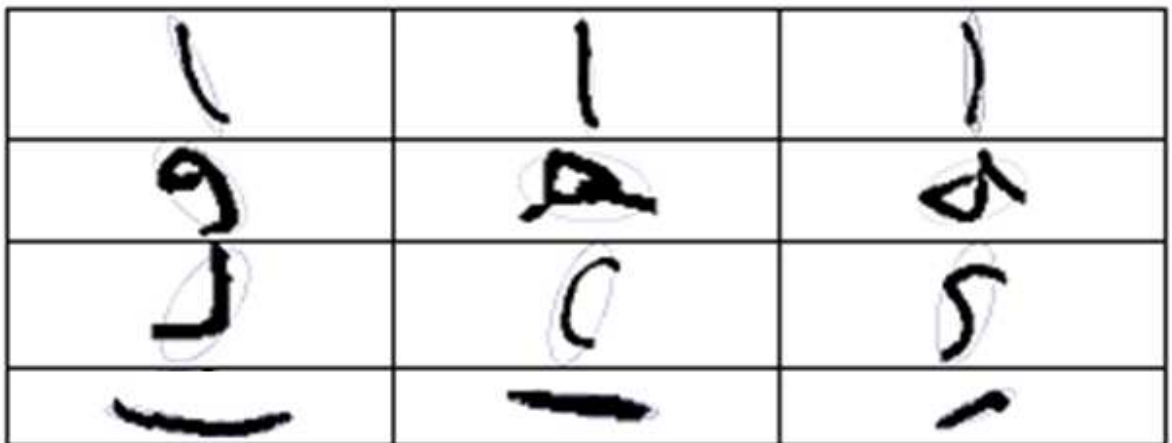
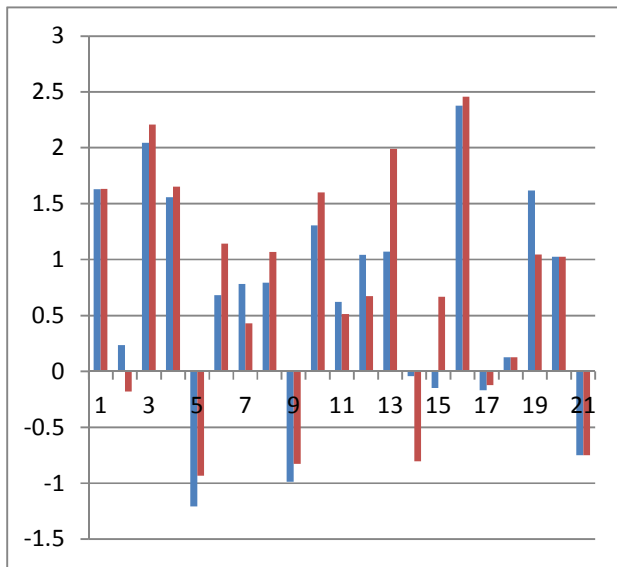


Figure 4-13: Samples of connected components of Alefs, circles, half circles, and horizontal segments, respectively.

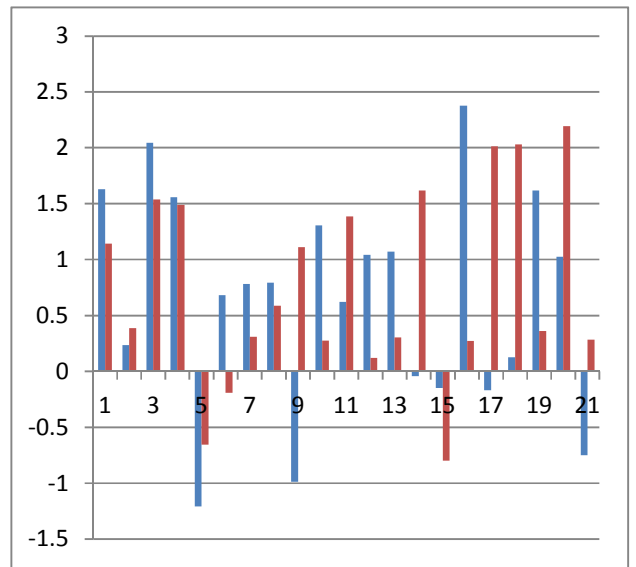
Twenty one structural and statistical features are estimated. These features are designed taking into consideration some of the features used by forensic experts for the Latin language but modified to suit Arabic script. The resulting features include:

- The mean of acute, obtuse, and reflex angles for all connected components.
- The mean area of all the Alephs.
- The percentage of Alef-like shapes, circles, half circles, and horizontal lines compared to the total number of segments.
- Histogram of -90, -30, 30, and 90 degrees for Alefs, horizontal lines, half circles, and circles.

Figure 4-14 (a, b) shows sample histograms for the CCF feature vectors for the same and different writers, respectively. The sum total of absolute differences for Figure 4-14 (a) is 6.15 and for Figure 4-14 (b) is 20.8. Readers can notice the similarity for intra-writer feature vectors and the differences for inter-writer feature vectors.



(a)



(b)

Figure 4-14: a) Sample histogram for CCF feature vector for the same writer (left), b) for different writers (right).

4.9 Conclusions

A novel approach is used to extract structural features that build on some of the main characteristics of the Arabic language. These features are dubbed the connected component features. Other statistical features are extracted (viz. gradient distribution features, contour chain code distribution features, density features, horizontal and vertical run lengths features, stroke features, and concavity features) from the sample image. Table 4-1 gives an overview of these features. The extracted features are tested and analyzed in the following chapters on writer identification of Arabic handwritten digits and text.

Table 4-1: Overview of features used in writer identification of Arabic handwritten digits and text.

Short Name	Feature Full Name	# of Variables	Used in
CCF	Connected Component Features	21	Text
OGDF	Overlapped Gradient Distribution Features	12	Digits
GDF	Gradient Distribution Features	23	Text
WGDF	Windows Gradient Distribution Features	23	Text
C ³ DF	Contour Chain Code Distribution Features	8	Digits & Text
WC ³ DF	Windowed Contour Chain Code Distribution Features	8	Text
DF	Density Features	1	Digits
HVRF	Horizontal and Vertical Run Features	2	Digits
SF	Stroke Features	4	Digits
CF	Concavity Features	5	Digits

CHAPTER 5

WRITER IDENTIFICATION USING ARABIC HANDWRITTEN DIGITS

In this chapter, Arabic handwritten digits are analyzed for writer identifiability. In addition to digit identifiability, the chapter presents digit recognition in order to validate the effectiveness of the features in digit recognition as well as writer identification. The digit image is divided into grids based on the distribution of the black pixels in the image. An extensive experimentation was carried out to estimate the suitable number of grid dimensions. Several types of features are extracted from the grid segments; viz. gradient, contour chain code, density, horizontal and vertical run lengths, stroke, and concavity features. K-Nearest Neighbor and Nearest Mean classifiers are used. A database of 70000 Arabic handwritten digit samples written by 700 writers is used in the analysis and experimentations. Many experiments are conducted using the nearest neighbor classifier.

Writer identifiability using isolated and combined digits was investigated. Analysis of the results indicates that writer identification systems using Arabic digits ‘٣’ (3), ‘٤’ (4), ‘٨’ (8), and ‘٩’ (9) are more identifiable than using other digits while writer identification systems using Arabic digit ‘٠’ (0) and ‘١’ (1) are the least identifiable. In addition, the chapter shows that combining the writer's digits increases the discriminability power of writer identification. Combining the features of all digits, K-

NN provided the best accuracy in text-independent writer identification with top-1 result of 88.14%, top-5 result of 94.81%, and top-10 results of 96.48%.

5.1 Introduction

Arabic handwritten digits identification is addressed as a first step in a comprehensive research in the effort of writer identification and verification using Arabic handwritten documents. These results of writer discriminability of Arabic digits for writer identification encouraged us to extend it to the writer identifiability using Arabic handwritten text which is addressed in the following chapter. We are not aware of any previous work in writer identification using Arabic handwritten digits. To our knowledge, only one work is reported on writer identification in Latin handwritten digits (Leedham & Chachra, 2003).

Leedham and Chachra used a database consisting of 15 writers. Each writer was asked to write random strings of 0 to 9 at least 10 times. The Hamming distance measure was used for identification. They published results of writer identification, verification, and forgery detection for different number of sets for training and testing. They reported an accuracy of 100% for writer identification using test sets of 3 and 4 samples per writer (while the remaining sets being used for training). With this limited number of writers, it is to be validated whether their features are effective with large number of writers

In this chapter, Section 2 presents a summarized description of the used features; the experimental results are detailed in Section 3; and finally, the conclusions are given in Section 4.

5.2 Features

In this work multiple types of features are used. Overlapped Gradient Distribution Features (OGDF), Contour Chain Code Distribution Features (C³DF), Density Features (DF), Horizontal and Vertical Run lengths Features (HVRF), Stroke Features (SF), and Concavity Features (CF) are implemented. Some of these features are classically known (Favata & Srikantan, 1996; Schomaker & Bulacu, 2004), and have been successfully implemented by the author in previous digit recognition tasks (Awaidah & S. A. Mahmoud, 2009; Sabri Mahmoud & Awaida, 2009). In our implementation, gradient features have been tweaked to improve results.

The first step in the feature extraction algorithm is to divide the image into $n \times m$ grids with equal number of foreground pixels for each of n rows, and equal number of foreground pixels for each of m columns. Figure 5-1 shows the Arabic numeral 6 divided into 3×3 , 4×4 , 5×5 and 6×6 divisions. As can be seen from the figure, the horizontal segments have equal number of black pixels and the vertical segments have equal number of black pixels.

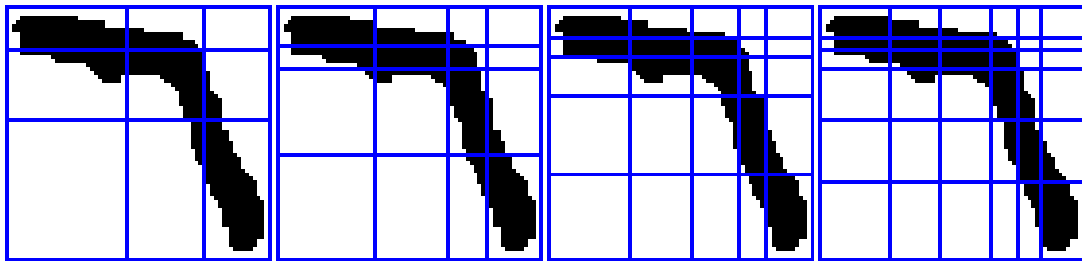


Figure 5-1: Arabic digit 6 divided into 4 different divisions.

The digit image is divided into $n \times m$ grids with equal number of black pixels in each of n rows, and for each of m columns. The features of the individual grid segments are extracted. The overlapped gradient distribution features produce $12 \times n \times m$ features representing the gradient feature vector (where n and m are the number of horizontal and vertical segments respectively). Contour chain code distribution features add $8 \times n \times m$ features. Medium scale features contribute $12 \times n \times m$ features. A detailed description of the extracted features is given in Chapter 4. A concise summary of these features is given below. These feature vectors are normalized to zero mean and variance of one before being presented to the classifier.

5.2.1 Overlapped Gradient Distribution Features (OGDF)

OGDF are computed by convolving the x and y Sobel operators with the binary image segment. The gradient angle of each pixel is calculated by computing the inverse tangent of $(\text{Gradient}_y/\text{Gradient}_x)$. Subsequently, the histogram of the gradient angles is calculated and stored as the feature vector.

A sliding window of a number of degrees (x) is used to estimate the histogram of gradient directions of the pixels in the window in the gradient features. The optimal value of x degrees will be calculated in the Feature Selection section (Section 5.5.2) and will be shown to be 45 degrees. The sliding window overlaps with the previous window by $1/3$ of the window range (i.e. 15 degrees). The total contributions of these features are 12 values for each image segment.

5.2.2 **Contour Chain Code Distribution Features (C³DF)**

The contour of each image segment is extracted and encoded using Freeman chain codes (Freeman, 1974). The histogram of each chain code is estimated and used as features. A total of 8 features are extracted for each image segment.

5.2.3 **Medium Scale Features (MSF)**

Medium Scale Features (MSF) can be broken down into four sub-classes of features: Density Features (DF), Horizontal and Vertical Run Features (HVRF), Stroke Features (SF), and Concavity Features (CF). The total contributions of these features are 12 values for each image segment. These features are all concatenated together and labeled MSF due to their relatively small size.

DF are extracted by calculating the average density of the black pixels in each image giving one feature value for each image. HVRF accumulates the number of horizontal and vertical run lengths in each image, giving 2 features for the image. SF first compute the run lengths of horizontal, vertical, left- and right-diagonal foreground pixels across the image segment. From this information, the presence of strokes is determined by storing the maximum horizontal, vertical, left- and right-diagonal run length in each segment. The stroke features contribute 4 values.

CF detect Upward/downward, left/right pointing concavities along with holes. These features are computed by convolving the image segment with a star like operator. This operator shoots rays in eight directions and determines what each ray hits. The class

of each pixel is determined by applying rules to the termination status patterns of the pixel. A total of 5 features are extracted for each image segment.

5.3 Arabic Digits Database

Abdleazeem et al. described their Arabic Digits dataBase (ADBase) in (Abdleazeem & El-Sherif, 2008). ADBase is composed of 70,000 digits written by 700 participants. Each participant wrote each digit (from '0' to '9') ten times. Images size in pixels varies from 3 by 5 pixels for the smallest image and up to 140 by 29 pixels for the largest image. Figure 5-2 shows samples of the ADBase. The database is partitioned into two sets for the purpose of digits recognition: a training set (60,000 digits with 6,000 images per class) and a test set (10,000 digits with 1,000 images per class). Writers of training and test sets are disjoint.

In order to use the database for writer identification, we divided the database into two sets: training set and testing set. The training set contains 49,000 digits (70% of the dataset), whereas the testing set contains 21,000 digits (30% of the dataset). For each writer, 70 random digits are selected for the training set (7 samples per digit for each writer), and the remaining 30 samples are selected for the testing set (3 samples per digit for each writer).

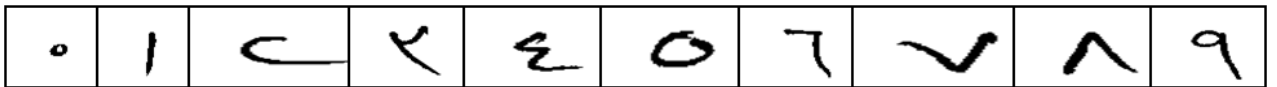


Figure 5-2: Samples of ADBase.

For selecting the optimal number of grids, optimal feature parameters, and feature combinations, the training set is further divided into initial-training and verification sets. 10,000 digits written by the first 100 participants are chosen for the initial-training and verification sets. Initial-training set contains 7000 samples (100 writers, 10 digits/writer, and 7 samples/digit) and the verification set contains 3000 samples (100 writers, 10 digits/writer, and 3 samples/digit).

5.4 Nearest Neighbor (NN) and Nearest Mean (NM) Classifier

Nearest Neighbor (NN) and Nearest Mean (NM) are simple classifiers that are used to measure the effectiveness of the extracted features and the writer identifiability using Arabic handwritten digits. The nearest neighbor is computed using an Euclidean distance formula given by:

$$E_i = \sqrt{\sum_{j=1}^k (M_{ij} - V_j)^2} \quad \text{Equation 1}$$

where:

- E_i is the distance between the input digit and model i .
- k is the total number of parameters in the feature vector.
- M_{ij} is the j^{th} feature of model i .
- V_j is the j^{th} feature of the input digit feature vector.

The distance (E_i) between the test sample and the feature vectors for all models are found. The argument of the minimum value found yields the recognized model i . This

model is considered as the writer class that matches most closely the obtained features vector of the unknown writer.

Writer identification researchers have preferred distance and dissimilarity measures over statistical classifiers like Hidden Markov Models (HMM) and Support Vector Machines (SVM) mainly because of the nature of the writer identification problem (Schlapbach & Bunke, 2007; Schomaker & Vuurpijl, 2000; Zaher & Abu-Rezq, 2010). The problem of writer identification usually involves large number of classes (i.e. writers) and few samples per class (e.g. digits per writer) compared to relatively few classes (i.e. number of distinct digits) and large number of samples per class (i.e. samples of images per digit) common in digit recognition scenarios. In addition, Schlapbach (Schlapbach, 2007) reported better results using the k-NN classifier over the SVM classifier.

The models for the NM classifier are taken as the mean of all the features of the training samples for each digit of each writer. This is done by averaging the features of samples of each digit of each writer and using them as the feature models for the writers. For the NM, the feature vectors for the training set for each digit and each writer are averaged. In our case, this reduces the number of feature vectors for training set from 49,000 training vectors into 7,000 averaged vectors (700 writers, 10 average digits per writer).

5.5 Feature Selection

Feature selection involves answering the following question. Given a combination of features, how can we choose the most important of them so as to reduce their dimensionality and at the same time improve (or at least preserve) their classifier's accuracy (Theodoridis & Koutroumbas, 2009)? Several types of experiments were performed to choose the optimal number of grid segments of the digit image, the best feature parameters, and the best combination of features.

A certain criterion is needed to judge if one feature type outperforms another feature type. In this case, the statistical significance measure is one of the acknowledged evaluation criteria in the pattern recognition field. The following derivations of the statistical significance measure are taken from (Plötz, 2005) and are presented here for ease of reference. A Bernoulli process can take only two values; 0 and 1. Hence, a writer identification result can be considered to be a Bernoulli process where the output is either a hit (H) or a miss (M). Thus, we calculate the confidence level for the accuracy percentage of a writer identification experiment, $\hat{\mathcal{E}}$, by evaluating the Binomial distribution. The local limit theorem states that $\hat{\mathcal{E}}$ can be estimated as asymptotically normally distributed for N number of experiments:

$$\frac{\hat{\mathcal{E}} - \varepsilon}{\sqrt{\frac{\varepsilon(1-\varepsilon)}{N}}} \approx N(0,1) \quad \text{Equation 2}$$

The level of confidence, denoted by $1-\alpha$ is computed. A level of confidence of $1 - 0.05 = 0.95$ (95%) is used. We can calculate the multiplier of the standard error, z , as:

$$z = \text{quantile of the normal distribution of } \left(100 \times \left(1 - \frac{\alpha}{2}\right)\right) = 1.96 \quad \text{Equation 3}$$

where the values of the quantile of the normal distribution can be looked up in statistical tables. Hence the lower and upper boundaries of the confidence interval can be defined as:

$$\varepsilon_{l/u} = \frac{N}{N+z^2} \left(\hat{\varepsilon} + \frac{z^2}{2N} \pm z \sqrt{\frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{N} + \frac{z^2}{4N^2}} \right) \quad \text{Equation 4}$$

Results that are outside the boundaries of the confidence interval can be considered statistically significant. Otherwise, these small differences in results can be considered insignificant and can be contributed to chance.

5.5.1 Optimal Number of Grid Segments

In order to estimate the optimal number of grid segments of the digit image, several experiments are conducted using divisions of 2 x 2 up to 8 x 8 on the initial-training and verification sets. Experimental results have shown that 5 x 5 divisions resulted in the highest recognition rate as shown in Figure 5-3. Figure 5-4 shows a sample of digit ‘9’ (9) divided into 5 x 5 divisions.

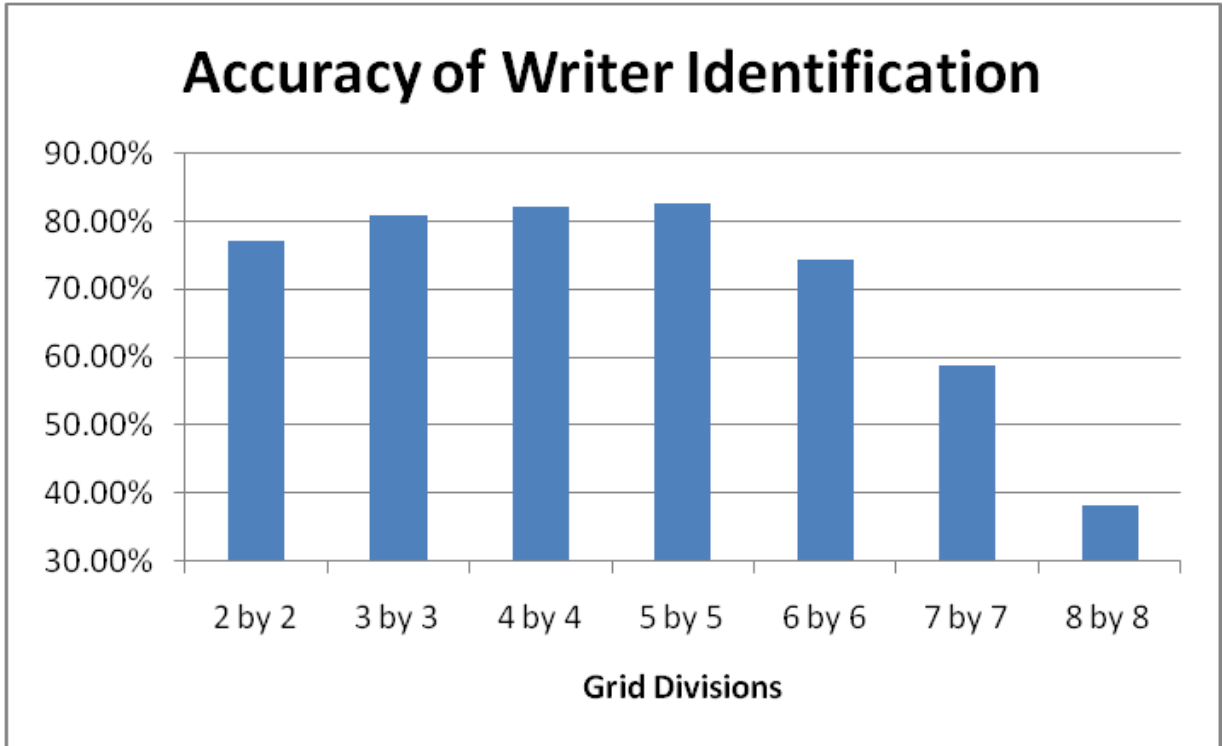


Figure 5-3: Writer identification accuracy at different divisions.

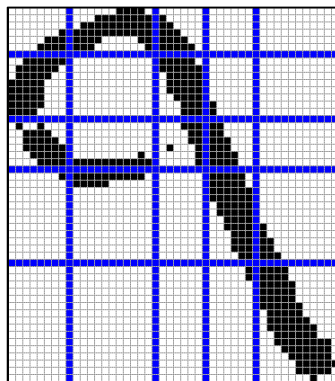


Figure 5-4: Digit '9' (9) divided into 5 x 5 divisions.

5.5.2 Selecting Overlapped Gradient Distribution Features (OGDF)

Sliding Window Size

For selecting the optimal feature parameters and feature combinations, four types of experiments were performed:

1. Exp1: Top-1, top-5 and top-10 writer-identification performance results as well as digit classification results for all samples using the NN classifier.
2. Exp2: Writer-identification and digit classification results for all samples using the NM classifier.
3. Exp3: The extracted features for each group of digits (0 to 9) are concatenated to form one feature vector.
4. Exp4: Each digit in the testing set is compared against all digits in the training and its rank is stored. This is done for all the digits (0-9) for that specific writer, and then the rank for each writer is added and the most probable writer is selected.

These experiments are explained in more details in the following section. For the OGDF features, as explained previously, a sliding window of a number of degrees (x) is used to estimate the histogram of gradient directions of the pixels in the window of the gradient features. So, if we use a sliding window of 10 degrees, then we would have 36 features' vector for each grid ($360/10 = 36$), and if we use a sliding window of 45 degrees, then we would have 8 features vector (78% reduction in feature size). Four different sliding window sizes (10, 20, 30, and 45 degrees) are tested on 100 writers for

all 4 experiments as shown below in Table 5-1, Table 5-2, Table 5-3, and Table 5-4. Results within the same confidence measure are highlighted and hence can be considered to be giving the same statistical accuracy and the difference in the results can be contributed to chance. Since using a sliding window size of 20, 30, and 45 degrees gives the same statistical accuracy result, and since using a sliding window of 45 degrees provides the least number of features, the sliding window size of 45 degrees is used for the OGDF features. It is also interesting to note that there is a positive correlation between the writer identification and digit recognition accuracy results. The gradient features seem to preserve the writer intrinsic properties while maximizing the different characteristic of each digit.

Table 5-1: Exp1 results (All samples using the NN classifier) for different sliding window sizes.

Sliding Window Size	Accuracy of Writer Identification			Accuracy of Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
10 Degrees	18.90%	40.60%	53.47%	97.13%	99.17%	99.43%
20 Degrees	21.43%	44.67%	56.93%	98.40%	99.27%	99.43%
30 Degrees	20.70%	44.90%	56.27%	98.27%	99.43%	99.67%
45 Degrees	21.20%	44.90%	56.87%	98.23%	99.30%	99.53%
Significance	1.50%	1.79%	1.76%	0.39%	0.21%	0.15%

Table 5-2: Exp2 results (All samples using the NM classifier) for different sliding window sizes.

Sliding Window Size	Accuracy of Writer Identification			Accuracy of Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
10 Degrees	23.23%	48.40%	60.13%	96.07%	97.77%	98.17%
20 Degrees	24.77%	50.43%	61.57%	96.83%	98.23%	98.47%
30 Degrees	25.47%	51.13%	62.53%	96.93%	98.23%	98.43%
45 Degrees	24.73%	50.70%	62.30%	96.67%	97.97%	98.47%
Significance	1.59%	1.79%	1.72%	0.56%	0.41%	0.38%

Table 5-3: Exp3 results (Group of digits, 0 to 9, are concatenated to form one feature vector) for different sliding window sizes.

Sliding Window Size	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
10 Degrees	67.67%	81.33%	85.33%	79.67%	89.67%	92.67%
20 Degrees	73.67%	82.67%	86.33%	83.33%	93.00%	95.00%
30 Degrees	73.00%	84.00%	89.00%	84.00%	93.33%	96.00%
45 Degrees	75.00%	84.33%	87.00%	84.33%	93.33%	95.00%
Significance	4.56%	3.68%	3.06%	3.68%	2.31%	1.70%

Table 5-4: Exp4 results (Each digit in testing set is compared against all digits in training and then distances are added for each group of digits) for different sliding window sizes.

Sliding Window Size	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
10 Degrees	84.67%	95.33%	97.67%	86.00%	95.00%	96.33%
20 Degrees	87.33%	96.67%	98.33%	89.00%	96.00%	97.33%
30 Degrees	89.67%	98.67%	99.00%	89.00%	96.00%	98.00%
45 Degrees	89.33%	97.33%	98.67%	89.00%	96.00%	98.33%
Significance	2.96%	0.81%	0.66%	3.06%	1.70%	0.95%

5.5.3 Feature Combinations Selection

Results for OGDF features using the optimal 45 degrees sliding window size as well as results for the C^3DF , and MSF features are shown. The three features type (OGDF, C^3DF , MSF) are concatenated with all possible combinations and there results are shown as well. Table 5-5, Table 5-6, Table 5-7, and Table 5-8 show the writer identification and digit recognition results for experiments 1, 2, 3, and 4, respectively for feature combinations. Results were also compared with previously published work in digit recognition labeled GSC features (Awaidah & S. A. Mahmoud, 2009).

It is clear from the tables that combining all the features gives the best accuracy results almost constantly. Even though different feature combinations gave the same statistically significant results on some experiments, using all features was the only combination that did it almost in all experiments except for Exp. 3 where it achieved similar results.

Table 5-5: Exp1 results (All samples using the NN classifier) for different feature combinations.

Features	Accuracy of Writer Identification			Accuracy of Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF (45 Degrees)	21.20%	44.90%	56.87%	98.23%	99.30%	99.53%
C ³ DF	16.00%	37.37%	50.77%	96.63%	99.03%	99.37%
MSF	23.37%	46.20%	57.80%	98.87%	99.70%	99.80%
OGDF, MSF	23.50%	46.77%	58.10%	98.80%	99.63%	99.73%
OGDF, C ³ DF	23.50%	46.77%	58.10%	98.80%	99.63%	99.73%
C ³ DF, MSF	23.50%	46.43%	58.30%	99.00%	99.73%	99.80%
All	24.37%	49.13%	60.50%	98.90%	99.70%	99.83%
GSC	23.37%	47.33%	59.57%	98.97%	99.63%	99.80%
Significance	1.57%	1.79%	1.73%	0.30%	0.13%	0.10%

Table 5-6: Exp2 results (All samples using the NM classifier) for different feature combinations.

Features	Accuracy of Writer Identification			Accuracy of Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF (45 Degrees)	25.80%	51.73%	63.47%	97.10%	98.20%	98.53%
C ³ DF	20.70%	46.40%	60.37%	95.97%	98.63%	99.10%
MSF	23.83%	49.60%	62.27%	97.03%	98.57%	99.00%
OGDF, MSF	26.37%	52.67%	64.03%	97.33%	98.43%	98.70%
OGDF, C ³ DF	25.13%	51.37%	63.27%	97.60%	98.63%	98.83%
C ³ DF, MSF	24.10%	50.00%	62.70%	97.20%	98.67%	99.13%
All	26.53%	52.90%	64.20%	97.67%	98.73%	98.90%
GSC	26.53%	52.57%	63.93%	97.10%	98.27%	98.67%
Significance	1.61%	1.78%	1.70%	0.48%	0.34%	0.28%

Table 5-7: Exp3 results (Group of digits, 0 to 9, are concatenated to form one feature vector) for different feature combinations.

Features	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF (45 Degrees)	75.00%	84.33%	87.00%	84.33%	93.33%	95.00%
C ³ DF	78.00%	91.33%	94.00%	90.00%	98.67%	99.33%
MSF	72.67%	84.33%	89.00%	86.00%	94.33%	96.67%
OGDF, MSF	74.33%	84.00%	88.00%	84.33%	94.33%	96.67%
OGDF, C ³ DF	74.00%	83.00%	86.67%	82.67%	93.00%	94.33%
C ³ DF, MSF	73.00%	84.33%	89.33%	86.00%	94.33%	97.33%
All	74.33%	84.00%	88.00%	84.33%	94.33%	96.67%
GSC	74.67%	85.33%	89.00%	85.33%	94.33%	96.67%
Significance	4.32%	2.68%	2.17%	2.91%	0.81%	0.49%

Table 5-8: Exp4 results (Each digit in testing is compared against all digits in training and then distances are added for each group of digits) for different feature combinations.

Features	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF (45 Degrees)	89.33%	97.33%	98.67%	89.00%	96.00%	98.33%
C ³ DF	89.00%	98.00%	98.33%	90.67%	99.00%	99.00%
MSF	87.00%	97.67%	98.33%	88.67%	97.00%	98.33%
OGDF, MSF	89.00%	98.67%	99.33%	89.67%	96.33%	99.00%
OGDF, C ³ DF	89.00%	97.67%	99.33%	88.67%	96.00%	98.00%
C ³ DF, MSF	87.67%	97.67%	98.33%	88.67%	97.00%	98.33%
All	89.00%	98.67%	99.33%	89.67%	96.33%	99.00%
GSC	89.00%	99.00%	99.33%	90.00%	97.00%	98.67%
Significance	3.01%	0.81%	0.49%	2.80%	0.66%	0.66%

5.6 Experimental Results

After smoothing the images, the Overlapped Gradient Distribution Features (using 45 degrees sliding window size), Contour Chain Code Distribution Features, and the Medium Scale Features are extracted for each set (e.g. for a 5×5 division, the concatenation of all of the features resulted in an 800-dimensional feature vector, viz. 300 OGDF, 200 C3DF, and 300 MSF).

With 5×5 grid divisions, training and testing is performed on the ADBase. Using the NN classifier and the above features we tested writer identification and digit recognition using Arabic handwritten digits for all 21000 samples. Table 5-9 shows the top-1, top-5 and top-10 writer-identification performance results as well as digit classification results.

For the NM, the feature vectors for the training set for each digit and each writer are averaged. This reduces the number of feature vectors of the training set from 49,000 training vectors into 7,000 averaged vectors (700 writers, 10 average digits per writer). Table 5-10 shows the results for each digit. The table shows that averaging vectors have reduced digit recognition rate as expected due to the decrease in inter-digit variations (e.g. an Arabic three digit would look more like an Arabic two digit).

Table 5-9: Writer identification and digit recognition accuracy for each digit using NN.

Digit	Accuracy of Writer Identification			Accuracy of Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
0 (०)	2.86%	8.81%	14.19%	98.90%	99.86%	99.90%
1 (१)	4.29%	11.76%	16.48%	99.10%	99.71%	99.86%
2 (२)	12.10%	28.05%	36.95%	99.43%	99.90%	99.90%
3 (३)	19.81%	38.19%	47.67%	98.90%	99.48%	99.48%
4 (४)	17.10%	33.81%	42.76%	99.10%	99.71%	99.86%
5 (५)	14.14%	29.52%	38.81%	99.43%	99.81%	99.81%
6 (६)	8.86%	23.19%	30.76%	99.29%	99.86%	99.86%
7 (७)	13.10%	32.90%	43.90%	99.76%	99.90%	99.95%
8 (८)	15.52%	33.95%	43.52%	99.71%	99.81%	99.95%
9 (९)	15.29%	31.81%	41.00%	98.76%	99.62%	99.81%
Total	12.30%	27.20%	35.60%	99.24%	99.77%	99.84%

Table 5-10: Writer identification and digit recognition accuracy for each digit using NM.

Digit	Accuracy of Writer Identification			Digit Recognition		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
0 (໐)	4.05%	11.14%	17.33%	98.86%	99.38%	99.67%
1 (໑)	4.33%	11.29%	16.00%	96.81%	98.33%	98.62%
2 (໒)	14.67%	31.19%	40.86%	98.10%	99.33%	99.52%
3 (໓)	22.05%	42.10%	50.95%	98.24%	99.19%	99.33%
4 (໔)	17.52%	33.48%	41.95%	98.86%	99.48%	99.62%
5 (໕)	17.90%	35.71%	45.29%	99.29%	99.48%	99.48%
6 (໖)	11.57%	24.71%	33.86%	98.86%	99.62%	99.76%
7 (໗)	17.71%	37.62%	48.24%	99.57%	99.81%	99.86%
8 (໘)	18.81%	38.62%	49.24%	99.24%	99.52%	99.57%
9 (໙)	17.71%	34.48%	44.76%	98.67%	99.33%	99.33%
Total	14.63%	30.03%	38.85%	98.65%	99.35%	99.48%

In order to combine the discriminatory power for all digits, the extracted features for each group of digits (0 to 9) are concatenated to form one feature vector. This is implemented simply by concatenating the features of the different digits as the database consists of only isolated digits. This produced 4900 concatenated training vectors (700 writers, 7 feature vectors per writer) for the k-NN classifier and 700 averaged and concatenated training vectors (700 writers, 1 feature vector per writer) for the NM classifier, and 2100 feature vectors for testing (700 writers, 3 feature vectors per writer). These concatenated feature vectors are used in the analysis using one classifier. Since each digit feature vector is compared to its corresponding digit feature vector in the training set, we consider this approach to be text-dependent writer identification. Table 5-11 shows a summary of the writer identification results for the text-dependent approach.

Finally, we compare each digit in the testing set against all digits in the training set for each writer and store its writer identification rank. We do this for all the digits (0-9) for that specific writer, and then we add the rank for each writer and select the most probable writer, and hence we consider this approach to be text-independent writer identification. Table 5-12 shows a summary of the writer identification results for the text-independent approach. Combining all features using this method gives the best result of 88.14% for top-1 accuracy of writer identification using the NN classifier. These encouraging results show the system ability to identify the writer accurately using only the handwriting of 10 digits for each writer if written at least twice.

Table 5-11: Text-dependent writer identification.

Features	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF	66.19%	79.76%	84.14%	80.33%	89.76%	92.38%
C ³ DF	60.19%	78.71%	84.86%	74.90%	89.14%	92.62%
MSF	65.71%	79.86%	83.76%	82.67%	91.48%	94.19%
All	69.52%	81.67%	85.81%	81.33%	90.67%	92.86%
Significance	1.27%	1.06%	0.95%	1.03%	0.75%	0.62%

Table 5-12: Text-independent writer identification.

Features	NN Classifier			NM Classifier		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
OGDF	84.62%	93.62%	95.90%	85.24%	93.24%	95.14%
C ³ DF	67.38%	84.29%	89.67%	67.10%	83.76%	89.24%
MSF	86.43%	94.67%	96.19%	86.71%	93.90%	95.81%
All	88.14%	94.81%	96.48%	87.76%	94.10%	96.14%
Significance	0.88%	0.59%	0.48%	0.89%	0.63%	0.50%

5.7 Conclusions

This chapter addresses the identifiability of Arabic handwritten digits. Nearest Mean and Nearest Neighbors are used for classification. In addition to digit identifiability, the chapter presents digit recognition. Gradient, contour chain code, density, horizontal and vertical run lengths, stroke, and concavity features are used. A database of Arabic handwritten digits written by 700 different writers is used in the analysis.

A number of experiments were carried out to select the optimal number of digit divisions for the feature extraction phase. Combining all digits (by concatenating their corresponding features) and finding the NN provided the best accuracy in text-independent writer identification with top-1 result of 88.14%, top-5 result of 94.81%, and top-10 results of 96.48%. The analysis of the results indicates that writer identification systems using Arabic digits ‘٣’ (3), ‘٤’ (4), ‘٨’ (8), and ‘٩’ (9) are more identifiable than using other digits while writer identification systems using Arabic digit ‘٠’ (0) and ‘١’ (1) are the least identifiable. K-NN provided best accuracy for digit recognition with top-1 result of 99.24%, top-5 result of 99.77%, and top-10 results of 99.84%, with only 34 erroneously classified digits out of 21,000 test digits in the top-10 results.

These encouraging results demonstrate the discriminability of Arabic digits for writer identification. In addition, they indicate the suitability of these features for both writer identification and digit recognition. We will extend these features for writer identification using Arabic handwritten text in the next chapter.

CHAPTER 6

WRITER IDENTIFICATION OF ARABIC HANDWRITTEN TEXT

This chapter addresses writer identification of Arabic handwritten text. Several types of structural and statistical features are extracted from Arabic handwriting text. A novel approach is used to extract structural features that build on some of the main characteristics of the Arabic language. Connected component features for Arabic handwritten text as well as gradient distribution features, windowed gradient distribution features, contour chain code distribution features, and windowed contour chain code distribution features are extracted. Nearest Neighbor (NN) classifier is used with the Euclidean distance measure. Data reduction algorithms (viz. PCA, LDA, MDA, MDS, and forward/backward feature selection algorithm) are used. A database of 250 Arabic handwritten paragraphs written by 250 writers is used. The used paragraphs were randomly generated from a large corpus. Each writer wrote 250 Arabic paragraphs that were split into two parts; one part for training and another for testing. NN provided the best accuracy in text-independent writer identification with top-1 result of 88.0%, top-5 result of 96.0%, and top-10 result of 98.5% for the first 100 writers. Extending the work to include all 250 writers and with the backward feature selection algorithm (using 54 out of 83 features); the system attained a top-1 result of 75.0%, top-5 result of 91.8%, and top-10 results of 95.4%.

6.1 Introduction

Writer identification can be divided into two categories; text-dependent and text-independent. Text-dependent writer identification systems require certain pre-defined text to be written, whereas text-independent writer identification systems can work on any given text. In this chapter, text-independent writer identification of offline Arabic handwritten text is addressed.

To the best of the author's knowledge, only few researchers have addressed writer identification and verification specifically for Arabic text, with all such efforts reported only in the last four years (Abdi, Khemakhem, & Ben-Abdallah, 2009; Al-Dmour & Zitar, 2007; Al-Ma'adeed, Al-Kurbi, Al-Muslih, Al-Qahtani, & Al Kubisi, 2008; Al-Ma'adeed, Mohammed, & Al Kassis, 2008; Bulacu, Schomaker, & Brink, 2007; Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010; Chawki & Labiba, 2010; Gazzah & Ben Amara, 2006; 2007; 2008; Lutf, Xinge You, & H. Li, 2010; S. Srihari & G. Ball, 2008).

It is worth noting that except for Bulacu et al. (Bulacu, Schomaker, & Brink, 2007) who used a database of 350 writers of city names and Lutf et al. (Lutf, Xinge You, & H. Li, 2010) who used diacritics samples from 287 writers writing city names, none of the previously mentioned research works reviewed in Chapter 3 used a database consisting of more than 150 writers of Arabic text. Furthermore, only few researchers applied writer identification on full Arabic paragraph text, e.g. (S. Srihari & G. Ball, 2008), who used a paragraph dataset of only 10 writers. Researchers in (Abdi, Khemakhem, & Ben-Abdallah, 2009; Bulacu, Schomaker, & Brink, 2007) synthetically

created a document text by combining city names in one page. To overcome some of these limitations, our research work applies writer identification on Arabic handwritten text of paragraphs from 250 different writers.

Bulacu et al. (Bulacu, Schomaker, & Brink, 2007), Abdi et al. (Abdi, Khemakhem, & Ben-Abdallah, 2009), Chawki and Labiba (Chawki & Labiba, 2010), Lutf et al. (Lutf, Xinge You, & H. Li, 2010), and Chaabouni et al. (Chaabouni, Boubaker, Kherallah, Alimi, & Haikal El Abed, 2010) used the IFN/ENIT dataset (or one of its variations) which is limited to Arabic town and city names (Pechwitz et al., 2002). Most noticeably, Bulacu et al. used 350 writers, with 5 pages per writer, each page containing 5 city names. They reported a top-1 accuracy of 88%. Abdi et al. used only 40 writers with each writer having more than 100 words. They reported 92.5% accuracy. Chawki and Labiba used 650 handwriting documents collected from 130 different writers where they reported a top-1 accuracy of 82.62%. It should be noted that Bulacu et al.'s work cannot be compared directly to Abdi et al.'s and Chawki and Labiba's work due to the large variance of the number of writers and the size of the used text. Srihari and Ball (S. Srihari & G. Ball, 2008) used macro- and micro-features along with likelihood ratio computation, and reported 86% accuracy on their limited dataset of 10 writers, each contributing 10 pages.

In this chapter, Section 2 presents a summarized description of the used features; the experimental results are detailed in Section 3; and finally, the conclusions are given in Section 4.

6.2 Features

In this work multiple types of features are used. Connected Component Features (CCF), Gradient Distribution Features (GDF), Windowed Gradient Distribution Features (WGDF), Contour Chain Code Distribution Features (C³DF), and Windowed Contour Chain Code Distribution Features (WC³DF) are implemented. A detailed description of these features is given in Chapter 4. A concise summary of these features is given below for ease of reference.

6.2.1 Connected Component Features (CCF)

Twenty one structural and statistical features are estimated. The resulting features include: The mean of acute, obtuse, and reflex angles for all connected components, the mean area of all the Aleph letters, the percentage of the Aleph letter, circles, half circles, and horizontal lines compared to the total number of segments, and the Histogram of -90, -30, 30, and 90 degrees for the Aleph letter, horizontal lines, half circles, and circles.

6.2.2 Gradient Distribution Features (GDF)

The gradient distribution features are computed by convolving the x and y Sobel operators with the binary image. The gradient angle of each pixel is calculated by computing the inverse tangent of ($\text{Gradient}_y/\text{Gradient}_x$). Subsequently, the histogram of the 24 gradient angles is calculated and stored as the feature vector. Since the histogram

of zero degrees contains little information, we only calculate the histogram of the remaining 23 angles into 23 different and distinct bins.

6.2.3 Windowed Gradient Distribution Features (WGDF)

The WGDF features are extracted using a 3 by 3 sliding window on the gradient angles' image. The most frequent angle in each sliding window is calculated for the windowed gradient distribution features. A total of 23 features are extracted for each image.

6.2.4 Contour Chain Code Distribution Features (C³DF)

The contour of the image is extracted and encoded using Freeman chain codes (Freeman, 1974). The histogram of each chain code is estimated and used as features. A total of 8 features are extracted for each image.

6.2.5 Windowed Contour Chain Code Distribution Features (WC³DF)

We employ the sliding window technique on the C³DF explained previously. Using a 3 by 3 sliding window, the most frequent chain code of each sliding window is chosen. A total of 8 features are extracted for each image.

Table 6-1 shows sample vectors for the five features type for the paragraph image shown in Figure 6-1 (b). These feature vectors are normalized to zero mean and a variance of 1 before being presented to the classifier.

Table 6-1: Sample feature vector for Figure 6-1 (a).

Feature	CCF	GDF	WGDF	C ³ DF	WC ³ DF
1	384	1252	436	5729	1924
2	7.4	1089	116	3360	972
3	130	6328	1459	4916	1840
4	120	1683	164	2186	661
5	0.1	1844	22	5411	1645
6	0.1	7934	1414	3509	1035
7	0.2	1444	166	4919	1733
8	0.1	1282	46	2024	477
9	0.1	9355	2188		
10	0.7	3049	202		
11	26	3217	15		
12	32	8146	1136		
13	19	1365	98		
14	15	1118	22		
15	11	5729	1386		
16	162	1523	194		
17	59	1794	17		
18	3	7953	1388		
19	28	1554	160		
20	10	1376	41		
21	0	10000	2106		
22		2952	476		
23		3114	16		

السبب الأساسي في هذه المشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات وأنشطة وحيوات الحياة الاجتماعية اليومية . يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين

السبب الأساسي في هذه المشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات و أنشطة وحيوات الحياة الاجتماعية اليومية. يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين

<p>مشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات و أنشطة وحيوات الحياة الاجتماعية اليومية. يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين</p>	<p>السبب الأساسي في هذه المشكلات إنما يعزى إلى فشل المجتمع في التسامح والتقبل للاختلافات و الفروق بين المعاقين من المشاركة العادية في فعاليات وحيوات الحياة الاجتماعية اليومية. يمكن إكساب ذوي الاحتياجات الخاصة مختلف المعارف والاتجاهات و القيم و المهارات التي تؤهلهم للمشاركة الإيجابية الفعالة في مختلف أنشطة وفعاليات الحياة الإنسانية إلى أقصى حد تؤهله لهم إمكانياتهم وقدراتهم إضافة إلى تغيير ثقافة المجتمع نحو المعاقين</p>
--	---

Figure 6-1: (a) Paragraph database sample image (top). (b),(c) Vertical division of sample image (bottom).

6.3 Experimental Results

The developed features are implemented on a database of 500 Arabic handwritten paragraphs written by 250 writers. The following sections explain the database used for writer identification, the different classifiers that are tested, different feature combination and dimensionality reduction techniques, the effect of increasing the number of writers, and a comparison with other published results in the literature.

6.3.1 Database Used

In order to create a text-independent dataset that contains the natural distribution of Arabic character shapes, we used a large corpus of Arabic texts from various fields (sports, medicine, news ...). An automated program randomly selects paragraphs from the corpus. These paragraphs are distributed to volunteers in a form. A sample page of the form is shown in Figure 6-1 (a). A detailed description of the used database can be found in Chapter 2.

A total of 250 volunteers filled the forms. In order to use the database for writer identification, we divided each image vertically into two samples. Dividing the images vertically instead of horizontally is done to have almost the same text distribution on each image. Figure 6-1 (b) and (c) shows the two segmented parts of the sample image shown in Figure 6-1 (a).

6.3.2 Classifier Selection

The Nearest Neighbor (NN) classifier classifies a new pattern by measuring its distance from the training samples and choosing the nearest sample to which the nearest neighbor belong (Friedman & Kandel, 1999). It is a simple yet effective classifier in the field of writer identification and verification as it can utilize various distance measures; i.e. the Euclidean distance measure, the Mahalanobis distance measure, and the chi-squared X^2 distance measure. The following variables are defined:

- $D_X(i)$ is the distance between the input sample (V) and model i using distance measure X .
- k is the number of features in the feature vector.
- M_{ij} is the j_{th} feature of model i , and V_j is the j_{th} feature of the input sample feature vector.
- C is the covariance matrix.

- $$\bar{V} = \frac{\sum_{j=1}^k V_j}{k} \text{ and } \bar{M}_l = \frac{\sum_{j=1}^k M_{lj}}{k}$$

By studying the most common distance measures used in the literature and reported in Chapter 3, the following distance measures are selected; the Euclidean distance measure, the City Block distance measure, the Chebychev distance measure (which measures the maximum coordinate difference), the Cosine distance measure (one minus the cosine of the angle between feature points), the Hamming distance measure (percentage of the coordinates that differ), the Mahalanobis distance measure, the

correlation distance measure, and the chi-squared distance measure (which is useful in histograms). The distance measure equations are given in Table 6-2, and more details can be found in (S.-H. Cha, 2007).

For each sample image, all distance measures are estimated between this sample and the remaining 499 images. Top-1, top-5, and top-10 classes of the nearest samples are obtained and are considered as the writer classes that match most closely the obtained features vector of the unknown writers. Testing different distance measures, the Euclidean distance measure provided the best results, and these results are statistically significant over other distance measures. The Hamming distance measure reported the worst accuracy results, which is expected since it is best suited for binary features and not for real values. Using all features described in section 6.2, a summary of the results for the various distance measures is given in Table 6-3.

Support Vector Machines (SVM) and Hidden Markov Models (HMM) are tried on this dataset with no significant improvement in accuracy. This is in accordance with reports of achieving better results using the k-NN classifier over the SVM classifier (Schlapbach, 2007). The large number of classes (i.e. writers) and few samples per class (i.e. paragraphs per writer) common in writer identification tasks are frequently not sufficient for statistical classifiers like SVM and HMM to perform proper training that is needed for efficient identification.

Table 6-2: Distance measures equations.

Euclidean	$D_{Euc}(i) = \sqrt{\sum_{j=1}^k (V_j - M_{ij})^2}$
City Block	$D_{CB}(i) = \sum_{j=1}^k V_j - M_{ij} $
Chebychev	$D_{Cheb}(i) = \max_j V_j - M_{ij} $
Cosine	$D_{Cos}(i) = \frac{\sum_{j=1}^k V_j M_{ij}}{\sqrt{\sum_{j=1}^k V_j^2} \sqrt{\sum_{j=1}^k M_{ij}^2}}$
Hamming	$D_{Ham}(i) = \frac{\text{Number of } (V_j \neq M_{ij})}{k}$
Mahalanobis	$D_{Mah}(i) = (V - M_i)C^{-1}(V - M_i)'$
Correlation	$D_{Corr}(i) = \frac{(V - \bar{V})(M_i - \bar{M}_i)'}{\sqrt{(V - \bar{V})(V - \bar{V})'} \sqrt{(M_i - \bar{M}_i)(M_i - \bar{M}_i)'}}$
Chi-Squared (X^2)	$D_{SqChi}(i) = \sum_{j=1}^k \frac{(V_j - M_{ij})^2}{(V_j + M_{ij})}$

Table 6-3: Distance measures results on writer identification for 100 writers.

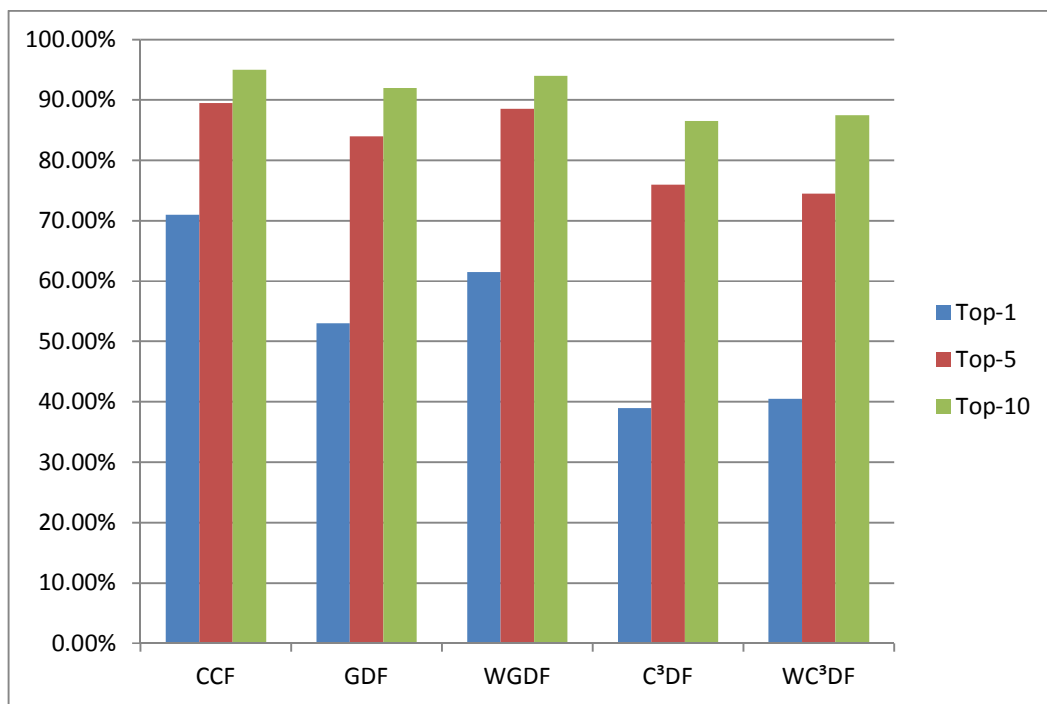
Distance Measure	Top – 1	Top – 5	Top – 10
Euclidean	83.5%	96.0%	97.5%
Cosine	77.5%	92.5%	95.5%
City Block	76.0%	95.5%	98.0%
Correlation	73.0%	90.0%	94.5%
Chebychev	53.0%	81.5%	86.5%
Squared X^2	41.5%	84.0%	90.0%
Mahalanobis	30.0%	47.5%	59.0%
Hamming	8.0%	22.5%	32.5%
Significance	6.03%	2.44%	1.73%

6.3.3 Feature Combination

In order to decide on the best feature combination, a randomly selected subset of 100 writers is used. This subset is also used to select the best dimensionality reduction technique. Figure 6-2 shows the writer identification recognition accuracy of the five types of implemented features implemented features for the first 100 writers. The statistical significance measure is shown for the top-1, top-5, and top-10 results. A 95% level of confidence is used. Results that statistically have the same significance are highlighted. Readers are referred to Chapter 5 for the derivation and explanation of the statistical significance measure.

The CCF feature vector provided the best top-1 accuracy of 71%, with the next-best feature vector (WGDF) almost 10% less accurate for the top-1 accuracy. The CCF features accuracy results are statistically significant compared to other features (albeit for the WGDF features in the top-5 and top-10). The windowed version of the remaining features (WGDF, WC³DF) attained better accuracy than their non-windowed counterparts (GDF, C³DF).

In order to combine the discriminatory power of each feature, the extracted features for each type are concatenated to form one feature vector. This resulted in a feature vector length of 83. Top-1 accuracy results for 100 writers are 83.5%, top-5 accuracy results are 96.0%, and top-10 results are 97.5%.



	CCF	GDF	WGDF	C³DF	WC³DF	Significance
Top-1	71.00%	53.00%	61.50%	39.00%	40.50%	8.02%
Top-5	89.50%	84.00%	88.50%	76.00%	74.50%	4.63%
Top-10	95.00%	92.00%	94.00%	86.50%	87.50%	2.85%

Figure 6-2: Writer identification recognition accuracy for different features (100 writers).

6.3.4 Dimensionality Reduction

Several feature reduction techniques are applied to reduce the dimensionality of the features and improve results. The methods considered are Principal Component Analysis (PCA) (Duda, Hart, & Stork, 2000), Linear Discriminant Analysis (LDA) (Duda et al., 2000), Multiple Discriminant Analysis (MDA) (Duda et al., 2000), Multi-Dimensional Scaling (MDS) (Agrafiotis, 2003), and sequential forward and backward feature selection techniques (Theodoridis & Koutroumbas, 2009). Results summarized in Table 6-4 show that the backward subset selection algorithm attained the highest accuracy results and with the same significance level as that of only the forward search and all features together. Using backward search also resulted in 35% reduction in number of features.

The backward search starts by using all features and then gradually removes features if this removal improves identification accuracy, while the forward search starts from the best single feature and then adds to it features that best improve results. Both methods are greedy techniques. Most of the other unsupervised techniques like PCA, MDS, and MDA gave inferior accuracy, yet they drastically reduced the dimensionality of the feature vector.

Table 6-4: Feature selection results on writer identification for 100 writers.

Subset Method	Top – 1	Top – 5	Top – 10	Number of features
All Features	83.5 %	96.0 %	97.5 %	83
PCA	54.0 %	79.0 %	87.5 %	10
MDA	52.5 %	82.5 %	91.5 %	15
MDS	52.0 %	79.5 %	87.5 %	20
LDA	71.0 %	91.5 %	96.5 %	40
Forward Search	84.5 %	96.5 %	98.5 %	57
Backward Search	88.0 %	96.0 %	98.5 %	54
Significance	5.02%	2.44%	1.16%	

6.3.5 **Writer Identification Accuracy vs. Number of Writers**

Using the selected features for 100 writers, we applied it to all 250 writers. This resulted in top-1 accuracy of 75%, top-5 accuracy of 91.8%, and top-10 accuracy of 95.4%. Figure 6-3 shows the results of top-1, top-5, and top-10 accuracy percentage vs. the number of writers ranging from 10 to 250 writers. Results indicate a decrease of 20% in top-1 accuracy, but only 8.2% and 4.6% decrease in top-5 and top-10 accuracy results, respectively for the 250 writers compared with 10 writers. Schomaker et al. (Schomaker & Bulacu, 2004) reported that “the target performance indicated by forensic experts would be 99 percent probability of finding the correct writer in the top-100 hit list, on a database of 20,000 samples.” We expect that the developed system for Arabic writer identification may provide comparable results when applied on such a large database. A reject criterion based on the difference of Euclidean distance between top-1 and top-2 result is implemented. The criterion would reject the sample if the difference of distance between top-1 and top-2 result is less than a threshold, thus taking into account doubtful samples. Applying this criterion resulted in top-1, top-5, and top-10 accuracy results of 80.5%, 94.0%, and 97.0%, respectively, with a rejection rate of 20% of the 250 writers.

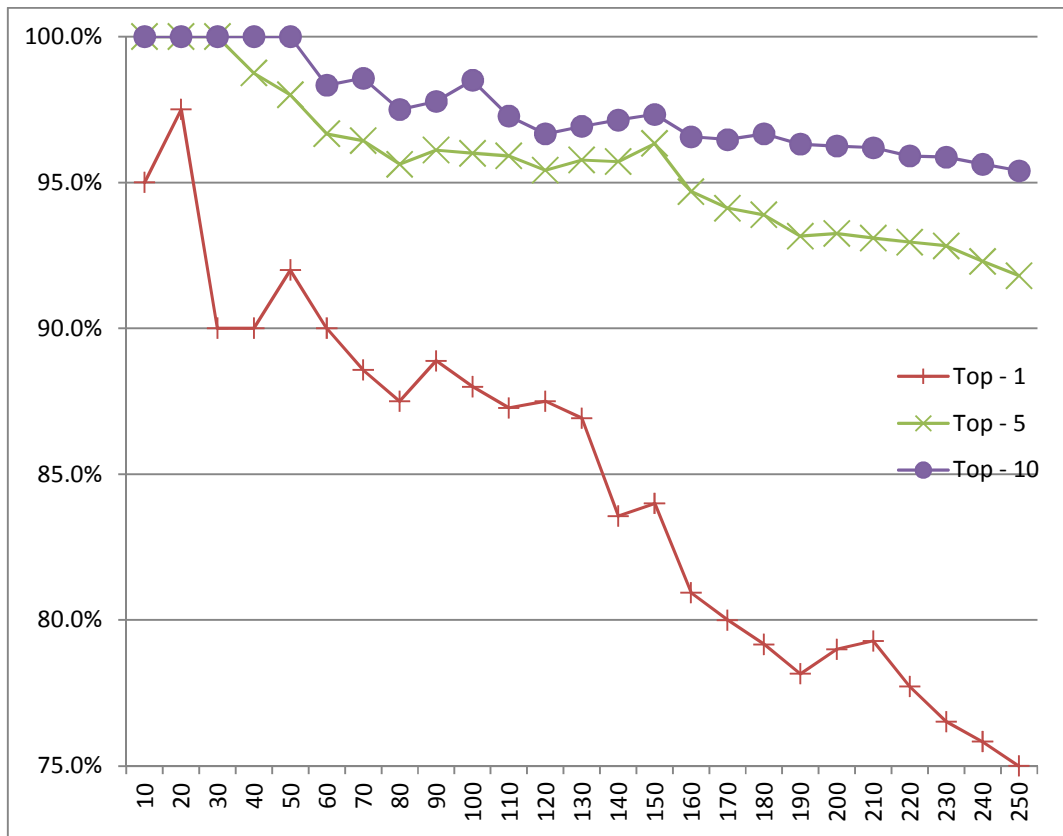


Figure 6-3: Top - 1, top - 5, and top - 10 accuracy % vs. the as dsa number of writers.

6.3.6 Comparison with Other Published Results

Our results are compared with the published research on writer identification having comparable number of writers. Table 6-5 shows the language, the number of writers, and the accuracy of other writer identification systems (top-1, top-5, and top-10). The results of our technique are comparable with other researchers, noting that not all researchers posted top-5 and top-10 results. The top-1 accuracy of Zhang (B. Zhang, 2003) is exceptional. However, his system was tested on a text-dependent database with manually extracted words for result improvements. Bulacu et al. (Bulacu, Schomaker, & Brink, 2007) worked on city names and not full text-independent Arabic paragraphs. Furthermore, they reported that the results obtained on Arabic are somewhat lower than the ones obtained on Western script, and that writer identification on Arabic script appears to be more difficult than its Latin counterpart.

Table 6-5: Comparison of writer identification systems.

Authors	Language	# of Writers	Top-1	Top-5	Top-10
Schomaker et al. (Schomaker, Bulacu, & M. van Erp, 2003)	Dutch	250	88.0%	98.0%	99.0%
Zhang et al. (B. Zhang, 2003)	English	1000	98.0%	-	-
Bulacuet al. (Bulacu & Schomaker, 2005)	Dutch	250	78.0%	-	92.6%
Bulacu et al. (Bulacu, Schomaker, & Brink, 2007)	Arabic	350	88.0%	-	99.0%
He et al. (He, X. You, & Tang, 2008b)	Chinese	500	39.2%	62.4%	77.2%
Siddiqi et al. (He, X. You, & Tang, 2008b)	English	200	86.0%	-	-
Our approach	Arabic	100	88.0%	96.0%	98.5%
Our approach	Arabic	250	75.0%	91.8%	95.4%
Our approach (with rejection)	Arabic	250	80.5%	94.0%	97.0%

6.4 Conclusions

The presented work addresses the writer identification of Arabic handwritten text. Several types of features are extracted (viz. connected component features, gradient distribution features, contour chain code distribution features and their windowed variations) from the handwriting samples. The connected component features are a combination of structural and statistical features that are extracted by studying common characteristics of the Arabic text.

A Nearest Neighbor classifier is utilized with the Euclidean distance measure. A database of 500 Arabic handwritten paragraphs written by 250 writers is used for analysis and experimentations. The Accuracy results for different number of writers are presented and analyzed. Several feature selection algorithms are used (viz. PCA, LDA, MDA, MDS, and backward/forward selection algorithms).

The Euclidean distance provided the best accuracy in text-independent writer identification for all 250 writers with top-1 result of 75.0%, top-5 result of 91.8%, and top-10 results of 95.4%. A reject criterion based on the difference of Euclidean distance between top-1 and top-2 result is implemented that improved the top-1 results by almost 6%. These encouraging results demonstrate the writer discriminability of Arabic text for writer identification. Finally, the developed system is compared to other Latin and Arabic writer identification systems with similar-sized databases in the literature.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this chapter we summarize the contributions of the thesis and analyze the state of the art in writer identification of Arabic text. Conclusions relevant to writer identification of Arabic text are discussed and future directions stated. In addition, we discuss future directions in writer identification & verification of Arabic text.

7.1 Conclusions

The main contribution in this thesis is the design and development of successful writer identification systems for handwritten Arabic digits and text. The second contribution is the design of statistical and structural features that are more adapted to the nature of the Arabic language. The design and implementation of a natural handwritten text Arabic database for a large number of writers in Chapter 3 can play a vital role in benchmarking between different writer identification systems and fills a gap in this area. A comprehensive literature survey for writer identification of Arabic and Latin and western languages has been needed in the research community, and Chapter 4 has satisfied this need. These contributions and others should help in advancing the field of writer identification and verification in general and for Arabic documents in particular.

Chapter 2 has shown that for Arabic, most of the databases are researcher generated for their own research with the exception of the IEF/ENIT database which consists of city names. So far there is no Arabic text database that is freely available for writer identification of Arabic text. The database of 250 writers that was developed as part of this work will be provided publicly for researchers.

In chapter 3, comparing the work on Arabic text with Latin indicates that a limited number of researchers are involved in writer identification of Arabic text. In addition, comparing features and classification approaches indicates that most of the previous work on Arabic text is based on features and classifiers used for English. It is clear that published work using Arabic text has reported lower writer identification accuracies than that using Latin text. We cannot conclude that Arabic text is less identifiable than Latin text although the general understanding is that Latin text gives better identification rates. We believe that one of the factors for reduced accuracy is that data used for Arabic text writer identification does not match in representation and naturalness that of the databases of Latin text. The database developed as part of this work will help alleviate this problem.

Chapter 4 described the development and design of several types of features (*viz.* connected component features, gradient distribution features, contour chain code distribution features, density features, horizontal and vertical run lengths features, stroke features, and concavity features) from the sample image. It also featured a novel approach to extract a combination of structural and statistical features that build on some of the main characteristics of the Arabic language. The chapter included a detailed description of these features along with figures indicating their discriminability power.

Writer identifiability for handwritten digits discussed in chapter 5 provided quantitative measurements for the discriminability of writers for each digit. These measurements have supported the common perception such as digits that contain few inter-writer difference like digit 0 ‘٠’ give poorer writer identification performance than digits that have more inter-writer differences, e.g. digit 3 ‘٣’. The research work has also shown that we can design a successful writer identification system from only the basic 10 Arabic digits given that other writers write the same digits.

In chapter 6, Writer identification using Arabic handwritten text was addressed, where several types of structural and statistical features were extracted. Writer identification performance using each feature type was compared using statistical significance. The effects of increasing the number of writers on the accuracy results were presented and analyzed. A reject criterion is implemented which improved top-1 results by 6%.

7.2 Future Directions

Most of the features previously applied in the literature on the identification and verification of Arabic text are features used originally for Latin or a modified version of them. In our effort to address this point we used connected components of Arabic text that takes the characteristic of the Arabic script in consideration.

More novel features that consider the intrinsic properties of the Arabic script are needed to improve the accuracy of writer identification and verification of Arabic documents. For example, diacritics can be used, which can prove to be valuable features

in this regard. Features based on diacritics include distance between diacritic marks and the main text, average size of diacritic marks, etc. In addition, inter-writer variance methods of writing the dots (as a stroke or as disconnected dots), location of the dots, and slant of the dots are other examples of features that utilize some of the known differences between writers while writing the Arabic script. Many variations between writers exist while writing Arabic, and a study of these variations should be utilized for writer identification and verification of Arabic documents.

Building on the success of the connected component features designed in this thesis, we hope that other researchers of writer identification and verification of Arabic text develop novel features that take the characteristics of Arabic text into consideration. Researchers have indicated that techniques used for Latin give lower rates when applied to Arabic due to some characteristics of the language.

There is a need for an Arabic text database with large number of writers for writer identification and verification. The designed database in Chapter 2 can serve as a basis for future Arabic handwritten database design considerations. In addition, conducting research competitions on the field of Arabic handwritten writer identification can help in advancing the field and comparing different research efforts.

Research for writer verification of handwritten Arabic text is very scarce. Researchers can use the developed database in Chapter 2 to develop writer verification systems. In addition, researchers can develop collect forged copies of the written paragraphs to further extend the work.

Finally, establishing research groups for Arabic text recognition and identification is essential for the prosperity of the field. This will enable building resources that the

research community can utilize. We hope that this thesis on writer identification of Arabic text with the survey on previous work will encourage more research contributions to this field.

NOMENCLATURE

ANN	Artificial Neural Network
C ³ DF	Contour Chain Code Distribution Features
CCF	Connected Component Features
CDF	Cumulative Distribution Functions
GDF	Gradient Distribution Features
GMM	Gaussian Mixture Models
GSC	Gradient, Structural and Concavity features
HMM	Hidden Markov Model
HMT	Hidden Markov Tree
KLD	Kull back Leibler Distance
k-NN	k Nearest Neighbours
LDA	Linear Discriminant Analysis
LDC	Linear Discriminant Classifier
LDC	Linguistic Data Consortium
LLR	Log-Likelihood Ratio
LOB	Lancaster-Oslo/Bergen corpus
MDA	Multiple Discriminant Analysis
MDS	Multi-Dimensional Scaling
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
NM	Nearest Mean
NN	Nearest Neighbor
OCR	Optical Character Recognition
PAW	Part of Arabic Word
PCA	Principal Component Analysis
PDF	Probability Distribution Function
ROI	Region Of Interest

SVM	Support Vector Machine
WC ³ DF	Windowed Contour Chain Code Distribution Features
WGDF	Windowed Gradient Distribution Features

REFERENCES

- Abdi, M. N., Khemakhem, M., & Ben-Abdallah, H. (2009). A Novel Approach for Off-Line Arabic Writer Identification Based on Stroke Feature Combination. *24th International Symposium on Computer and Information Sciences* (pp. 597-600). IEEE.
- Abdleazeem, S., & El-Sherif, E. (2008). Arabic handwritten digit recognition. *International Journal on Document Analysis and Recognition*, *11*(3), 127-141. doi: 10.1007/s10032-008-0073-5.
- Agrafiotis, D. K. (2003). Stochastic Proximity Embedding. *Journal of Computational Chemistry*, *24*(10), 1215-21.
- Al-Badr, B., & Mahmoud, S. (1995). Survey and bibliography of Arabic optical text recognition. *Signal Processing*, *41*(1), 49-77.
- Al-Dmour, A., & Zitar, R. A. (2007). Arabic Writer Identification Based on Hybrid Spectral-Statistical Measures. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(4), 307-332.
- Al-Ma'adeed, S., Al-Kurbi, A.-A., Al-Muslih, A., Al-Qahtani, R., & Al Kubisi, H. (2008). Writer Identification of Arabic Handwriting Documents Using Grapheme Features. *IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2008* (pp. 923-924).
- Al-Ma'adeed, S., Elliman, D., & Higgins, C. A. (2002). A data base for Arabic handwritten text recognition research. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition* (pp. 485-489).
- Al-Ma'adeed, S., Mohammed, E., & Al Kassis, D. (2008). Writer Identification Using Edge-Based Directional Probability Distribution Features for Arabic Words. *IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008)* (pp. 582-590).
- Al-Ohali, Y., Cheriet, M., B, M., & Suen, C. (2003). Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, *36*(1), 111-121.
- Amin, A. (1998). Off-line Arabic character recognition the state of the art. *Pattern Recognition*, *31*(5), 517-530.
- Awaidah, S. M., & Mahmoud, S. A. (2009). A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. *Signal Processing*, *89*(6), 1176-1184.

- Ball, G. R., & Srihari, Sargur N. (2008). Writer adaptation in off-line Arabic handwriting recognition. *Document Recognition and Retrieval* (Vol. 6815, p. 4).
- Bar-Yosef, I., Beckman, I., Kedem, K., & Dinstein, I. (2007). Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. *International Journal on Document Analysis and Recognition*, 9(2), 89.
- Bensefia, A., Nosary, A., Paquet, T., & Heutte, L. (2002). Writer identification by writer's invariants. *Eighth International Workshop on Frontiers in Handwriting Recognition* (pp. 274-279). Ontario, Canada: IEEE Comput. Soc.
- Bensefia, A., Paquet, T., & Heutte, L. (2003a). Information Retrieval Based Writer Identification. *Seventh International Conference on Document Analysis and Recognition* (p. 946). Edinburgh, Scotland: IEEE Computer Society.
- Bensefia, A., Paquet, T., & Heutte, L. (2003b). Grapheme based writer verification. *11th Conference of the International Graphonomics Society (IGS2003)* (pp. 274-277). Arizona, USA: in Proceedings of the 11th Conference of the International Graphonomics Society (IGS2003).
- Bensefia, A., Paquet, T., & Heutte, L. (2004). Handwriting Analysis for Writer Verification. *Ninth International Workshop on Frontiers in Handwriting Recognition* (pp. 196-201). Tokyo, Japan: IEEE.
- Bensefia, A., Paquet, T., & Heutte, L. (2005). Handwritten Document Analysis for Automatic Writer Recognition. *Electronic Letters on Computer Vision and Image Analysis*, 5(2), 72-86.
- Bensefia, A., Paquet, T., Thierry, & Heutte, Laurent. (2005). A writer identification and verification system. *Pattern Recognition Letters*, 26(13), 2080-2092.
- Bradford, R. R., & Bradford, R. B. (1992). *Introduction to Handwriting Examination and Identification*. Burnham, Inc.
- Brink, A., Bulacu, M., & Schomaker, L. (2008). How Much Handwritten Text Is Needed for Text-Independent Writer Verification and Identification. *19th International Conference on Pattern Recognition, ICPR* (pp. 1-4).
- Brink, A., Niels, R., Batenburg, R. A. van, Den Heuvel, C. E. van, & Schomaker, L. (2010). Towards Robust Writer Verification by Correcting Unnatural Slant. *Pattern Recognition Letters*, 32(3), 449-457.
- Brown (ed.), K. (2006). *Encyclopedia of Language & Linguistics (Second Edition)*, 14-Volume Set. Elsevier.

- Buckwalter, T. (2002). Arabic Word Frequency Counts. Retrieved February 4, 2011, from www.qamus.org/wordlist.htm.
- Bulacu, M. (2007). Statistical Pattern Recognition for Automatic Writer Identification and Verification. *Ph.D. dissertation, Dept. of Behav. and Soc. Sci., Univ. of Groningen, Netherlands*, 140. Netherlands.
- Bulacu, M., Koert, R. van, Schomaker, L., & Der Zant, T. van. (2007). Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 1, pp. 357-361).
- Bulacu, M., & Schomaker, L. (2005). A Comparison of Clustering Methods for Writer Identification and Verification. *Eighth International Conference on Document Analysis and Recognition (ICDAR '05)* (pp. 1275-1279 Vol. 2).
- Bulacu, M., & Schomaker, L. (2006). Combining Multiple Features for Text-Independent Writer Identification and Verification. In Guy Lorette (Ed.), *In Proc. of 10th IWFHR* (p. 281–286). La Baule (France): Suvisoft.
- Bulacu, M., & Schomaker, L. (2007a). Text-Independent Writer Identification and Verification Using Textural and Allographic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4), 701-717.
- Bulacu, M., & Schomaker, L. (2007b). Automatic Handwriting Identification on Medieval Documents. *14th International Conference on Image Analysis and Processing* (pp. 279-284).
- Bulacu, M., Schomaker, L., & Brink, A. (2007). Text-Independent Writer Identification and Verification on Offline Arabic Handwriting. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 769-773).
- Bulacu, M., Schomaker, L., & Vuurpijl, L. (2003). Writer Identification Using Edge-Based Directional Features. *ICDAR '2003: International Conference on Document Analysis and Recognition* (p. 937). IEEE Computer Society.
- Central Intelligence Agency, C. (2003). The World Factbook. *ISSN: 1553-8133*. Retrieved October 4, 2011, from <https://www.cia.gov/library/publications/download/download-2003/index.html>.
- Cha, S., & Srihari, S. (2000). Assessing the Authorship Confidence of Handwritten Items. *Fifth IEEE Workshop on Applications of Computer Vision (WACV'00)* (pp. 42-47). California, USA.

- Cha, S. H. (2001). Use of distance measures in handwriting analysis. *Ph.D. dissertation, Dept. of Comp. Sci. and Eng., St. Univ. of N.Y at Buffalo, NY*, 239. Retrieved January 30, 2010, from <http://portal.acm.org/citation.cfm?id=933434>.
- Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. Retrieved November 29, 2010, from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.8446>.
- Chaabouni, A., Boubaker, H., Kherallah, M., Alimi, A. M., & Abed, Haikal El. (2010). Fractal and Multi-fractal for Arabic Offline Writer Identification. *20th International Conference on Pattern Recognition* (pp. 3793-3796). IEEE. doi: 10.1109/ICPR.2010.924.
- Chawki, D., & Labiba, S.-M. (2010). A Texture Based Approach for Arabic Writer Identification and Verification. *2010 International Conference on Machine and Web Intelligence* (pp. 115-120). IEEE.
- Cong, S., Xiao-Gang, R., & Tian-Lu, M. (2002). Writer identification using Gabor wavelet. *Proceedings of the 4th World Congress on Intelligent Control and Automation (Cat. No.02EX527)* (pp. 2061-2064). IEEE. doi: 10.1109/WCICA.2002.1021447.
- Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)* (pp. 886-893). IEEE.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2nd Edition)* (p. 654). Wiley-Interscience.
- El Abed, H., & Märgner, V. (2007). The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems (pp. 1-4).
- Favata, J., & Srikantan, G. (1996). A multiple feature/resolution approach to handprinted digit and character recognition. *International Journal of Imaging Systems and Technology*, 7(4), 311, 304.
- Fornes, A., Lladós, J., Sanchez, G., & Bunke, H. (2008). Writer Identification in Old Handwritten Music Scores. *The Eighth IAPR International Workshop on Document Analysis Systems, DAS '08* (pp. 347-353). doi: 10.1109/DAS.2008.29.
- Franke, Katrin, Bünнемeyer, O., & Sy, T. (2002). Ink Texture Analysis for Writer Identification. *Proceedings. Eighth International Workshop on Frontiers in Handwriting Recognition* (p. 268). Los Alamitos, CA, USA: IEEE Computer Society. doi: <http://doi.ieeecomputersociety.org/10.1109/IWFHR.2002.1030921>.

- Franke, Katrin, & Koppen, M. (2001). A Computer-Based System to Support Forensic Studies on Handwritten Documents. *International Journal on Document Analysis and Recognition*, 3(4), 218-231.
- Franke, Katrin, Schomaker, L., Veenhuis, C., Taubenheim, C., Guyon, Isabelle, Vuurpijl, L., et al. (2003). WANDA: a generic framework applied in forensic handwriting analysis and writer identification. *Design and application of hybrid intelligent systems. Proceedings 3rd International Conference on Hybrid Intelligent Systems (HIS03)* (pp. 927-938). IOS Press.
- Freeman, H. (1974). Computer Processing of Line-Drawing Images. *ACM Computing Surveys*, 6(1), 57-97. New York: ACM.
- Friedman, M., & Kandel, A. (1999). *Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches* (p. 329). World Scientific Publishing Company.
- Gazzah, S., & Ben Amara, N. (2006). Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection. In J. Wang, Z. Yi, J. Zurada, B.-L. Lu, & H. Yin (Eds.), *Advances in Neural Networks - ISNN 2006* (Vol. 3972, pp. 271-276). Springer Berlin / Heidelberg.
- Gazzah, S., & Ben Amara, N. (2007). Arabic Handwriting Texture Analysis for Writer Identification Using the DWT-Lifting Scheme. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 1133-1137).
- Gazzah, S., & Ben Amara, N. (2008). Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script. *The International Arab Journal of Information Technology*, 5(1), 13-75.
- Gibbons, M., Yoon, S., Cha, S.-H., & Tappert, C. (2005). Evaluation of Biometric Identification in Open Systems. In T. Kanade, A. Jain, & N. Ratha (Eds.), *Audio- and Video-Based Biometric Person Authentication* (Vol. 3546, pp. 823-831). Springer Berlin / Heidelberg.
- Gonzalez, R. C., & Woods, R. E. (2007). *Digital Image Processing (3rd Edition)* (p. 976). Prentice Hall.
- Graff, D. (2007). *Arabic Gigaword third edition - Published as LDC publication: LDC2007T40*. Philadelphia, USA.
- Grosicki, E., Carré, M., Brodin, J.-M., & Geoffrois, E. (2008). RIMES Evaluation Campaign for Handwritten Mail Processing. *Proceedings 11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Montreal, Canada.

- Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., & Janet, S. (1994). UNIPEN project of on-line data exchange and recognizer benchmarks. *12th IAPR International Conference on Pattern Recognition* (pp. 29-33). Jerusalem, Israel: IEEE Comput. Soc. Press.
- He, Z., Bin, F., Jianwei, D., Yuan Yan, T., & Xinge, Y. (2005). A Novel Method for Off-line Handwriting-based Writer Identification. *Eighth International Conference on Document Analysis and Recognition (ICDAR '05)* (pp. 242-246). IEEE.
- He, Z., & Tang, Y. (2004). Chinese handwriting-based writer identification by texture analysis. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics* (pp. 3488-3491). IEEE.
- He, Z., Tang, Y., & You, X. (2005). A contourlet-based method for writer identification. *2005 IEEE International Conference on Systems, Man and Cybernetics* (Vol. 1, pp. 364-368 Vol. 1).
- He, Z., You, X., & Tang, Y. (2008a). Writer identification of Chinese handwriting documents using hidden Markov tree model. *Pattern Recognition*, 41(4), 1295-1307.
- He, Z., You, X., & Tang, Y. (2008b). Writer Identification Using Global Wavelet-Based Features. *Neurocomputing*, 71(10-12), 1832-1841.
- Helli, B., & Moghaddam, M. E. (2008a). Persian Writer Identification Using Extended Gabor Filter. In A. Campilho & M. Kamel (Eds.), *Image Analysis and Recognition* (Vol. 5112, pp. 579-586). Springer Berlin / Heidelberg.
- Helli, B., & Moghaddam, M. E. (2008b). A Text-Independent Persian Writer Identification System Using LCS Based Classifier. *2008 IEEE International Symposium on Signal Processing and Information Technology* (pp. 203-206). IEEE.
- Helli, B., & Moghaddam, M. E. (2009). A Writer Identification Method Based on XGabor and LCS. *IEICE Electronics Express*, 6(10), 623-629.
- Helli, B., & Moghaddam, M. E. (2010). A Text-Independent Persian Writer Identification Based on Feature Relation Graph (FRG). *Pattern Recognition*, 43(6), 2199-2209.
- Hochberg, J., Bowers, K., Cannon, M., & Kelly, P. (1999). Script and Language Identification for Handwritten Document Images. *International Journal on Document Analysis and Recognition*, 2(2), 45-52. Springer Berlin / Heidelberg.
- Huber, R. A., & Headrick, A. M. (1999). *Handwriting Identification: Facts and Fundamentals* (p. 456). CRC Press.
- Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5), 550-554.

- Int. Unipen Foundation. (2011). Int. Unipen Foundation - iUF. Retrieved 2011, from <http://unipen.org/products.html>.
- Johansson, S., Leech, G. N., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*. Department of English, University of Oslo, Oslo.
- K. D. (2007). Questioned Document Examination using CEDAR-FOX. *Journal of Forensic Document Examination*, 18(2), 1-20.
- Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd ed., p. 312). New York, NY, USA: Springer-Verlag New York, Inc.
- Leedham, G., & Chachra, S. (2003). Writer Identification Using Innovative Binarised Features of Handwritten Numerals. *Proceedings. Seventh International Conference on Document Analysis and Recognition* (pp. 413-416 vol.1).
- Li, X., & Ding, X. (2009). Writer Identification of Chinese Handwriting Using Grid Microstructure Feature. In M. Tistarelli & M. Nixon (Eds.), *Advances in Biometrics* (Vol. 5558, pp. 1230-1239). Springer Berlin / Heidelberg.
- Li, X., Wang, X., & Ding, X. (2006). An Off-line Chinese Writer Retrieval System Based on Text-sensitive Writer Identification. *18th International Conference on Pattern Recognition, 2006. ICPR 2006*. (pp. 517-520). IEEE Computer Society.
- Liu, C.-L., Dai, R.-W., & Liu, Y.-J. (1995). Extracting individual features from moments for Chinese writer identification. *Proceedings of the Third International Conference on Document Analysis and Recognition* (Vol. 1, pp. 438-441 vol.1). Los Alamitos, CA, USA: IEEE Computer Society.
- Lowe, D. G. (1999). Object Recognition From Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision* (pp. 1150-1157 vol.2). IEEE.
- Lutf, M., You, Xinge, & Li, H. (2010). Offline Arabic Handwriting Identification Using Language Diacritics. *20th International Conference on Pattern Recognition* (pp. 1912-1915). IEEE.
- Maaten, L., & Postma, E. (2005). Improving automatic writer identification. *Proc. of 17th Belgium-Netherlands Conference on Artificial Intelligence* (pp. 260-266). Proc. of 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2005).
- Mahmoud, Sabri, & Awaida, S. (2009). Recognition of Off-Line Handwritten Arabic (Indian) Numerals Using Multi-Scale Features and Support Vector Machines vs. Hidden Markov Models. *Arabian Journal for Science and Engineering (AJSE)*, 34, 429-444.

- Mar, S. H., & Thein, N. L. (2005). Myanmar Character Identification of Handwriting Between Exhibit and Specimen. *Proceedings. 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, APSITT 2005* (pp. 95-98).
- Margner, V., & Pechwitz, M. (2001). Synthetic data for Arabic OCR system development. *Proceedings of Sixth International Conference on Document Analysis and Recognition* (pp. 1159-1163). IEEE Comput. Soc. doi: 10.1109/ICDAR.2001.953967.
- Marti, U.-V., & Bunke, H. (2002). The IAM-Database: an English Sentence Database for Offline Handwriting Recognition. *International Journal on Document Analysis and Recognition*, 5(1), 39-46.
- Matsuura, T., & Qiao, Y. (1989). Writer identification using an impulse response of the system characterizing handwriting motion. *IEE Colloquium on Character Recognition and Applications* (pp. 2/1-2/8).
- Märgner, V., & El Abed, H. (2007). ICDAR 2007 - Arabic Handwriting Recognition Competition. *9th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1274 - 1278). Curitiba - Paraná - Brazil.
- Märgner, V., & El Abed, H. (2009). ICDAR 2009 Arabic Handwriting Recognition Competition. *10th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1383-1387). IEEE.
- Märgner, V., & El Abed, H. (2010). ICFHR 2010 - Arabic Handwriting Recognition Competition. *12th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 709-714). IEEE.
- Märgner, V., & El Abed, H. (2011). ICDAR 2011 Arabic Handwriting Recognition Competition. *To appear in the 11th International Conference on Document Analysis and Recognition (ICDAR)*. Beijing, China.
- Märgner, V., Pechwitz, M., & Abed, H.E. (2005). ICDAR 2005 Arabic handwriting recognition competition. *Eighth International Conference on Document Analysis and Recognition* (pp. 70-74 Vol. 1).
- Niels, R., Vuurpijl, L., & Schomaker, L. (2007). Automatic allograph matching in forensic writer identification. *International Journal of Pattern Recognition and Artificial Intelligence*, 21, 61-81.
- Panagopoulos, M., Papaodysseus, C., Rousopoulos, P., Dafi, D., & Tracy, S. (2009). Automatic Writer Identification of Ancient Greek Inscriptions. *IEEE transactions on pattern analysis and machine intelligence*, 31(8), 1404-14. Published by the IEEE Computer Society.

- Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., & Amiri, H. (2002). IFN/ENIT - Database of Handwritten Arabic Words. *7th Colloque International Francophone sur l'Écrit et le Document , CIFED 2002* (p. 129--136). Hammamet, Tunis.
- Plamondon, R. (1994). *Progress in Automatic Signature Verification* (p. 180). River Edge, NJ, USA: World Scientific Publishing Co., Inc.
- Plamondon, R., & Lorette, G. (1989). Automatic Signature Verification and Writer Identification — The State of the art. *Pattern Recognition*, 22(2), 107-131.
- Plamondon, R., & Srihari, S.N. (2000). Online and Off-Line Handwriting Recognition: a Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 63-84.
- Plötz, T. (2005). *Advanced stochastic protein sequence analysis*. Bielefeld University.
- Ram, S. S., & Moghaddam, M. E. (2009a). Text-independent Persian Writer Identification Using Fuzzy Clustering Approach. *2009 International Conference on Information Management and Engineering* (pp. 728-731). IEEE.
- Ram, S. S., & Moghaddam, M. E. (2009b). A Persian Writer Identification Method Based on Gradient Features and Neural Networks. *2009 2nd International Congress on Image and Signal Processing* (pp. 1-4). IEEE.
- Rey, A., Ziegler, J. C., & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *International Journal of Cognitive Science*, 75(1), B1 - B12.
- Rodríguez-Serrano, J. A., Perronnin, F., Sánchez, G., & Lladós, J. (2010). Unsupervised Writer Adaptation of Whole-Word HMMs with Application to Word-Spotting. *Pattern Recognition Letters*, 31(8), 742-749.
- Said, H. E. S., Baker, K. D., & Tan, T. (1998). Personal identification based on handwriting. *Proceedings. Fourteenth International Conference on Pattern Recognition* (Vol. 2, pp. 1761-1764 vol.2).
- Schlapbach, A. (2007). Writer Identification and Verification. *Ph.D. dissertation, Inst. of Comp. Sci. and App. Math., Bern Univ., Netherlands*, 161. The Netherlands.
- Schlapbach, A., & Bunke, H. (2004a). Off-Line Handwriting Identification Using HMM Based Recognizers. *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)* (Vol. 2, pp. 654-658 Vol.2).
- Schlapbach, A., & Bunke, H. (2004b). Using HMM Based Recognizers for Writer Identification and Verification. *Ninth International Workshop on Frontiers in Handwriting Recognition, IWFHR-9* (pp. 167- 172).

- Schlapbach, A., & Bunke, H. (2006). Off-line Writer Identification Using Gaussian Mixture Models. *18th International Conference on Pattern Recognition, ICPR 2006* (Vol. 3, pp. 992-995).
- Schlapbach, A., & Bunke, H. (2007). A Writer Identification and Verification System Using HMM Based Recognizers. *Pattern Analysis & Applications*, 10(1), 33-43.
- Schlapbach, A., Kilchherr, V., & Bunke, H. (2005). Improving writer identification by means of feature selection and extraction. *Proceedings. Eighth International Conference on Document Analysis and Recognition* (pp. 131-135 Vol. 1).
- Schomaker, L., & Bulacu, M. (2004). Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6), 787-798.
- Schomaker, L., Bulacu, M., & Erp, M. van. (2003). Sparse-Parametric Writer Identification Using Heterogeneous Feature Groups. *Proceedings. International Conference on Image Processing, ICIP 2003* (Vol. 1, pp. I-545-8 vol.1).
- Schomaker, L., Bulacu, M., & Franke, K. (2004). Automatic writer identification using fragmented connected-component contours. *Ninth International Workshop on Frontiers in Handwriting Recognition, IWFHR-9* (pp. 185-190).
- Schomaker, L., Franke, Katrin, & Bulacu, M. (2007). Using codebooks of fragmented connected-component contours in forensic and historic writer identification. *Pattern Recognition Letters*, 28(6), 719-727.
- Schomaker, L., & Vuurpijl, L. (2000). *Forensic Writer Identification: A Benchmark Data Set and a Comparison of Two Systems (research report)*. Nijmegen.
- Shahabi, F., & Rahmati, M. (2006). Comparison of Gabor-Based Features for Writer Identification of Farsi/Arabic Handwriting. *10th International Workshop on Frontiers in Handwritten Recognition (IWFHR)* (pp. 550, 545).
- Shahabi, F., & Rahmati, M. (2007). A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 829-833).
- Siddiqi, I., & Vincent, N. (2007). Writer Identification in Handwritten Documents. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 108-112). IEEE.
- Siddiqi, I., & Vincent, N. (2008). Combining Global and Local Features for Writer Identification. *Proceedings of the Eleventh International Conference on Frontiers in Handwriting Recognition*. Montreal, Canada.

- Siddiqi, I., & Vincent, N. (2009). A Set of Chain Code Based Features for Writer Recognition. *2009 10th International Conference on Document Analysis and Recognition* (pp. 981-985). IEEE.
- Srihari, S. (2000). Distance between histograms of angular measurements and its application to handwritten character similarity. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (pp. 21-24). IEEE Comput. Soc.
- Srihari, S., & Ball, G. (2008). Writer Verification of Arabic Handwriting. *The Eighth IAPR International Workshop on Document Analysis Systems (DAS '08)* (pp. 28-34).
- Srihari, S., & Ball, G. (2009). Comparison of Statistical Models for Writer Verification. *Proceedings Document Recognition and Retrieval XVI* (pp. 7247OE 1-8). San Jose, CA, USA.
- Srihari, S., Beal, M. J., Bandi, K., Shah, V., & Krishnamurthy, P. (2005). A Statistical Model for Writer Verification. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (pp. 1105-1109 Vol. 2). IEEE.
- Srihari, S., Cha, S. H., Arora, H., & Lee, S. (2002). Individuality of Handwriting. *Journal of Forensic Sciences (JOFS)*, 47(4).
- Srihari, S., Huang, C., Srinivasan, H., & Shah, V. (2007). Biometric and Forensic Aspects of Digital Document Processing. *Digital Document Processing, Advances in Pattern Recognition* (pp. 379-405). Springer London.
- Su, T., Zhang, T., & Guan, D. (2007). Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *International Journal of Document Analysis and Recognition (IJ DAR)*, 10(1), 27-38. Springer Berlin / Heidelberg.
- Tan, G., Viard-Gaudin, Christian, & Kot, A. (2008). Online Writer Identification Using Fuzzy C-means Clustering of Character Prototypes. *International Conference on Frontiers in Handwriting Recognition* (p. 6). Montréal : Canada.
- Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition, Fourth Edition* (p. 984). Academic Press.
- Tomai, C. I., & Srihari, S. (2004). Discriminatory Power of Handwritten Words for Writer Recognition. *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004* (pp. 638-641 Vol.2). IEEE.
- Ubul, K., Hamdulla, A., Aysa, A., Raxidin, A., & Mahmut, R. (2008). Research on Uyghur off-line handwriting-based writer identification. *9th International Conference on Signal Processing* (pp. 1656-1659). IEEE.

- Viard-Gaudin, C., Lallican, P., Binter, P., & Knerr, S. (1999). The IRESTE On/Off (IRONOFF) Dual Handwriting Database. *Fifth International Conference on Document Analysis and Recognition (ICDAR)* (pp. 455-458).
- Wang, X., Ding, X., & Liu, H. (2003). Writer Identification Using Directional Element Features and Linear Transform. *Proceedings. Seventh International Conference on Document Analysis and Recognition* (pp. 942 - 945). IEEE Computer Society.
- Yoshimura, I. (1988). Writer identification based on the arc pattern transformation. *9th International Conference on Pattern Recognition* (pp. 35-37 vol.1).
- Zaher, A. A., & Abu-Rezq, A. (2010). A Hybrid ANN-Based Technique for Signature Verification. *Proceedings of the 4th WSEAS International Conference on Computational Intelligence* (pp. 13-19). Bucharest, Romania.
- Zhang, B. (2003). Handwriting Pattern Matching and Retrieval with Binary Features. *Ph.D. dissertation, Dept. of Comp. Sci. and Eng., St. Univ. of N.Y at Buffalo, NY*, 172.
- Zhang, B., Srihari, S., & Lee, S. (2003). Individuality of Handwritten Characters. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* (pp. 1086-1090). IEEE Comput. Soc.
- Zhang, D., Jain, A., Tapiador, M., & Sigüenza, J. A. (2004). Writer Identification Method Based on Forensic Knowledge - Biometric Authentication. *Lecture Notes in Computer Science* (Vol. 3072, pp. 555-561-561). Springer Berlin / Heidelberg.
- Zhu, Y., Tan, T., & Wang, Y. (2000). Biometric Personal Identification Based on Handwriting. *Proceedings. 15th International Conference on Pattern Recognition* (Vol. 2, pp. 797 - 800). Los Alamitos, CA, USA: IEEE Computer Society.
- Zois, E. N., & Anastassopoulos, V. (2000). Morphological Waveform Coding for Writer Identification. *Pattern Recognition*, 33(3), 385-398.

VITA

- Sameh Mohammad Awaida.
- Born in Amman, Jordan on December 29, 1981.
- Nationality: Jordanian.
- Received Bachelor of Science (B.S.) in Electrical Engineering from University of Hartford, CT, U.S.A in 2003, with Magna Cum Laude with a GPA of 3.7/4.0.
- Received Master of Engineering in Electrical Engineering from University of Hartford, CT, U.S.A in 2004, with a GPA of 4.0/4.0.
- Joined the Computer Engineer Department as Lecturer at Princess Sumaya University (PSUT), Amman, Jordan from 2005-2007.
- Joined the Computer Engineering Department as a Lecturer-B at King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in February 2007.
- Completed Doctor of Philosophy (Ph.D.) in Computer Science and Engineering from King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in May 2011, with a GPA of 4.0/4.0.
- Permanent Address: P.O Box 17015, Amman 11195, Jordan.
- Emails: sameho@kfupm.edu.sa, samehemail@gmail.com.