Palestine Polytechnic University
Deanship of Graduate Studies and Scientific Research
Master of informatics

Offline Sub-Word Handwritten Recognition for Arabic Historical Manuscripts

Submitted by:

Ahlam Hasan Al-Bashiti

Thesis submitted in partial fulfillment of requirements of the degree Master of Science in Informatics
Sept, 2014

The undersigned hereby certify that they have read, examined and recommended to the Deanship of Graduate Studies and Scientific Research at Palestine Polytechnic University the approval of a thesis entitled: **Offline Sub-Word Handwritten Recognition for Arabic Historical Manuscripts** , submitted by **Ahlam H. Al-Bashiti** in partial fulfillment of the requirements for the degree of Master in Informatics.

Graduate Advisory Committee:

Prof. Jihad El-Sana (Supervisor), Triangle Research and Development Center.

Signature:________ Date:________

Dr. Hashem Tamimi (Internal committee member), Palestine Polytechnic University.

Signature: Hashem Tamimi Date: 16/4/2015

Dr.Raid Saabni (External committee member), Triangle Research and Development Center

Signature:________ Date:________

Thesis Approved

Dr. Sameer Khader
Dean of Graduate Studies and Scientific Research
Palestine Polytechnic University

Signature:________ Date: 20.04.2015

DECLARATION

I declare that the Master Thesis entitled " **Offline Sub-Word Handwritten Recognition for Arabic Historical Manuscripts** " is my original work, and herby certify that unless stated, all work contained within this thesis is my own independent research and has not been submitted for the award of any other degree at any institution, except where due acknowledgement is made in the text.

Ahlam Hasan Yehya Al-Bashiti

Signature:____________________ Date: 16/4/2015

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for the master degree in Informatics at Palestine Polytechnic University, I agree that the library shall make it available to borrowers under rules of the library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made.

Permission for extensive quotation from, reproduction, or publication of this thesis may be granted by my main supervisor, or in his absence, by the Dean of Graduate Studies and Scientific Research when, in the opinion of either, the proposed use of the material is for scholarly purposes.

Any coping or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Ahlam Hasan Yehya Al-Bashiti

Signature:__________ Date: 16/4/2015

DEDICATION

To my husband "Rabea", to the one who always stood by my side and supported me all the way from the first beginning to the end. Thank you for your unconditional support and patience.

To my parents, who raised me, protected me, encouraged me, and walked by my side until I became the person who I am now, God bless you both.

To my father and mother in law, who have been supportive and encouraging from the first beginning.

To my siblings and friends especially "Rihabb and Haneen".

To my precious and beloving daughters "Noor and Lojien", you are truly the pupils of my eyes.

ACKNOWLEDGEMENT

As my thesis work comes to an end, I truly and proudly appreciate prof Jihad's sustainable effort and time during my journey. I'm humble to say it is such an honour to meet you as well as working with you, and I also want to thank you and Dr Hashim as well for the amazing idea that facilitated the creation of my thesis and cleared the horizon for such a great and wonderful thesis that I already accomplished.

I would like to thank the thesis's examiner Dr. Raid Saabni and Dr. Hashim Tamimi for their valuable suggestions and corrections to this work, which greatly helped me to improve in various aspects.

الملخص

نستعرض في هذه الرسالة بحوثنا المتعلقة في كيفية التعرف الالي على مقاطع الكلمات المكتوبة بخط اليد والمأخوذة من المخطوطات التاريخية القديمة. التعرف الالي للنص المكتوب بخط اليد له استخدامات عدة من اهمها: معالجة النماذج المكتوبة والرموز البريدية...الخ. تم نشر عدد من الابحاث في هذا المجال مؤخرا. تتميز اللغة العربية بخواص معينة تزيدها صعوبة على اللغات الاخرى. فمثلا النقاط والتشكيل وتداخل الحروف والكلمات مع بعضها البعض. كل هذه العوامل جعلت الابحاث التي تعتمد على اللغات الاخرى اكثر انتشارا. اضافة الى عدم وجود قواعد بيانات معيارية.

الابحاث المتعلقة بالنصوص المأخوذة من المخطوطات قليلة جدا. لعدم توافر قواعد بيانات خاصة في هذه النصوص، ووجود نصوص اضافية غير النص الاصلي كالكتابات في الحواشي، واختلاف الخطوط ، واخيرا مع تقدم الزمن قد تفقد بعض الكلمات احرافا بسبب الغبار او اية عوامل اخرى.

نستحدث في هذه الرسالة اساليب وخوارزميات جديدة في التعرف الالي للنصوص المكتوبة في خط اليد. خوارزمية اختيار الخصائص قبل عملية التصنيف. تهدف هذه الخوارزمية الى اختيار مجموعة الخصائص التي تعبر عن النصوص والخصائص غير المناسبة يتم تجاهلها. ادت هذه الخوارزمية الى تحسين النتائج ورفع كفاءة المصنف عن طريق حذف الخصائص الغير ملائمة.

لقد تم تطبيق هذه الرسالة على قاعدة بيانات خاصة تم بناؤها من نصوص مأخوذة من المخطوطات التاريخية. وحصلنا على معدلات نجاح منافسة.

Abstract

In this thesis, we address Arabic offline handwritten recognition for historical documents. The Automation of the handwritten recognition has many applications, such as zip coding, forms processing, indexing and retrieving historical manuscripts and so on. Recognition for Arabic handwritten script lags far compared to other languages such as Latin, and Chinese texts. The challenges for Arabic language raise from its nature such as overlapping characters, cursive texts, and lack of benchmark databases.

In this work, we introduced new techniques for different phases of the offline Arabic handwritten recognition. First it addresses the recognition of the Arabic handwritten for historical manuscripts not contemporary scripts. In addition, the feature selection algorithm is presented. This work aims to select appropriate features and remove irrelevant ones. The relevant features are those which enhance the results and give a higher success rates. We depend on probabilistic classifier not statistically such as HMM. Naive Bayesian classifier is used for training and classification.

The presented work was applied and tested on private database, which was collected from historical manuscripts. A competitive recognition rates were achieved.

The results show that applying feature selection prior classification gives haghier success rates than classification without feature selection.

Table of Contents

List of Figures

List of Figures

List of Tables

List of Abbreviations

PAW	Part of Arabic Word
OCR	Optical Character Recognition
HMM	Hidden Markov Model
NN	Neural Network
SVM	Support Vector Machine
KNN	k Nearest Neighbors

Chapter 1

Introduction

1.1 Introduction

Offline handwritten recognition is the process of determining the textual representation of words or characters in a digital image of handwritten text [38]. It is a sub-field of Optical Character Recognition (OCR) that deals with printed and handwritten texts.

This work is restricted to Arabic language, which script is a challenging and a complex script compared to other scripts, such as Latin. This complexity results from the way Arabic characters are combined to form words, which consist of one or multiple continuous word parts. Characters are connected to form continuous word parts, i.e., one stroke. In addition, a typical Arabic character has multiple forms ,and that depends on its position in a word. Moreover, about half of the Arabic letters include essential dots.

The scripts to be recognised may come from handwritten historical manuscripts, contemporary handwritten, or printed texts. In the case of historical manuscripts, the recognition is more complex. Historical manuscripts are usually of low quality due to aging and often include elements such as ornamentation, seals,

Chapter 1

Introduction

1.1 Introduction

Offline handwritten recognition is the process of determining the textual representation of words or characters in a digital image of handwritten text [38]. It is a sub-field of Optical Character Recognition (OCR) that deals with printed and handwritten texts.

This work is restricted to Arabic language, which script is a challenging and a complex script compared to other scripts, such as Latin. This complexity results from the way Arabic characters are combined to form words, which consist of one or multiple continuous word parts. Characters are connected to form continuous word parts, i.e., one stroke. In addition, a typical Arabic character has multiple forms ,and that depends on its position in a word. Moreover, about half of the Arabic letters include essential dots.

The scripts to be recognised may come from handwritten historical manuscripts, contemporary handwritten, or printed texts. In the case of historical manuscripts, the recognition is more complex. Historical manuscripts are usually of low quality due to aging and often include elements such as ornamentation, seals, and spots. The lines between pages are narrow spaced, which leads to overlapping and touching among characters. Characters may be degraded with ageing and noises. These factors usually affect the recognition process [36]. Over the last few decades, a lot of work have been reported for handwritten recognition, especially for Latin script. However, there is a few publications in the recognition of the Arabic historical manuscripts. manuscripts.

There are many applications for automatic handwritten recognition such as:

- Processing bank checks [51].
- Post Office address and zip code recognition [31].
- Handwritten forms processing [69].
- Indexing and retrieval of manuscripts.

1.2 Thesis objective

This thesis studies the performance of various features on the recognition of the historical Arabic documents. Our database was built to apply this methodology. We apply pre-processing such as diacritics removing, features extraction (structural and statistical), feature selection, training and classification. Applying feature selection and choosing a probabilistic classifier (Naive Bayesian classifier) enhances the results.

1.3 Contributions

The following summarizes the main contributions of this thesis:

- A probabilistic classifier gives the posterior probability for each class, not only one result which is the proposed class. This will enhance the results by predicting not only the highest posterior but also the first two or three options. We can depend on more than one posterior probability for each tested sample. This idea has shown good results

- Studying a set of features (structural and statistical), the results showed the strength of these features to represent each PAW. Features include solidity, holes, moment invariant, theta, and so on.
- Applying our algorithm of selecting best features before training and classification enhances the results. The aim of this algorithm is to select the appropriate features that enhance the recognition rates and remove the irrelevant features.
- Data from historical Arabic manuscripts, it was collected from historical Arabic manuscripts, which we used to build our small database.

1.4 Thesis Organization

The remaining parts of the thesis are organized as follows: chapter 2 describes the theories and basic concepts that are needed to understand the rest of the thesis. Chapter 3 contains a summary of the previous works that are closely related. Chapter 4 explains the methodology used in this thesis to enhance the accuracy of the recognition. Chapter 5 demonstrates the experiments and analyses the obtained results. Finally, Chapter 6 concludes the work and suggests directions for the future work.

Chapter 2

Background

This chapter gives a theoretical background that is needed for understanding the thesis. In the first section we discuss the motivation behind studying the Arabic script, the challenges that encounter the Arabic language, printed versus handwritten Arabic text, model of handwritten recognition, research approaches, phases of the handwritten recognition system, and finally feature selection.

2.1 Motivation

Arabic is the mother tongue of more than 300 million native speakers. It has been used by more than 1.5 billion Muslims across the world. Furthermore,written Arabic has been adopted for use in a wide variety of languages such as Kurdish, Malay, Persian and Urdu. Thus, the automation of the Arabic scripts would have widespread advantages [1]. On the other hand, Arabic handwritten recognition enables the automatic reading and searching for the Arabic historical document. This process enables important knowledge to be accessible to the general public, while protecting the historical documents from deterioration by frequent handling. To optimally utilize the digital availability of these documents, it is important to develop searching and indexing mechanisms. Currently, indexing is manually constructed and the search is performed on the scanned pages, one by one. Since this procedure is expensive and time consuming, an automation process is desirable. Off-line handwriting recognition can be used to convert these images into text files.

2.2 Arabic Handwriting Recognition: Challenges

Arabic script is more complex than other scripts such as Latin. The main characteristics of the Arabic can be summarized as follows [5]:

1. Arabic is a cursive script written from right to left. The characters are usually connected along the baseline.

2. It consists of 28 characters. Each one has two to four shapes; depending on the character's position within its word. Figure 4.1 [12] shows each shape for each character.

3. There are disconnect six characters. Their existence segment the word into different parts. Some publications call them sub-words Pieces of Arabic Words or PAWs. Figure 4.2 .

4. Arabic script contains ligatures, which are characters composed from combining two or more characters in an accepted manner. Ligatures are difficult to segment into component characters and may be treated as separate characters figure 4.3 [29]. Also, there are diacritics which represent short vowels or sound, such as fat-ha, dumma, and sukkun, see figure 2.4 [5], which are normally omitted from handwritings.

5. The segmentation of the word into its component letters or PAWs for the connected Arabic word is a challenging task. Sometimes, mismatch dots (above or below) characters differs the meaning of the characters, ligatures, and also PAWs make segmentation a difficult task [48].

6. Arabic has complex and different fonts and writing styles, see figure 2.5 [18].

2.3 Printed versus handwritten Arabic text

Optical Character Recognition (OCR) is "the mechanical or electronic conversion of scanned images of typewritten or printed text into machine-encoded/computer-readable text". Offline text can be either printed or handwritten. The recognition of handwritten text, compared to printed text, is considered more difficult due to the following reasons [58]:

- Mono-spaced text: in the case of printed text,each character has the same distance apart of any other character. Handwritten characters are connected or are not spread out equally.

- United height and width: in handwritten text the width and height of characters are not united, but in printed text, the width and height of characters are the same. These irregularities complicate the recognition of text and can make other important related tasks, such as segmentation more difficult.

- Stable Base-Line: Text characters switch on the same horizontal base-line. In handwriting the characters vary up and down on an invisible baseline. This necessitates determine the writing line at the word and line levels.

- Handwriting Styles: In handwritten, there are several styles such as Naskh, Hijazi, Thuluth, and others. In several instances, writers mix between these writing styles. While in printed text there is one style for printing the text.

  In the case of the historical manuscripts, the process is even more complex. These manuscripts are generally degraded, vulnerable to noise, and use complex aesthetic calligraphic styles. Sometimes they have notes on the margins, which make the recognition task more difficult.

2.4 Model of handwriting recognition systems

The diagram in figure 2.6 [46] represents the general model of the offline Arabic handwritten recognition. As we see, the input to the system is a scanned page contains the text. The page may need to go through some preprocessing steps for enhancement purposes, such as noise removal, skew detection. Then text-line and/or word segmentation are applied to extract the words in the input image. In our work, we segment the word into parts of words. Then, features are extracted for training and testing using the Naive Bayesian classifier. Finally post processing is applied to improve the recognition by refining the decision of the classifier by using a lexicon.

The handwritten recognition system is composed of different phases. Therefore, the performance of the system depends on the accuracies of each one. Each one has its own challenges and difficulties. Segmentation phase itself is difficult and may introduce errors if it is not treated carefully. Feature extraction is also a challenging task, as they must be distinct and accurate to give good results.

2.5 General Research Approaches

There are several research approaches for handwritten recognition. Next, we explain these approaches.

2.5.1 Online versus Offline Recognition

Handwritten recognition can be either online [39], [4], [16] or offline [67], [21], [15]. Online handwriting is represented as a sequence of (x, y) coordinates that represent each point on the handwriting. The usage of online data gives additional information about the content of the handwriting, which includes the time of writing by storing the points as they are written in sequence. The time each segment/stroke is considered an additional feature about that segment or stroke which helps in enhancing recognition. Generally the position of the writing, velocity, and acceleration are functions of time [57].

Offline systems aim to recognize texts represented in image format. Mainly the texts are written on paper or using stylus supported devices (typewritten or print), then these writings are stored as images. Offline handwriting recognition is often called Optical Character Recognition (OCR). OCR includes machine-printed text and handwritten text.

Offline handwritten recognition is observed to be harder than online handwritten recognition. In online, features can be extracted directly from the pen movement, however, in offline only static images are available. The raw data of the images do not contain information about the writing order [7].

In this thesis, we will deal with the offline handwritten recognition for Arabic manuscript. The advances in electronic storage and digital scanning have driven the automation of historical documents for keeping and the analysis of cultural heritage. Currently, indexing is manually constructed and the search is performed using off-line handwriting recognition to convert these images into text files.

2.5.2 Segmentation free versus Segmentation based approach

Handwritten Word Recognition techniques use analytic or holistic methods for recognition. In Holistic (segmentation free) strategies features are extracted from the entire word image. Thus eliminates the segmentation problem. Another approach uses implicit segmentation, where a sliding window scans through the line of text and features are extracted from the portion of the text line within the window. The analytical (segmentation based) strategies segment the word into characters or strokes. However, segmentation algorithm is not available for exactly extracting characters from a given word. Features are extracted for the segmented components [49].

2.6 Offline Arabic handwritten word recognition phases

Next, the offline handwritten recognition phases are explained. They include image acquisition and preprocessing and segmentation, features extraction, and recognition.

2.6.1 Image acquisition and preprocessing and Segmentation

Offline data can be collected from handwritten pages. Scanners or cameras are used to convert these pages into images. In our work, we used Arabic search is performed using off-line handwriting recognition to convert these images into text files.

2.5.2 Segmentation free versus Segmentation based approach

Handwritten Word Recognition techniques use analytic or holistic methods for recognition. In Holistic (segmentation free) strategies features are extracted from the entire word image. Thus eliminates the segmentation problem. Another approach uses implicit segmentation, where a sliding window scans through the line of text and features are extracted from the portion of the text line within the window. The analytical (segmentation based) strategies segment the word into characters or strokes. However, segmentation algorithm is not available for exactly extracting characters from a given word. Features are extracted for the segmented components [49].

2.6 Offline Arabic handwritten word recognition phases

Next, the offline handwritten recognition phases are explained. They include image acquisition and preprocessing and segmentation, features extraction, and recognition.

2.6.1 Image acquisition and preprocessing and Segmentation

Offline data can be collected from handwritten pages. Scanners or cameras are used to convert these pages into images. In our work, we used Arabic historical manuscripts images for recognition. Handwritten images may need some preprocessing techniques to facilitate the recognition. Common techniques are Binarization, for converting the image into black/white model, skeleton , and segmentation, for segmenting the word into PAWs.

2.6.2 Feature Extraction Approaches

Features are the information extracted from the image to represent the shape of the word. The main goal of the feature extraction is to remove the redundant data and to produce a set of mathematical properties, shape data or pixels features. These features are mapped to the classifier to determine the corresponding class. Many features have been developed by various researchers. These features can be divided into two types: statistical and structural features. Statistical features: are numerical measures computed over images or regions of images. They include histogram of chain code directions, moments, pixel densities and others. Structural features are intuitive parts of writing, such as end points, loop, start point, and dots [38].

2.6.3 Classification Approaches

There are several classifiers have been used for recognition, such as Hidden Markov Model (HMM), Neural Network (NN), Support Vector Machine (SVM), Naive Bayesian, and others. In our work we used the Naive Bayesian classifier. This classifier depends on probability. It performs well when the dataset is small. We are encouraged to test it's performance because we have small data, and we want the probability for a testing sample not exact value.

Naive Bayesian classifier

The Naive Bayesian classifier is a supervised learning method, also it is a statistical method for classification. It is based on Bayes theorem with independence assumptions between predictors. It is simple, and widely used, because it is often outperforms more complicated classification methods [66]. Naive Bayes model is applied to inferential statistics and decision making that deals with probability inference. It uses the knowledge of prior events to predict future events. Parameter estimation for this model uses the method of maximum likelihood. It needs a small amount of training data to estimate the parameters.

Naive Bayes classifier assumes that the effect of the value of a predictor x on a given class c is independent of the values of other predictors. This assumption is called class conditional independence [45].

Derivation:

D : Set of rows.

- Each row is an n dimensional attribute vector.
- X : $(x_1, x_2, x_3, .x_n)$ Let there be m Classes : $c_1, c_2, c_3 c_m$ Nave Bayes classifier predicts X belongs to Class C_i if $P(C_i/X) > P(C_j/X)$ for $1 <= j <= m, j <> i$ Maximum Posterior Hypothesis.
- $P(C_i/X) = P(X/C_i)P(C_i)/P(X)$
- Maximize $P(X/C_i)P(C_i)$ as $P(X)$ is constant Nave Assumption of "class conditional independence"

2.7 Feature Selection

Feature selection is the process of selecting a subset of appropriate features for model construction. Feature selection often increases classification accuracy by eliminating irrelevant features. In fact there are two problems which may be evoked by those features. First, the irrelevant features will result greater computational cost. Second, the irrelevant input features may lead to overfitting [17].

Feature selection methods can be classified into wrapper methods and filter methods. The wrapper method evaluates the features using the learning algorithm (classifier) that will be employed. While filter method, examines essential properties of the data to evaluate the features before learning tasks [23].

The main benefits of feature selection are follows [32]:

- Reducing the measurement cost and storage requirements.
- Coping with the degradation of the classification performance due to the finiteness of training sample sets.
- Reducing training and utilization time
- Facilitating data visualization and data understanding.

2.7.1 Image Opening

Morphological operations affect the structure, shape, or form of an object. It is applied on the binary images. The two main morphological operations are erosion and dilation. Erosion shrinks objects by eroding the boundaries. Dilation makes objects to expand, thus potentially connecting disjoint objects and filling in small holes. Morphological operations take two arguments, the binary image A and the structure element B. The structure element is a binary image (or mask) that allows us to define arbitrary neighborhood structures each with value zero or one. [13]

Dilation

The dilation process is performed by laying the structuring element B on the image A and sliding it across the image in a manner similar to convolution [13].

1. If the origin of the structuring element coincides with a 'white' pixel in the image, there is no change; move to the next pixel.
2. If the origin of the structuring element coincides with a 'black' in the image, make black all pixels from the image covered by the structuring element.

Erosion

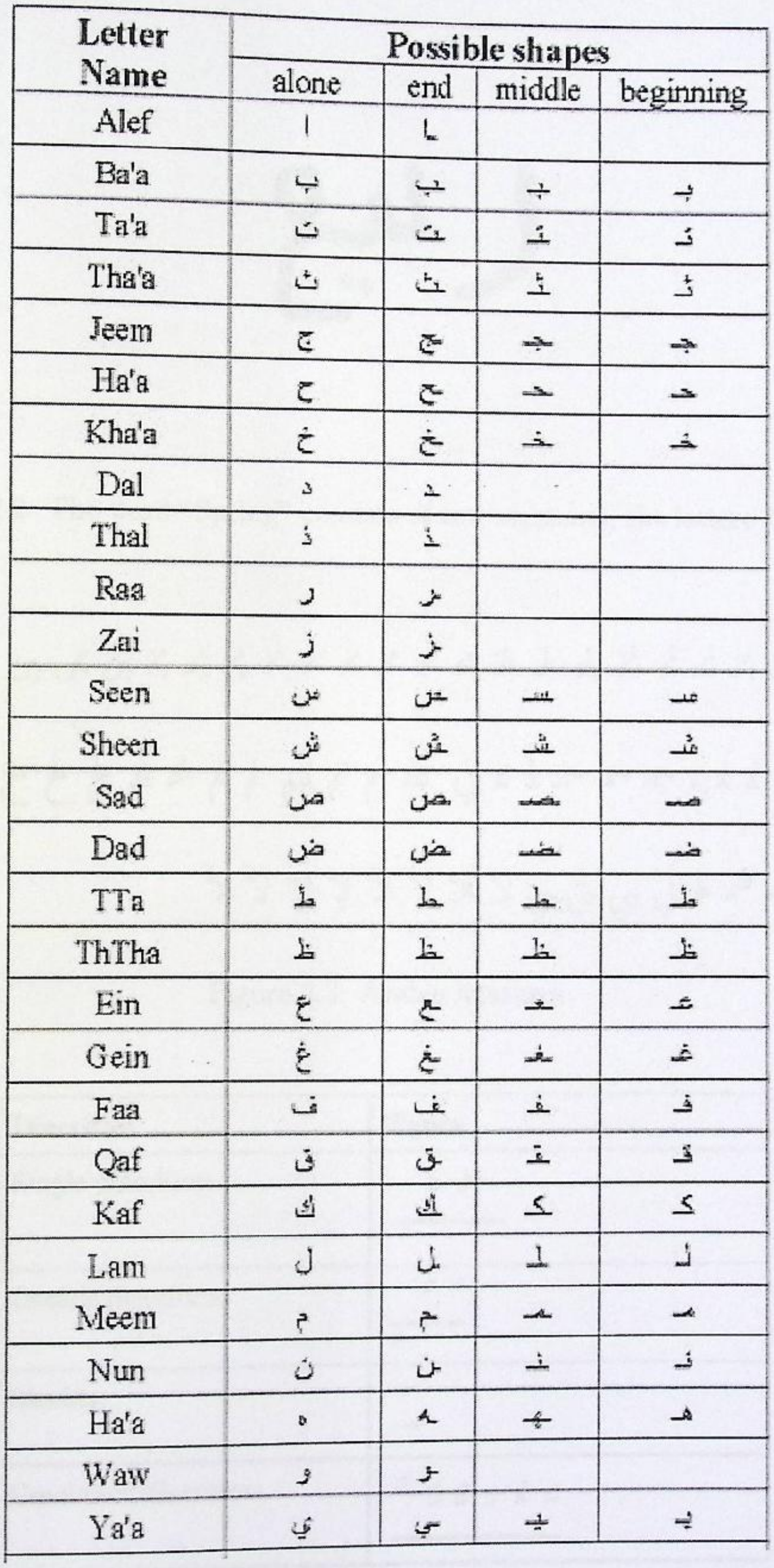
The same as dilation except converting pixels into white not black. Slide the structuring element across the image to do the following [13]:

1. If the origin of the structuring element coincides with a 'white' pixel in the image, there is no change; move to the next pixel.
2. If the origin of the structuring element coincides with a 'black' pixel in the image, and at least one of the 'black' pixels in the structuring element falls over a white pixel in the image, then change the 'black' pixel in the image from black to a 'white'.

Opening

Opening consists of an erosion followed by a dilation and can be used to eliminate all pixels in regions that are too small to contain the structuring element. In this work, the image opining is applied to connect diconnected pixels in some cases [13].

Figure 2.7 represents the original image and figure 2.8 represents the image after opening.



Letter Name	Possible shapes			
	alone	end	middle	beginning
Alef	ا	ـا		
Ba'a	ب	ـب	ـبـ	بـ
Ta'a	ت	ـت	ـتـ	تـ
Tha'a	ث	ـث	ـثـ	ثـ
Jeem	ج	ـج	ـجـ	جـ
Ha'a	ح	ـح	ـحـ	حـ
Kha'a	خ	ـخ	ـخـ	خـ
Dal	د	ـد		
Thal	ذ	ـذ		
Raa	ر	ـر		
Zai	ز	ـز		
Seen	س	ـس	ـسـ	سـ
Sheen	ش	ـش	ـشـ	شـ
Sad	ص	ـص	ـصـ	صـ
Dad	ض	ـض	ـضـ	ضـ
TTa	ط	ـط	ـطـ	طـ
ThTha	ظ	ـظ	ـظـ	ظـ
Ein	ع	ـع	ـعـ	عـ
Gein	غ	ـغ	ـغـ	غـ
Faa	ف	ـف	ـفـ	فـ
Qaf	ق	ـق	ـقـ	قـ
Kaf	ك	ـك	ـكـ	كـ
Lam	ل	ـل	ـلـ	لـ
Meem	م	ـم	ـمـ	مـ
Nun	ن	ـن	ـنـ	نـ
Ha'a	ه	ـه	ـهـ	هـ
Waw	و	ـو		
Ya'a	ي	ـي	ـيـ	يـ

Figure 2.1: The Arabic letters. The shape of each lettere depending on its position in the word

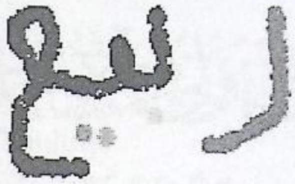


Figure 2.2: The word "Spring" consists of tow segments, the lettere "R" and "Bie'"

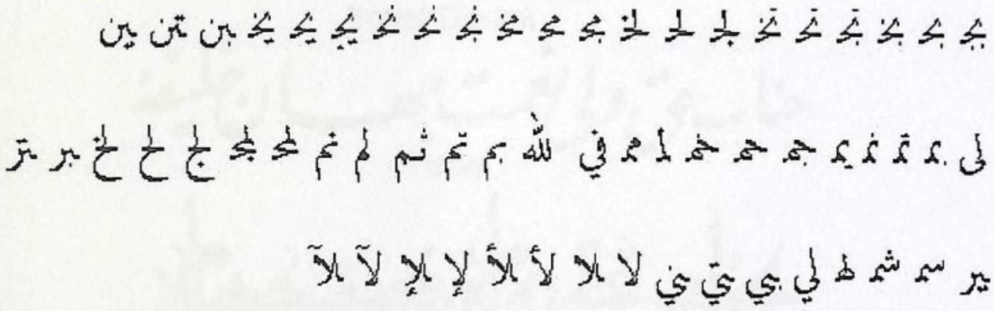


Figure 2.3: Arabic ligatures

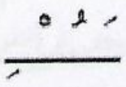
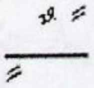
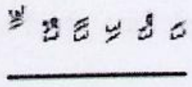
Diacritics	Figure
Single diacritics:	
Double diacritics:	
Shadda:	
Combined diacritics:	

Figure 2.4: Arabic diacritics

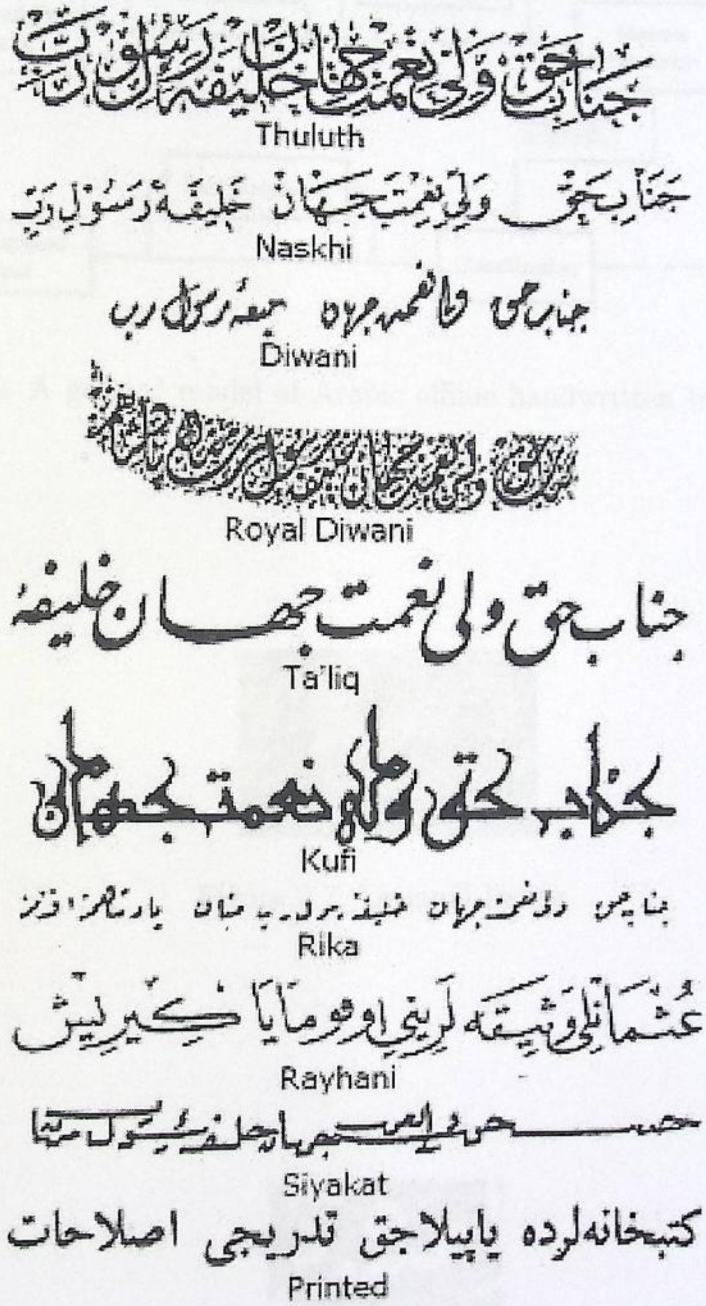


Figure 2.5: Arabic fonts and styles

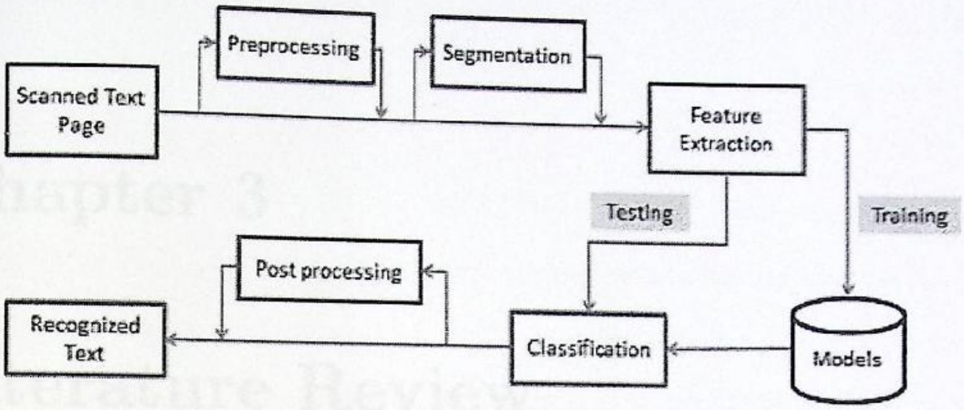


Figure 2.6: A general model of Arabic offline handwritten text recognition system

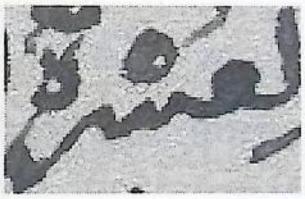


Figure 2.7: Original Image

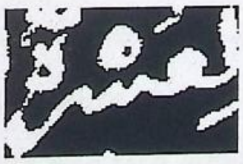


Figure 2.8: Image after opening

Chapter 3

Literature Review

This chapter contains a summary of some important contributions related to our work. This chapter includes researches for historical Arabic manuscripts, feature extraction, statistical features, structural features, and classification.

3.1 Historical Arabic manuscripts

Recently scientists are interested in analyzing and studying the historical manuscript, to help historian and researchers for retrieving information in order to facilitate the editing, indexing, and archiving process. Historical manuscripts are written in various languages (Arabic, Latin, Hebrew, etc). Arabic manuscripts are considered to be more complex than other manuscripts written in other languages. In general any manuscript faces different problems such as low quality images, character degradation, noise, stains, etc. Arabic manuscripts also face other problems related to the nature of the language such as (cursiveness of the text, character overlapping, diacritic, decorations, and so on). Unfortunately, the published researches for handwritten recognition for historical documents is little compared to the handwritten recognition for scripts that are not historical. We don't know the exact reasons for that, maybe the scarcity of databases or the problems of the manuscripts as we mentioned previously. The situation of the Arabic manuscripts is more complicated. We searched a lot and we found only few researches.

Abd Al-Aziz et al. [2] proposed a recognition system for Arabic manuscripts. The purpose of their work is to discriminate between historical documents of different writing styles to three different ages: Contemporary age, Ottoman age and Mamluk age. They depend on a Spatial Gray-level Dependence technique that provides eight different texture features for each sample document. They achieved results up to 95.83 correct classification.

Khorsheed [61] proposed a holistic approach for words handwritten recognition in historical Arabic manuscripts. The 2D Fourier transform is applied to polar word images. Multiple Hidden Markov Models are applied for recognition. The recognition of an unknown word is based on finding the likelihood probability of that word against each word model and giving the word model with the highest probability.

On the other hand,Farag [59] introduced a system to recognize text from historical documents. He automatically segmented the image into lines using horizontal projection, lines into words using boundary box and words into characters using vertical projection with no overlap among text lines, and finally he recognizes the characters using advanced neural-network training algorithm with optimized error, that was based on modified Gram Schmidt with Reorthogonalization algorithm for increasing the performance of recognition.

3.2 Feature Extraction

Features are the information extracted from the image to represent the shape of the word (see chapter 2). In general, features can be divided in to two types: statistical and structural features. Statistical features: are numerical measures computed over images or regions of images. Structural features: are aspects of writing, such as loops, end points, and dots. Next we will review these types of features.

3.2.1 Statistical Features

(Pechwitz and Maergner,2003) [43] used the pixel values as a basic features using rectangular window. Their recognition rate was 90 percent. Mozaffari et al(2005) [64] found the average and variance of X, and Y changes in portions of the image skeleton. Spectral features was used by Khorsheed et al (2007) [61]. Concavity features and distribution features such as density, derivative and etc were used by Mohamad et al (2009) [54]. Hamdani et al (2009) [40] used pixel values, density, moment, distribution and concavity In addition, directional, contour and density features were tested by Kessentini et al(2010)) [73]. SIFT descriptors and Harris corners were used by Rothacker et al (2012) [37].

3.2.2 Structural Features

Structural features such as dots, loops, ascender, descender, curves, end point, branch point, start point, and etc. next we review some of researches used structural features for recognition. Almuallim and Yamaguchi in 1987 [27] proposed the first Arabic handwritten recognition for words. They used the skeleton of the word to get the structural features. They got 91 percent recognition rate(1). (Amin,2001) [8] traced thinned image to find primitives(lines , curves, and loops . Then the structure of each primitive is described, for example line could be , small , medium, and large size. The recognition rate was 86.65 percent. Other work was done by Maddouri and Amiri (2002) [63]. They found the dots (below or above) the word, ascending, desending, and loops. Also they found local features using Fourier descriptors. Their recognition rate wes 98 percent. (Khorsheed,2003) [60] they described a loop feature as simple, complex and double loop. End points, turning points, branch point, and cross point.

3.2.3 Structural and Statistical Features

Some researchers used both structural and statistical features for recognition. Eraq and Abdelazeem [24] computed the gradient of the image to produce eight directional sub-images. For each sub-image a sliding window is used to extract these features: The density of foreground pixels of the sub-image, the density of the window foreground pixels, the window centroid, and the distance between the uppermost foreground pixel of the window and the lowest one. They also depended on the horizontal baseline to extract the width and height (the distances from the baseline to the upper and lower foreground pixels) to recognize diacrtices. IFN/ENIT database was used for evaluation.Haboubi et al [68] used ascenders, descenders , Loops , diacritic dots,and the position of diacritic dot as structural features. Also Fourier Descriptors, Gabor filter as statistical features. Zavorin et al [28] used dots, branch point, end points ,cross points, and loops as structural features. They used Aspect Ratio, Centroid Location Relative to Width, Centroid Location Relative to Height, Density, Maximum Transitions in the Horizontal Direction, Top Edge to Bottom Edge Ratio , Right Edge to Left Edge Ratio, and etc as statistical features. Azizi et al [47] used number of dots and their positions below or above the baseline,stroke, Concavity features, Descendants and ascendants as statistical features. They also used black pixels density and the variance as statistical features. Chen et al [30] used Gabor features that operates directly on gray-level images without needing image binarization as statistical features. They also compute GSC structural features (Gradient ,Structure,Concavity). Their experiments showed that Gabor features enhances the results more than Graph features. They are slightly better than GSC features for PAW. Combining Gabor and GSC give a significant error rate over using Gabor or GSC alone. El-hajj [55] divided the word image into vertical overlapping windows (or frames), then each frame is divided into cells for each frame, there are 24 features such as Density, densities of black (foreground) which are pixels for each vertical column of pixels in each frame, the density of foreground pixels over and under the Lower baselines for each frame, and concavity features.

3.3 Feature Selection

In general increasing the size of the feature vector slowing down the learning process as well as causing the classifier to overfit the training data.

A redundant feature does not add anything new to describing the target concept. Redundant features might possibly add more noise than useful information in describing the concept of interest. Feature selection is selecting s ubset features and removing irrelevant ones.

A feature selection framework generally consists of two parts: a searching engine used to determine the promising feature subset candidates, and a criterion used to determine the best candidate [33].

3.3.1 Searching Strategies

- Greedy and heuristic search

  Applying forward or backward sequential schemes, which always provide a sub-optimal solution. Forward strategies usually provide a nested rank of variables, with the drawback of conditioning the m selected features given the previous m-1 selected. The backward strategy starting from the whole set of variables and discarding one at the time to get to the subset of m desired features. Researches such as Wang et al. (2009) [72] , Tang and Mao (2007) [70]

- Optimization based search

  The feature selection problem can be considered as an optimization problem, researchers have used: Genetic algorithms [25] , Ant Colony Optimization [3], and Particle Swarm Optimization [9] [33].

3.3.2 Evaluation Criteria

The researchers focus on the design of performance measures to determine the relevance between features and decision. Distance [62], consistency [35], correlation [11], and mutual information [20] are usually used [33].

3.4 Classification

Researchers have utilized different classifiers for the recognition of Arabic handwritten words, characters and numerals. These classification algorithms include Hidden Markov model (HMM) [52], [42], support vector machines (SVM) [6], artificial neural networks (NN) [10], [22] , k nearest neighbors (kNN) [14], Bayesian networks [41] and others. Combinations of different classifiers to improve recognition have also been investigated in [19], [65].

3.5 Thesis Contribution

Reviewing previous literature showed that there are few works have been reported for recognition of Arabic historical manuscripts. On the other hand, few other studies have focused on using feature selection algorithms before the classification to improve the classification accuracy.

This thesis is concerned with studying the performance of various features on the recognition of the historical Arabic manuscripts. Structural and statistical features will be studied.

An algorithm is built to improve the accuracy of the classifier by selecting the best features. The probabilistic Naive Bayesian classifier is used for recognition.

Chapter 4

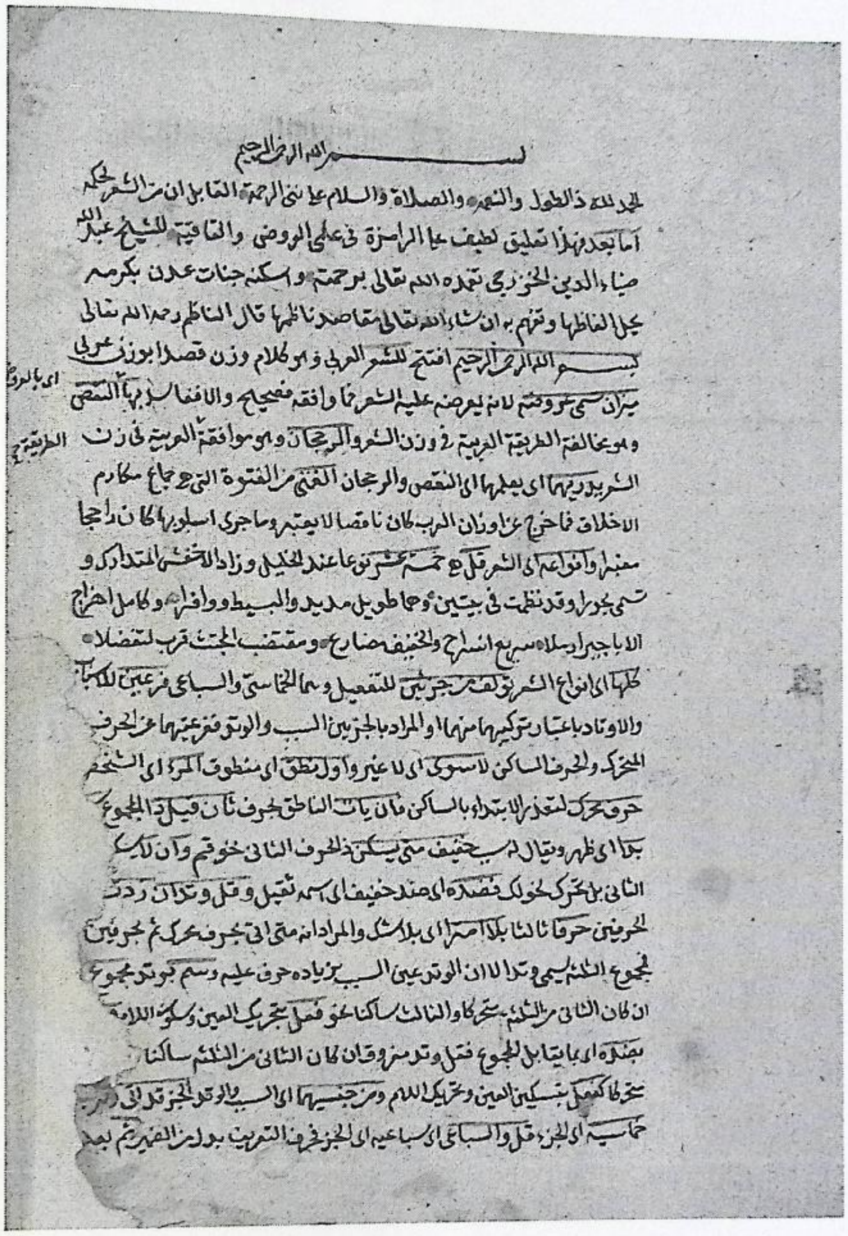
Data and Methods

This chapter covers the methodology used in this thesis that aims at enhancing the performance of the recognition process. Section one represents the structure and organization of the data set, section two prepossessing,section three feature extraction, section four classification, and section five feature selection.

4.1 Structure and organization of the Data Set

4.1.1 Historical Arabic manuscripts

Historical Arabic manuscripts were used to build the database. First step, we collect several pages of Arabic manuscript for the same writer. Each page has many words for example 600 word. See figure 4.1. Figure 4.2 represents general Arabic script not historical manuscript. The recognition for historical manuscript is more complex than general scripts.See chapter 2

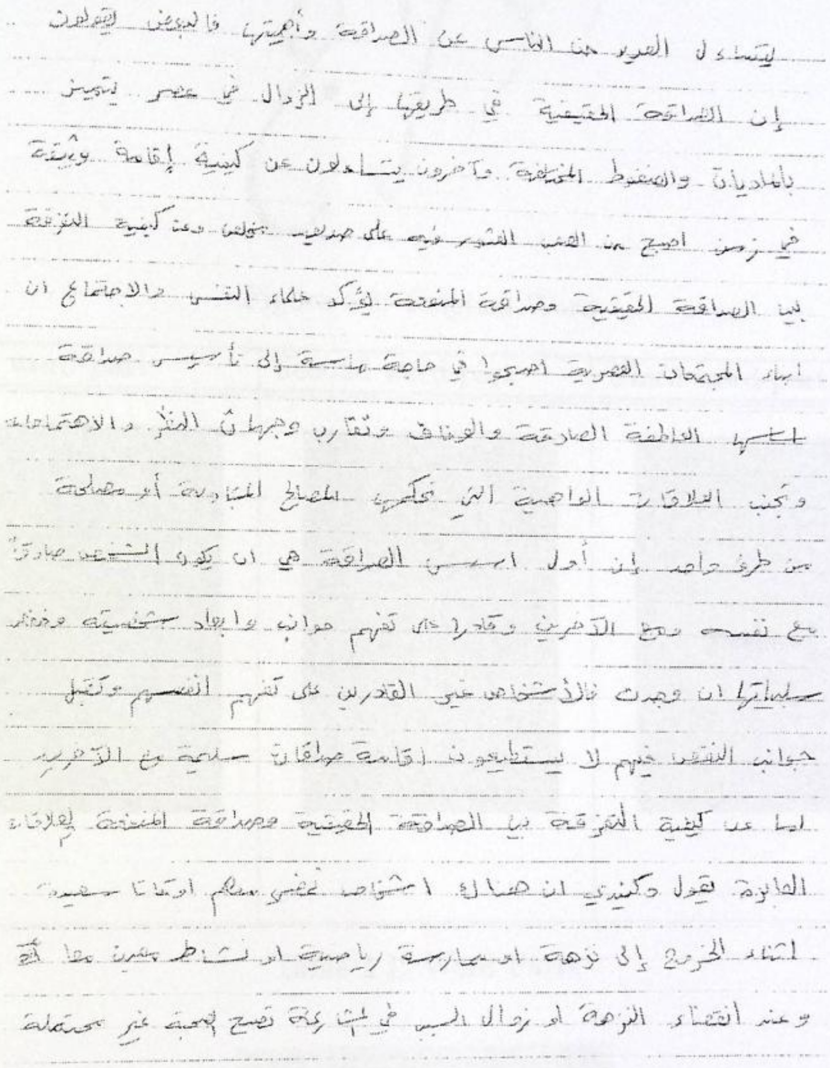


بسم الله الرحمن الرحيم
الحمد لله ذي الطول والنعم والصلاة والسلام على نبي الرحمة القائل ان من الشعر لحكمة
اما بعد فهذا تعليق لطيف على الرامزة في علمي العروض والقافية للشيخ عبد الله
ضياء الدين الخزرجي تغمده الله تعالى برحمته واسكنه جنات عدن بكرمه
يحل الفاظها ويتمم به ان شاء الله تعالى مقاصد ناظمها قال الناظم رحمه الله تعالى
بسم الله الرحمن الرحيم افتتح للشعر العربي وهو كلام وزن قصدا بوزن عربي
ميزان سمي عروضا لانه يعرض عليه الشعر فما وافقه فصحيح والا ففاسد بها النقص اي بالعروض
وهو مخالفة الطريقة العربية في وزن الشعر والرجحان وهو موافقة العربية في وزن الطريقة
الشعر يدريهما اي يعلمهما اي النقص والرجحان الفتى من الفتوة التي هي جماع مكارم
الاخلاق فما خرج عن اوزان العرب كان ناقصا لا يعتبر وما جرى اسلوبها كان راجحا
معتبرا وانواعه اي الشعر قل هي خمسة عشر نوعا عند الخليل وزاد الاخفش المتدارك و
تسمى بحورا وقد نظمت في بيتين وهما طويل مديد والبسيط وافر وكامل اهزاج
الارجاز ارملاء سريع انسراح والخفيف مضارع ومقتضب المجتث قرب لتفضلا
كلها اي انواع الشعر تولف من جزئين للتفعيل وهما الخماسي والسباعي من عين للكبار
والاوتاد باعتبار تركيبهما منهما والمراد بالجزئين السبب والوتد فمن عينهما من الحرف
المتحرك والحرف الساكن لا سوى اي لا غير واول منطق اي منطوق المرء اي الشخص
حرف محرك لتعذر الابتداء بالساكن فان يات الناطق بحرف ثان فقيل ذا المجموع
بدا اي ظهر ويقال له سبب خفيف متى يسكن ذا الحرف الثاني نحو قم وان لا يسكن
الثاني بل تحرك نحو لك فضده اي ضد خفيف اي اسمه ثقيل وقل وتد ان زدت
الحرفين حرفا ثالثا بلا امترا اي بلا شك والمراد انه متى اتى بحرف محرك ثم بحرفين
المجموع الثلاثة يسمى وتدا لا ان الوتد عين السبب بزيادة حرف عليه وسم وتدا مجموعا
ان كان الثاني من الثلاثة متحركا والثالث ساكنا نحو فعل بتحريك العين وسكون اللام
بضده اي بما يقابل المجموع فقل وتد مفروق ان كان الثاني من الثلاثة ساكنا
متحركا كفعل بتسكين العين وتحريك اللام ومن جنسيهما اي السبب والوتد الجزء قد اتى ذا
خماسية اي الجزء قل والسباعي اي سباعيه اي الجزء في التركيب بواسط الضمير ثم بعد

Figure 4.1: Historical Arabic Manuscript

4.1.2 Segmentation of the word

The OCR system of this thesis used the sub-word level (PAW) recognition. Every word contains more than one part is segmented into different parts. The word in figure 4.3 contains mainly three parts. The result of the segmentation is three sub words (PAW). See table 4.1 the first column contains the first word-part, second column is the second word-part, and the third one is the third word-part. Not all words need segmentation, some words have



يتساءل الكثير من الناس عن الصداقة وأهميتها فالبعض يقولون
إن الصداقة الحقيقية في طريقها إلى الزوال في عصر يتميز
بالماديات والضغوط المختلفة وآخرون يتساءلون عن كيفية إقامة وثيقة
في زمن أصبح من الصعب العثور فيه على صديق ويتساءلون عن كيفية التفرقة
بين الصداقة الحقيقية وصداقة المنفعة يؤكد علماء النفس والاجتماع ان
أبناء المجتمعات العصرية أصبحوا في حاجة ماسة إلى تأسيس صداقة
أساسها العاطفة الصادقة والوفاق وتقارب وجهات النظر والاهتمامات
وتجنب العلاقات الواهية التي تحكمها المصالح المتبادلة أو مصلحة
من طرف واحد إن أول أسس الصداقة هي ان يكون الشخص صادقا
مع نفسه ومع الآخرين وقادرا على تفهم جوانب وأبعاد شخصيته [illegible]
سلبياتها ان وجدت فالأشخاص غير القادرين على تفهم أنفسهم وتقبل
جوانب النقص فيهم لا يستطيعون اقامة علاقات سليمة مع الآخرين
اما عن كيفية التفرقة بين الصداقة الحقيقية وصداقة المنفعة والعلاقات
العابرة تقول د.كنيدي ان هناك اشخاص تجمعهم اوقاتا سعيدة
اثناء الخروج إلى نزهة او ممارسة رياضية او نشاط معين معا لكن
وعند انتهاء النزهة او زوال السبب في المشاركة تصبح الصحبة غير محتملة

Figure 4.2: General Arabic script

only one part. See figure4.4

As mentioned in chapter 2 there are six characters split the word into different PAWs. See Figure 4.5. Table 4.2 represents examples for words contain splitting characters.

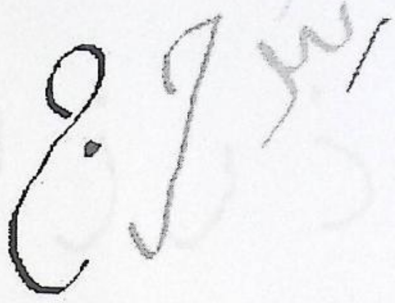


Figure 4.3: Original Word

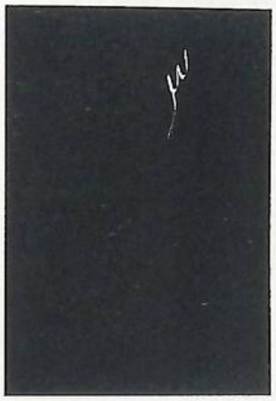
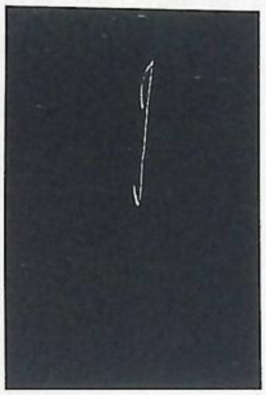
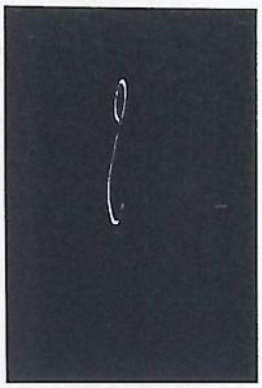
First word-part	Second word-part	Third word-part
		

Table 4.1: Word Parts

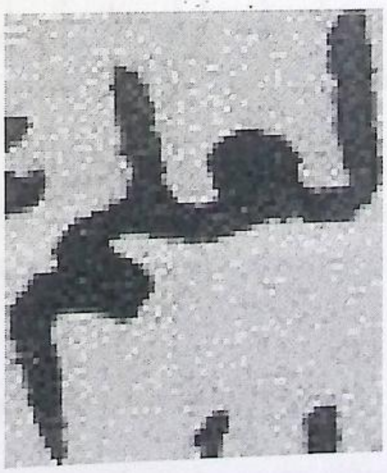


Figure 4.4: The word has only one part

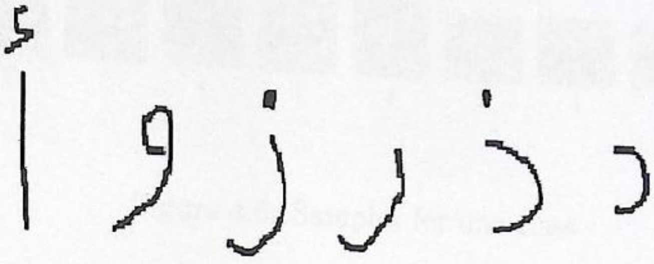


Figure 4.5: Splitting characters

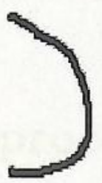
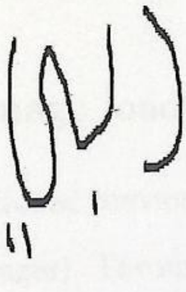
First Character	Second Character	Third Character
		
First Word	Second Word	Word
		

Table 4.2: Splitting characters and words

4.1.3 Structure of the database

After the segmentation process, there are many different PAWs. We used ten different Arabic manuscripts; each one approximately has from 590 to 680 PAWs. We collect all samples for each PAW and saved them in one folder. We

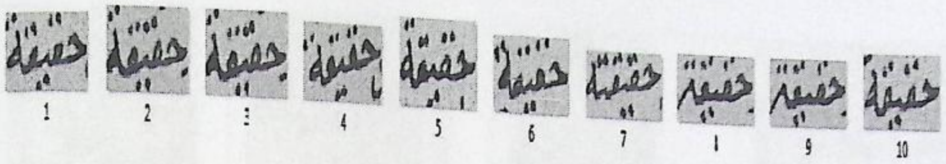


Figure 4.6: Samples for one class

faced a problem that most of PAWs only have lower than four samples. There are about 50 classes each class has 10 samples. So the database contains 500 samples. From 5900-6800 PAWs only 500 PAWs were used for learning and testing. Each folder is named with it's class number. Figure 4.6 represents the tens samples for class number 5.

4.2 Preprocessing

Handwritten recognition system needs to go through preprocessed steps before features extraction and recognition. These steps may include, binarization, thinning, and so on. Next the stpes that are included in pre-processing.

4.2.1 Image loading

As we mentioned previously, there are many classes. Each class has many samples (images). The input to the system is an image contains the PAW. See figure 4.7. This image is loaded through the system program to go through several steps and finally for learning and recognition.

4.2.2 Binarization

The image is converted from grayscale color model into black/white (binary model). The foreground pixels are black (foreground is the exact text), and the background pixels are white. This process is done since some features

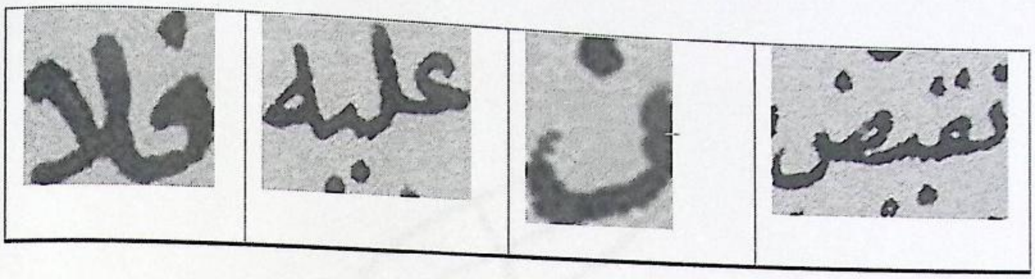


Figure 4.7: Samples for the input of the system

can't be extracted using grayscale color model. See table 4.3 .

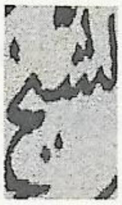
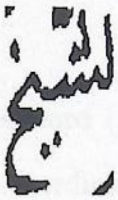
Grayscale Image	Binary Image
	

Table 4.3: Color model converting

4.2.3 Diacritics removing

PAW may has some type of diacritics, such as dots, tanween , shadda, and so on. All diacritics are removed. Only the main part (connected component) is kept. We depend on the area fro each component, as follows:

- Calculate area for each component.
- Select the maximum area as the main component.

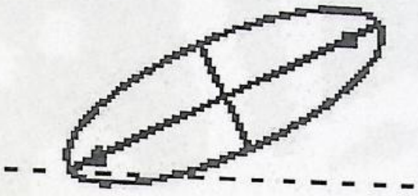


Figure 4.8: Ellipse axises

4.3 Feature Extraction

- Word Center

  This feature represents the center of the word (X_c, Y_c). X_c is the x-coordinate of the center, and Y_c is the y-coordinate of the center.

- Area

  The actual number of pixels in the region of the word.

- The length of the Major diameter of the ellipse

  Scalar specifying the length (in pixels) of the major (longest) axis of the ellipse that enclose the PAW of the image. See Figure 4.8, the taller axis is measured.

- The length of the Minor diameter of the ellipse

  Scalar specifying the length (in pixels) of the minor (shortest) axis of the ellipse that surrounds the PAW of the image. See Figure 4.8, the shorter axis is measured.

- Theta

  The angle (ranging from -90 to 90 degrees) between the the major axis of the ellipse and x-axis.

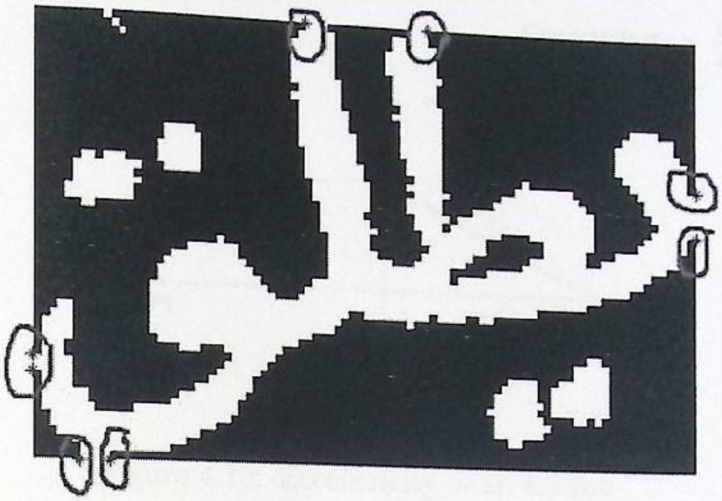


Figure 4.9: Extreme points

- Rectangle

  The smallest rectangle containing the PAW.

- Boundary

  The distances between each neighboring pair of pixels around the boundary of the PAW.

- Extreme Points

  The eight extreme points of the PAW.(left-bottom, left-top,top-left,bottom-left,top-right, right-top, right-bottom bottom-right). See figure 4.9

- Eccentricity of an Ellipse

  "A measure of how out of round an ellipse is" [56]. *Eccentricity* $= c/a$ Where a is the distance between a focus and a vertex,and c is the distance between the center and a focus. See figure 4.10 [56]. Figure 4.11 describes how increasing the eccentricity affects the ellipse. [50].

- Circle Diameter

  The diameter of a circle containing the PAW.

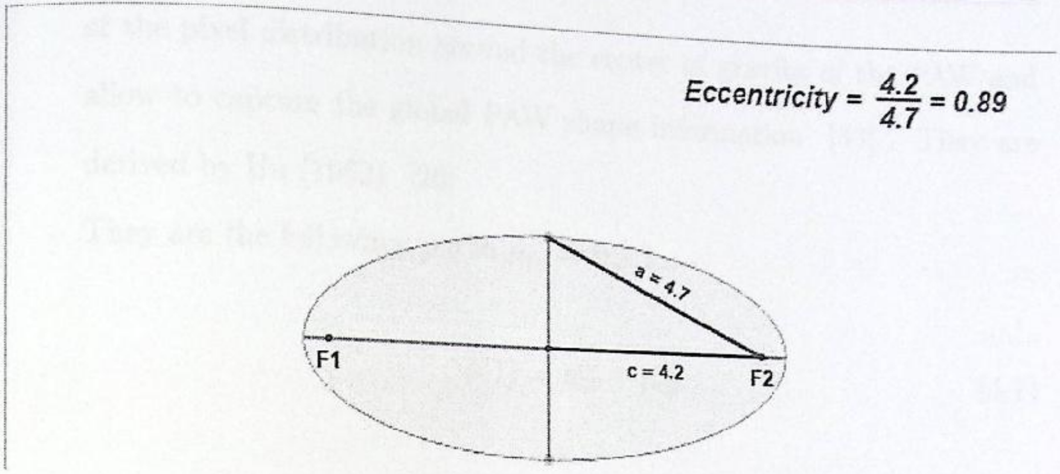


Figure 4.10: Eccentricity of an Ellipse

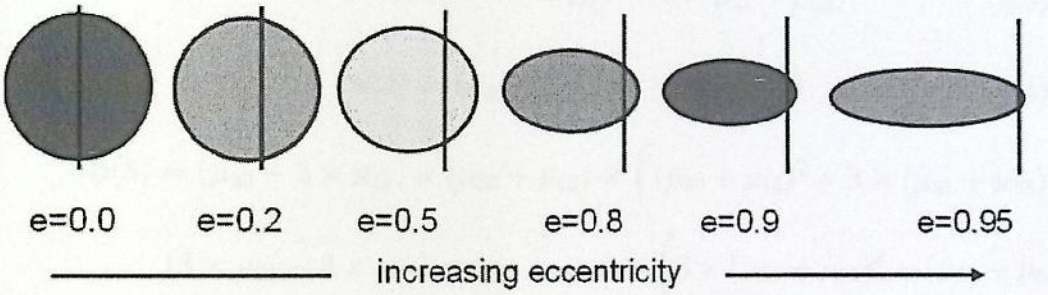


Figure 4.11: Increasing Eccentricity of an Ellipse

- Aspect ratio

  The ratio between major axis and the minor axis.

- Extent of pixels

  The ratio between the area of all image to the area of the smallest rectangle that contains the PAW.

- Solidity

  The ratio between the area of the pixel to the area of the polygon that contains the PAW.

- 7-Moment Invariant

  The moments invariant are well known to be invariant under translation, rotation,scaling and reflection.. They are pure statistical measures of the pixel distribution around the center of gravity of the PAW and allow to capture the global PAW shape information [53] . They are derived by Hu (1962) [26]

They are the following, p,q in $\mu_{pq} = 0,1,2..$:

$$\phi(1) = \mu_{20} + \mu_{02} \tag{4.1}$$

$$\phi(2) = (\mu_{20} - \mu_{02})^2 + 4 \times \mu_{11}^2 \tag{4.2}$$

$$\phi(3) = (\mu_{30} - 3 \times \mu_{12})^2 + (3 \times \mu_{21} - \mu_{03})^2 \tag{4.3}$$

$$\phi(4) = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \tag{4.4}$$

$$\begin{aligned}\phi(5) = (\mu_{30} - 3 \times \mu_{12}) \times (\mu_{30} + \mu_{12}) \times \left((\mu_{30} + \mu_{12})^2 - 3 \times (\mu_{21} + \mu_{03})^2 \right) + \\ (3 \times \mu_{21} - 3 \times \mu_{03}) \times (\mu_{21} + \mu_{03}) \times \left(3 \times (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right)\end{aligned} \tag{4.5}$$

$$\begin{aligned}\phi(6) = (\mu_{20} - \mu_{02}) \times \left((\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right) + 4 \times \mu_{11} \\ \times (\mu_{30} + \mu_{12}) \times (\mu_{21} + \mu_{03})\end{aligned} \tag{4.6}$$

$$\begin{aligned}\phi(7) = (3 \times \mu_{21} - \mu_{03}) \times (\mu_{30} + \mu_{12}) \times \left((\mu_{30} + \mu_{12})^2 - 3 \times (\mu_{21} + \mu_{03})^2 \right) \\ -(\mu_{30} - 3 \times \mu_{12}) \times (\mu_{21} + \mu_{03}) \times \left(3 \times (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right)\end{aligned} \tag{4.7}$$

4.4 Classification

4.4.1 Naive Bayesian classifier

Naive Bayesian classifier is designed for use in supervised learning tasks, in which the performance goal is to accurately predict the class of test instances and in which the training instances include class information. It is a probabilistic classifier depends on Bayesian theorem [34].See chapter 2.

$$P(c_j|x_1, x_2, ...x_n) = P(x_1, x_2, ...x_n|c_j)P(c_j) \tag{4.8}$$

Where $P(c_j)$ is the class probability,and $P(x_1, x_2, ...x_n|c_j)$ is the probability of vector x belongs to class c_j [34].

Gaussian Naive Bayesian classifier

When dealing with continuous features, and these features are normally distributed for each class, the $P(x_1, x_2, ...x_n|c_j)$ can be computed using the equation 4.9 [71] of a Normal distribution . Where m_j is the mean of the features associated with class c_j, cov_j is the covariance matrix of the features associated with class c_j.

$$P(x_1..x_n|c_j) = \log c_j - \frac{1}{2}\log|cov_j| - \frac{1}{2}(x - m_j)^t cov_j^{-}1(x - m_j) \tag{4.9}$$

Naive Bayesian classifies data in two steps:

- Training step: computing the parameters of the probability distribution for each class.

- Predicting step: for unseen test data, the classifier finds the posterior probability of that data belonging to each class. The class of the largest posterior probability is chosen as the recognized class.

4.5 Selecting best features

Feature selection simply, selecting relevant features from the feature lists. This process omitting irrelevant features that affect the accuracy of the results. Irrelevant features allow the redundancy and over-fitting.

In machine learning there are several algorithms and methods for feature selection that give good results see section 3.3. In our thesis we built a small simple algorithm to choose best features. We have limited time so we can't study and use the existing machine learning- feature selection algorithms. The proposed algorithm is summarised in Algorithm 1

4.5.1 Algorithm

Next the algorithm of the feature selection is described

- Combination Step

  We combine the feature list up to NC combinations, $NC = 3$. For $NC = 1$ the feature list has $featureList = \{f_1, f_2, f_3, \cdots, f_d\}$ where d is number of features.

  For $NC = 2$, the feature list then contains a combination of two features. $feature_list = \{(f_1, f_2), (f_1, f3) \ldots (f_{d-1}, f_d)\}$

  For $NC = 3$ the feature list has combination of three features $feature_list = \{(f_1, f_2, f_3), (f_1, f_2, f_4), \ldots (f_{d-1}, f_d d-2, f_d)\}$

- Selecting Initial Features

  For each previous feature list ($NC = 1, 2, 3$), we test all of its features to

Algorithm 1 The proposed algorithm

Combination Step

```
  Input: The feature list (FeatureList)
    Number of combination(NC)
  Output: Combination of features
for c=1 to NC do
  comp(c)= $(combinationoffeatures)_c$ c :number of features in combination
end for
```

Selecting Initial Features

```
  Input: (Dataset)
    Number of folds(N)
  Output: Intial Features
for c=1 to NC do
  featureList=[comp(c)]
  for i=1 to size(featureList) do
    Train classifier based on the training data for N folds using i
    Classify the testing data for N folds using i
    rates(i)= success rate for feature(i)
  end for
  [selected(c)]=select the features based on best rates
end for
select the initial features that have the highest rate in selected(c)
```

Constructing feature vector

```
  Input: Initial Features
    features=FeatureList - Initial Features
    Feature Vector=Initial Features
  Output: Feature Vector
for each f in features do
  testedFeature=Initial Features + f
  Train classifier based on the training data for N folds using testedFeature
  Classify the testing data for N folds using testedFeature
  rates= success rate for testedFeature
  if rates > maxRecentlyRate then
    Add f to Feature Vector
  end if
end for
```

check their performance. We train and test the classifier using N folds according the inputted feature in each list. The evaluating is done using success rate. We choose the best combination in each feature list(the maximum success rate). We have three results the best when $NC = 1$

,$NC = 2$, and $NC = 3$. The best of them is chosen and is considered the initial features.

- Constructing Feature Vector

  We want to examine all features to remove the irrelevant features that affect the results. First, we put the initial features of the previous step to the feature vector, then each feature is added to this vector, and the success rate is computed for the new vector, if the rate is enhanced, then keep it otherwise remove this feature form the vector. Finally, we got the feature vector that contains best features.

Chapter 5

Experiments and Results

In this chapter we present the experimental results of applying our approach. First section describes the environments that were used, second section explains what the results of the classifier mean, then last section discuss the results before and after applying feature selection algorithm.

5.1 Environments

The database that we built contains 50 classes, each class has 10 samples, so we have 500 samples. Section 4.1 describes the database in details. There are 32 features (structural and statistical) were extracted from each sample. The size of the database is 500*32.

N-folds cross validation for testing and comparing the results. We used the Matlab Statistics Toolbox for the Naive Bayesian classifier implementation [44].

5.2 Results for classifier

Naive Bayesian classifier is a probabilistic classifier that gives the results as a matrix that contains probabilities for all samples. Table 5.1 discusses the meaning of the output of this classifier. As we can see, there is a matrix contains samples, classes, and probabilities. We have m classes, class1, class2, class3, class m. The samples are n samples, sample1, sample2, sample3, sample n. For example sample1 has 0.22 probability as class1, 0.44 probability as class2, 0.80 as class3, and 0.01 as class m. By selecting the maximum probability which is 0.80, the predicting class for this sample is class3. If we don't want to depend on the maximum value only, but also the second maximum value, the results will give class3,then class2 (0.80,044) are the candidates classes for sample1, and so on.

Table 5.1: Naive Bayesian classifier probabilities

Sample	Class 1	Class 2	Class3	Class m
Sample 1	0.22	0.44	0.80	0.01
Sample 2	0.90	0.77	0.55	0.80
Sample 3	0.70	0.67	0.40	0.30
Sample 4	0.66	0.33	0.20	0.50
Sample n	0.23	0.37	0.55	0.42

Table 5.2 shows the results depending on number of probabilities (posteriors).

$$Success\ Rate = \frac{Number\ of\ samples\ that\ are\ correctly\ classified}{Total\ samples} \tag{5.1}$$

Table 5.2: Naive Bayesian classifier posteriors testing

Number of posteriors	Success rate
1	0.8543
2	0.9202
3	0.942
4	0.9600

5.3 Results from all features

The success rate for all features that are discussed in section 4.3 is 0.8527. This results taken from one posterior.

5.4 Results after selecting best features

After we applied our algorithm that is proposed in section 5.4 to select best features. The results were as followed.

5.4.1 Combination

There are 32 features, We make combination of one feature (comb1), the combination of two features(comb2), and combination of three features(comb3). The success rate is calculated for each group of combination to select initial features for next step.

5.4.2 Selecting initial features

The initial features are the features that gave the maximum rate. We calculate the maximum for each combination and then choose the maximum of them. Table 5.3 represents the maximum of comb1, table 5.4 represents the maximum of comb2, and table 5.5 represents the maximum of comb3. The initial features are feature number 4,1, and 10.

Table 5.3: Best feature from comb1

Feature number	Success rate
15	0.0371

Table 5.4: Best feature from comb2

Feature number 1	feature number2	Success rate
1	14	0.93

Table 5.5: Best feature from comb3

Feature number1	Feature number2	Feature number 3	Success rate
4	1	10	0.96

5.4.3 Constructing feature vector

The initial features from previous step are feature 1, 4, and 10. For finding the next features,success rate is calculated after adding each feature to the initial vector, if the new rate is larger than previous rate, we preserve this feature otherwise it is removed.

Table 5.6 represents the final feature vector, only seven features are selected. Table 5.7 represents the success rates in which the features are selected as discussed in the algorithm. The best features were:

1. Aspect Ratio:

$$Aspect\ Ratio = \frac{The\,length\,of\,the\,major\,diameter\,of\,the\,ellipse}{The\,length\,of\,the\,minor\,diameter\,of\,the\,ellipse} \tag{5.2}$$

2. Area : Actual number of pixels.

3. Boundary: The distances between each neighboring pair of pixels around the boundary of the PAW.

4. Extreme Points :The eight extreme points of the PAW.(left-bottom,left-top,top-left,bottom-left,top-right, right-top, right-bottom bottom-right).

See figure 4.13.

5. The first Moment invariant feature. See equation 4.1

Table 5.6: Selected Features

f1	f2	f3	f4	f5	f6	f7
4	1	10	3	15	17	26

Table 5.7: Success Rates

S1	S2	S3	S4	S5
0.960	0.9657	0.9714	0.9743	0.9829

5.5 Discussion of results

We have 5o class, each one has 10 samples. These samples were taken from different styles or writers. Class 32 is the worst for recognition which is the character (ذ). The best were (أ) (صطلا)

According to the features we have 32 features. Some of them are invariant under translation, rotation, scaling and reflection such as 7-Moment invariant, and extreme points. Other features are not like theta, solidity and extent.

Chapter 6

Conclusion and Future Work

In this chapter, we summarize the contributions of this thesis. The results are discussed and analyzed. Finally, we present future research directions.

6.1 Conclusion

In this thesis, we have conducted research on offline Arabic handwritten recognition for historical manuscripts. The Automation of the offline handwritten recognition has many applications, such as cheques processing, writer identification and verification, mail sorting, digitizing historical manuscripts and so on.

There are several issues that researches of the Arabic handwritten recognition need to consider. Cursive written of the scripts, absence or displacement of dots, overlapping ligatures and characters, and lack of databases, and so on. In case of historical manuscripts, there exist additional challenges, such as large variation in handwriting and noise distribution models. All these issues caused the lack of research on offline Arabic handwritten recognition.

Previous research on Arabic handwritten recognition has focused on recognition for contemporary handwritten. Very few efforts have been reported on the recognition of handwritten historical manuscripts. Moreover, researches are based on statistical methods, such as Hidden Markov Models (HMM), Neural Networks, and so on.

In this work, we address Arabic handwritten recognition for historical manuscripts. Moreover, our approach is based on a probabilistic method not statistical. We used a Naive Bayesian classifier for training and recognition.

In the following we present the contributions of the thesis to offline Arabic handwriting recognition using a probabilistic classifier for historical manuscripts.

- The classification method, using a probabilistic classifier will give the posterior probability for each class, not only one result which is the proposed class. This will enhance the results by predicting not only the highest posterior but also the first two or three posteriors. We can depend on more than one posterior probability for each tested sample. This idea has shown good results

- Studying a set of features (structural and statistical). The results showed the strength of these features to represent each PAW. All diacritics are removed before feature extraction. Features include solidity, holes, moment invariant, theta, and so on.

- Feature selection, Applying our algorithm of feature selection before training and classification led to enhance the results. Feature selection is to select appropriate features and remove irrelevant features. By applying this algorithm, the result is a subset of rellevent features is selected. When we test this subset the success rate increased from 0.85% to reach 0.98%.

- Data from historical Arabic manuscripts, data was collected from historical Arabic manuscripts. Which we used to build our small database.

6.2 Future Work

In this section we provide some directions of future research for our work.

- Build a larger database, we don't have a benchmark database for Arabic handwritten historical manuscripts. the database that we built is small. In the future we can enlarge the database by collecting other classes.
- Studying a gray scale-features, in this work we extracted features from the binary images. In the future we may test features using gray scale images, taking into account not all features can work under gray scale images, some features need binarization before extraction.
- We may extend the work to test word-level. The input to the system is a single word. Then we may use a lexicon for results verifications.
- We may utilize from the probability of the naive Bayesian classifier by suggesting other PAWs or words for the entered PAWs or words. The suggestion is done using the database of the system.

Bibliography

[1] Arabic language. http://en.wikipedia.org/wiki/Arabic$_l$*anguage*, 2014.

[2] Sayed F Ayman Abd AlAziz Ahmad, Gheith Mervat. Recognition for old arabic manuscripts using spatial gray level dependence (sgld). *Egyptian Informatics*, 2011.

[3] Ani A Al. Feature subset selection using ant colony optimization. *Int J Comput Intell 2(1):5358*, 2005.

[4] Usher .M Al-Emami . S. On-line recognition of handwritten arabic characters. *IEEE*, 1990.

[5] Somaya A S Al-Maadeed. *Recognition of Off-line Handwritten Arabic Words.* PhD thesis, The University of Nottingham, 2004.

[6] Pal U Alaei A, Nagabhushan P. Fine classification of unconstrained handwritten persian/arabic numerals by removing confusion amongst similar classes. *In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR)*, 2009.

[7] Jurgen Schmidhuber Alex Graves. Offline handwriting recognition with multidimensional recurrent neural networks. *MIT*, 2009.

[8] Adnan Amin. Recognition of hand-printed characters based on structural description and inductive logic programming. *IEEE*, 2001.

[9] Chung YY Liu S-L Bae C, Yeh W-C. Feature selection with intelligent dynamic swarm and rough set. *Expert Syst Appl 37:70267032*, 2010.

[10] A.Kacem Ben Cheikh, A Belad. A novel approach for the recognition of a wide arabic handwritten word lexicon. *IEEE*, 2008.

[11] Bapi RS Bhavani SD, Rani TS. Feature selection using correlation fractal dimension: issues and applications in binary classification problems. *Appl Soft Comput 8:555563*, 2008.

[12] Hani Safadi Bilal Alsallakh. Arapen: an arabic online handwriting recognition system. *IEEE*, 2006.

[13] Raluca Brehar. Morphological operations on binary images, 2013.

[14] Varnoosfaderani M. R Broumandnia A., Shanbehzadeh J. Persian/arabic handwritten word recognition using m-band packet wavelet transform. *BROUMA*, 2008.

[15] Gorsky .D. Experiments with handwriting recognition using holo-graphic representation of line images. *Patern Recognition*, 1994.

[16] Jamous . H Daifallah . K, Zarka . N. Recognition-based segmentation algorithm for on-line arabic handwriting. *International Conference on Document Analysis and Recognition*, 2009.

[17] Kan Deng. *Omega: On-Line Memory-Based General Purpose System Classifier*. PhD thesis, Carnegie Mellon University, 1998.

[18] Pinar Duygulu Esra Ataer. Matching ottoman words: An image retrieval approach to historical document indexing. *ACM*, 2007.

[19] M. Cheriet F. Menasri N, Vincent E. Augustin. Shape-based alphabet for off-line arabic handwriting recognition. *n. In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007.

[20] Fleury J Gomez-Verdejo V, Verleysen M. Information-theoretic feature selection for functional data classification. *Neurocomputing 72:35803589*, 2009.

[21] Krishnamurthy . R Govindaraju . V. Holistic handwritten word recognition using temporal features derived from off-line images. *Patern Recognition*, 1996.

[22] Aghagolzade A Harifi .A. A new pattern for handwritten persian/arabic digit recognition. *Int. J .Inf. Technol.*, 2004.

[23] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *In NIPS*. MIT Press, 2005.

[24] Sherif Abdelazeem Hesham M. Eraqi. Hmm-based offline arabic handwriting recognition using new feature extraction and lexicon ranking techniques. *Acm*, 2010.

[25] Cho S-B Hong J-H. Efficient huge-scale feature selection with speciated genetic algorithm. *Pattern Recognit Lett 27:143150*, 2006.

[26] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE*, 1962.

[27] SHOICHIRO YAMAGUCHI HUSSEIN ALMUALLIM. A method of recognition of arabic cursive handwriting. *IEEE*, 1987.

[28] Ericson Davis Anna Borovikov Kristen Summers Ilya Zavorin, Eugene Borovikov. Combining different classification approaches to improve off-line arabic handwritten word recognition. *SPIE-IST*, 2008.

[29] Dr. International. Middle eastern language issues, 2001.

[30] Rohit Prasad Jin Chen, Huaigu Cao. Gabor features for offline arabic handwriting recognition. *Acm*, 2010.

[31] Shridhar M Chen Z Kimura F, Tsuruoka S. Contextdirected handwritten word recognition for postal service applications. *the USPS Advanced Technical Conference*, 1992.

[32] S. B. Kotsiantis. Feature selection for machine learning classification problems: a recent overview. *Springer*, 2011.

[33] S. B. Kotsiantis. Feature selection for machine learning classification problems: a recent overview. *Springer*, 2011.

[34] George H John Pat Langley. Estimating continuous distributions in bayesian classifiers. *the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[35] Anthony L Lashkia G. Relevant, irredundant feature selection and noisy example elimination. *IEEE Trans Syst Man Cybern B Cybern 34(2):888897*, 2004.

[36] Bruno Taconet Laurence Likforman-Sulem, Abderrazak Zahour. Text line segmentation of historical documents: a survey. *Springer*, 2006.

[37] Gernot A Fink Leonard Rothacker, Szilard Vajda. Bag-of-features representations for offline handwriting recognition applied to arabic script. *IEEE*, 2012.

[38] Venu Govindaraju Liana M. Lorigo. Offline arabic handwriting recognition: A survey. *IEEE*, 2006.

[39] Alimi A M. An evolutionary neuro-fuzzy approach to recognize on-line arabic handwriting. *International Conference Document Analysis and Recognition*, 199.

[40] Monji Kherallah Adel M Alimi Mahdi Hamdani, Haikal El Abed. Combining multiple hmms using on-line and off-line features for off-line arabic handwriting recognition. *IEEE*, 2009.

[41] Mohamed Ali Mahjoub, Nabil Ghanmy, Khlifia Jayech, and Ikram Miled. Multiple models of bayesian networks applied to offline recognition of arabic handwritten city names. *CoRR*, 2013.

[42] Haikal El Abed Mario Pechwitz, Volker Maergner. Comparison of two different feature sets for offline recognition of handwritten arabic words. *In Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006.

[43] Volker Maergner Mario Pechwitz. Hmm based approach for handwritten arabic word recognition using the ifn/enit - database. *IEEE*, 2003.

[44] The Mathworks. Statistical toolbox r2013a. `http://www.mathworks.com/help/stats/naivebayes-class.html?refresh=true`.

[45] Dr. Ing. Cristian Mihaescu. Naive-bayes classification algorithm, 2014.

[46] Sabri A. Mahmoud Mohammad Tanvir Parvez. Offline arabic handwritten text recognition: A survey. *ACM*, 2013.

[47] Mokhtar Sellami Abde Ennaji Nabiha Azizi, Nadir Farah. Using diversity in classifier set selection for arabic handwritten recogniti. *Springer*, 2010.

[48] Fatos T. Yarman-Vural Nafiz Arica. Optical character recognition for cursive handwriting. *IEEE*, 2002.

[49] Fatos T. Yarman-Vural Nafiz Arica. Optical character recognition for cursive handwriting. *IEEE*, 2002.

[50] Swinburne University of Technology. Orbital eccentricity.

[51] Lecourtier Y Paquet T. Handwritten recognition: Application on bank cheques. *Proceedings of the 1 st International Conference on Document Analysis and Recognition (ICDAR)*, 1991.

[52] Rohit Prasad Ehry MacRostie Krishna Subramanian Prem Natarajan, Shirin Saleem. Multi-lingual offline handwriting recognition using hidden markov models: A script-independent approach. *Springer*, 2008.

[53] R. J. Ramteke. Invariant moments based feature extraction for handwritten devanagari vowels recognition. *International Journal of Computer Applications*, 2010.

[54] Chafic Mokbel Ramy Al-Hajj Mohamad, Laurence Likforman-Sulem. Combining slanted-frame classifiers for improved hmm-based arabic handwriting recognition. *IEEE*, 2009.

[55] Chafic Mokbel Ramy El-Hajj, Laurence Likforman-Sulem. Arabic handwriting recognition using baseline dependant features and hidden markov modeling. *IEEE*, 2005.

[56] Math Open Reference. Eccentricity an ellipse, 2009.

[57] Sargur N Regean Plamondon. Online and offline handwriting a comprehensive survey. *IEEE*, 2000.

[58] Chris Riley. Hand-print or handwriting, makes a big difference, 2014.

[59] Farag M S. Handwrittentext recognition system for automatic reading of historical arabic manuscripts. *International Journal of Computer Applications*, 2012.

[60] Khorsheed M S. Recognising handwritten arabic manuscripts using a single hidden markov model. *Elsevier*, 2003.

[61] Khorsheed M S. Hmm-based system for recognizing words in historical arabic manuscript. *International Journal of Robotics Automation*, 2007.

[62] Piramuthu S. Evaluating feature selection methods for learning in data mining applications. *Eur J Oper Res 156:483494*, 2004.

[63] H. Amiri S. Snoussi Maddouri. Combination of local and global vision modelling for arabic handwritten words recognition. *IEEE*, 2002.

[64] Majid Ziaratban Saeed Mozaffari, Karim Faez. Structural decomposition and statistical description of farsi/arabic handwritten numeric characters. *IEEE*, 2005.

[65] Hamid Amiri Sameh Masmoudi Touj, Najoua Essoukri Ben Amara. A hybrid approach for off-line arabic handwriting recognition based on a planar hidden markov modeling. . *In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR07)*, 2007.

[66] Dr. Saed Sayad. Naive bayesian, 2014.

[67] BEDDA . M SEPTI . M. Contribution to the recognition of hand arabic word based on neural network. *IEEE*, 2006.

[68] Noureddine Ellouze Haikal El-Abed Sofiene Haboubi, Samia Maddouri. Invariant primitives for handwritten arabic script: A contrastive study of four feature sets. *IEEE*, 2009.

[69] Bruel T. Design and implementation of a system for recognition of handwritten responses on us census form. *Proceedings of the IAPR Workshop on Document Analysis System*, 1994.

[70] Mao KZ Tang W. Feature selection algorithm for mixed data with both nominal and continuous features. *Pattern Recognit Lett 28:563571*, 2007.

[71] Geogios P Vangelis M, Ion A. Spam filtering with naive bayes - which naive bayes? *Third Conference on Email and Anti-Spam*, 2006.

[72] Ni J Huang S Wang Y, Li L. Feature selection using tabu search with long-term memories and probabilistic neural networks. *Pattern Recognit Lett 30:661670*, 2009.

[73] AbdelMajid Ben Hamadou Yousri Kessentini, Thierry Paquet. Off-line handwritten word recognition using multi-stream hidden markov models. *ScienceDirect*, 2010.