CHARACTER RECOGNITION SYSTEM FOR ARTISTIC ARABIC STYLE

By

Abed Alssalam Aqel Khaleel Abu Odeh

Supervisor

Dr. Moayad A. Fadhil

This Thesis Was Submitted In Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduate Studies

Philadelphia University

December-2008

Committee Decision

Successfully defended and: approved on _____

Examination Committee	Signature
Dr Moayad A. Fadhil Chairman Academic Rank: Associated Prof.	
Dr. Said Ghoal member. Academic Rank Professor	
Dr. Rashid Al-Zubaidy member. Academic Rank: Associative Prof	
Dr. Venus W. Samawi. External Member. Academic Rank: Associated Prof	

جامعة فيلادلفيا

نموذج تفويض

أنا عبد السلام عقل خليل أبوعودة ، أفوض جامعة فيلادلفيا بتزويد نسخ من رسالتي للمكتبات أو المؤسسات أو الهيئات أو الأشخاص عند طلبها.

التوقيع :

التاريخ :

Philadelphia University

Authorization Form

I, Abed Alssalam Aqel Khaleel Abu Odeh, authorize Philadelphia University to supply copies of my thesis to libraries or establishments or individuals upon request.

Signature:

Date:

CHARACTER RECOGNITION SYSTEM FOR ARTISTIC ARABIC STYLE

BY

ABED ALSSALAM AQEL KHALEEL ABU ODEH

SUPERVISOR

DR. MOAYAD A. FADHIL

THIS THESIS WAS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

MASTER'S DEGREE

IN

COMPUTER SCIENCE

DEANSHIP OF ACADEMIC RESEARCH AND GRADUATE STUDIES

PHILADELPHIA UNIVERSITY

December-2008

Committee Decision

Successfully defended and: approved on _____

Examination Committee	Signature
Dr Moayad A. Fadhil — Chairman Academic Rank: Associated Prof.	
Dr. Said Ghoal — member. Academic Rank Professor	
Dr. Rashid Al-Zubaidy — member. Academic Rank: Associative Prof	
Dr. Venus W. Samawi. External Member. Academic Rank: Associated prof	

Dedication

I fully dedicate this thesis to my whole family, my professors, my friends, and to all learning students.

Abed Alssalam Abu Odeh

Acknowledgment

I would like to express my regards and appreciation to Dr. Moayad A. Fadhil, who has evaluated my work from the beginning and supported me, and to all of those who have helped and encouraged me.

Abed Alssalam Abu Odeh

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning of Abbreviation
ACR	**Automatic Character Recognition**
BMP	**Bitmap Format**
DP	**Dynamic Programming**
DPI	**Dots Per Inch**
HT	**Hough Transform**
HTV	**Height Threshold Value**
OCR	**Optical Character Recognition**
PCA	**Principal Components Analysis**
TTF	**True Type Font**
WTV	**Width Threshold Value**

Character Recognition System for Artistic Arabic Style

By

Abed Alssalam Aqel Khaleel Abu Odeh

Supervisor

Dr. Moayad A. Fadhil

Abstract

Despite the tremendous achievement which has been made on the reading mechanism in the field of desktop publishing, the door remains open to the great efforts should be done in this field for the Arabic language contrary to the languages based on Latin and Chinese characters.

The characteristics of the Arabic text, such as the complexity of characters and overlapping, and some characters do not accept the contact only one side in addition to the varied forms of characters in accordance with its position in the middle, beginning or the end of the word, all these factors increase the specificity and uniqueness of the research in this area.

The recognition of Arabic characters pass through several steps; the preprocessing, segmentation, feature extraction, recognition and post processing. However, the Arab character recognition depends mainly on the segmentation, because the segmentation has many problems resulting from the characteristics of Arabic characters such as the complexity of characters and overlap, a number of researches have emerged dealing with this problem in the printed and handwritten text, and reached a series of solutions and proposals and recommendations to solve the problem of segmentation.

There are many obstacles facing researchers working on Arabic character recognition mainly in segmentation process such as overlapping, ligatures and over-segmentation.

So, this thesis address the subject of this research concerning segmentation of artistic Arabic characters, using Thulth font that have the mentioned obstacles above, trying to segment the artistic words into characters and recognizing these characters.

Segmentation based on colors have been used in the research to solve the problem of the segmentation, which has passed into three steps, the first step coloring each connection part by a separate color, the second step segmentation of colored text to connection parts, and the third step segmentation of each connection part into its characters. The outcome of the segmentation gave successful results.

There are several ways of character recognition, such as Matrix matching and feature extraction. Matrix matching has been used because it is widely used and the easiness of its usage.

Matrix Matching compares the scanned picture of the character with the matrix library or pattern characters, and if the scanned picture of the character match one of the pictures in the matrix library within a known limit of similarity, the program considers that picture as identical ASCII character.

During the application of the system, the result of recognition was high and exceeds 89.19 % for the words tested in the system which are around 400 words.

CHAPTER ONE
INTRODUCTION

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Arabic language is the mother of Semite languages. The antiquities discovered of the Arabic script belong to late period of the history of the Arabic language. The Arabic language primarily inherited through speaking before it inherited through writing.

The importance of manipulating the Arabic language by computer is not luxury or secondary matter, but it is very important because the future of the language and Arab stature in the culture depends on it, also their future economic and scientific situation.

Many new fields appeared under the umbrella of artificial intelligence, processing of natural language was the most important field. This field has been developed to include many aspects of processing natural language such as writing, correct spelling, grammar, vocabulary, interpretation, and also eloquence and poetry. The field of identification forms, images and models involved in many applications such as identification of sites, military applications, mineral prospecting, sailing ships in the oceans, space and so on. The recognition of writing is one of these applications.

1.2 History of Arabic Type Evolution

The origin of Arabic script alphabets belong to the first alphabets created by the Phoenicians that were living on the coastal areas of the Mediterranean such as Lebanon, Syria, Palestine, which affected at all the Mediterranean nations.

The early Arabic alphabet created in the Kufa (Iraq), where it consists of 17 characters without dots or accents, and then diacritic dots and accents were putted to help pronunciation, which lead to the increase of characters to 29 characters.

By the emerging of Islam, the Holy Quran was the main reason of reformation of all Arabic scripts present in Arabic countries. In the seventh century A.D. a new Arabic script developed consist of 29 characters to enhance the Quranic scripts, where Holy Quran was written firstly using Quranic Kufi then by using Quranic Naskh Styles.

After that, Arabic script spread all over Middle East, Northern Africa, even reached Spain, because of the Islamic conquests. In the different Arab cities many new scripts and many writing techniques were developed such Kufi, Thuluth, Diwani and Diwani Djeli, Naskh, Persian, Ruqaa and Maghrebi. Figure (1-1) illustrates some of those types of scripts.

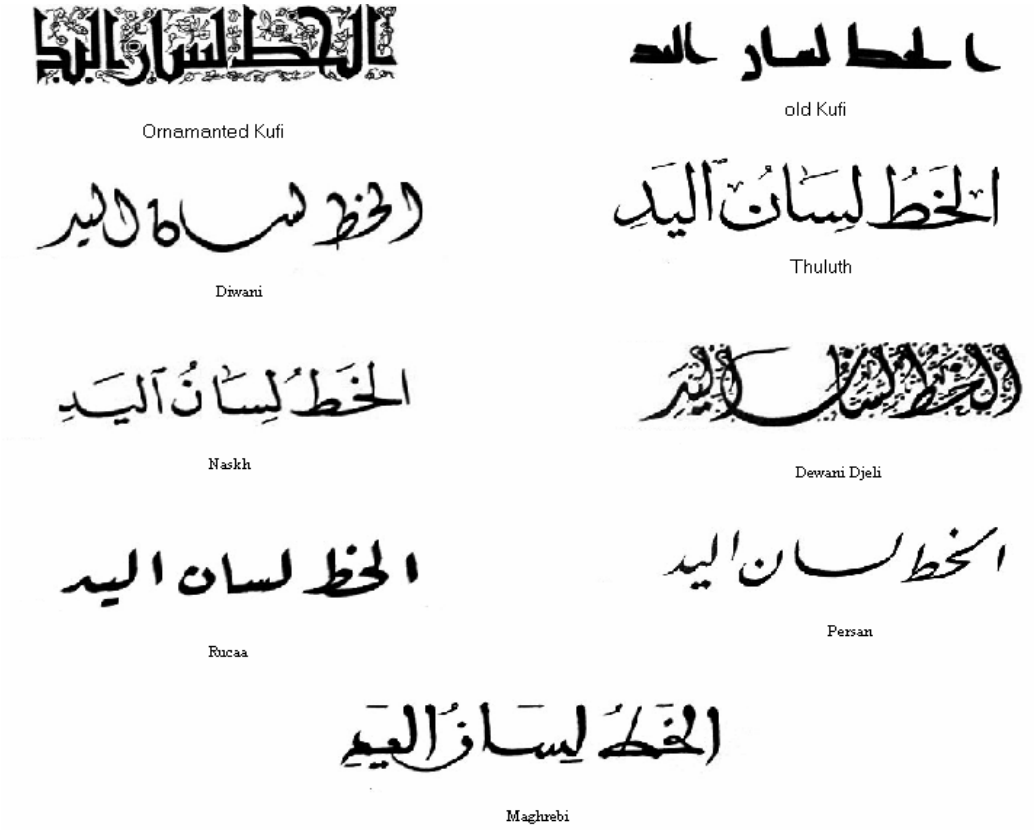


Figure (1-1): The most known Arabic calligraphy styles

The Arab characters consist of 29 characters and 11 vocalization mark in the shape of accent, the basic construction of the character can be found in 19 basic shapes, because of the changing of the character shape upon its position in the word (initial, medial, final, and isolated), accordingly the set of characters shapes would be 106 shapes including 23 characters each of which has 4 alternative shapes, and 7 characters including hamzah each of which has 2 alternative shapes, in addition to (لا) character the total would become an 108 shapes.

Since the Arabic characters can be used in languages other, not only Arabic, such as Afghanistan, Iran, Malta, India, Pakistan, Kazakhstan, some Malaysian newspapers, Indonesia and Somalia using the Arabic language. In Africa, the Berber, Hausa, Swahili

languages use Arabic characters sometimes. For this reason of that many amendments will be on the characters to represent each non-Arab phonetics, Furthermore, if the hand-written characters used the set of characters shapes will start from 130 shapes and ends in hundreds or more.

The four typographic elements of the Arabic script can be demonstrated in the Figure (1-2) (Schmieg, 2007).

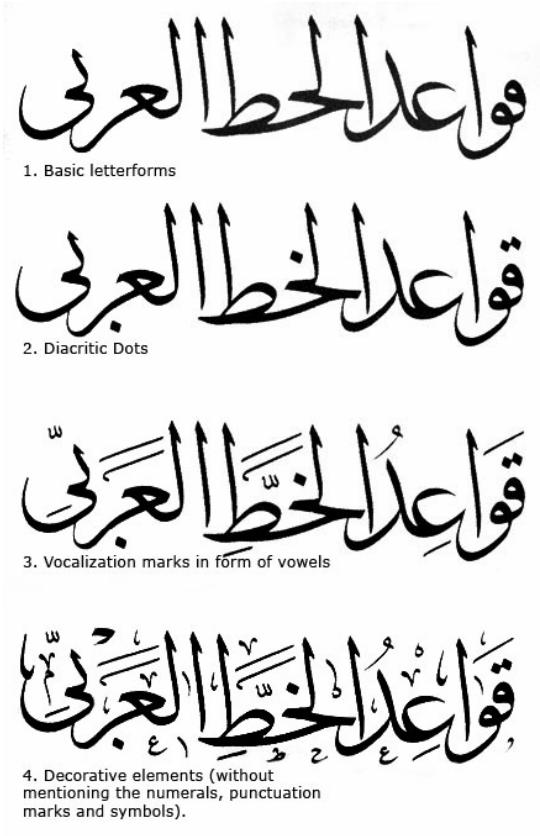


Figure (1-2): Four typographic elements of the Arabic script

1. Basic letter forms.
2. Diacritic dots.
3. Vocalization marks in form of vowels.
4. Decorative elements (without mentioning the numerals, punctuation marks and symbols).

Usually the daily writing uses diacritic dots and some vocalization mark only as shown in the Figure (1-2).

During the second half of the 20th century, shortly after the spread of computers it had been possible to introduce languages that use Latin script to the computer firstly, then, emerged the idea of introducing other languages after several developments in the installation and programming of the computer. The Arabic characters, words and sentences were crammed to the computer through the programs so as to be written as entered, and then this situation had been developed.

The main use of computer in accordance to different languages is still writing and several major ideas should be taken in consideration for Arabic writing by computer. The computer work must be a means to serve the Arabic language and accurately manipulate its various aspects and not to be incapable, and have to adapt to the Arabic language to suit the design of computer itself. This does not mean the stalemate on the current status of the Arabic language, but assigned to scientists and researchers and calligraphers, who are not forced to follow the opinions and assiduity of computer specialists or designers.

The second item should be available is the honesty representation of the Arabic language by computer, this requires full understanding of the possibility of the entry of the Arabic language into a computer, and the possibility of manipulation in details and accuracy in the mean time and all prospects in the future, in addition to the possibility of printing normally taking into consideration the aesthetic situation of the computer output. In another hand, there are four different fields of the relationship between Arabic characters and the computer.

These fields are the data entry, output of data, processing and aesthetic aspect of the Arab characters. The entry of Arabic characters can done by two ways; through keyboard and optical character recognition known as (OCR), the keyboard should be prepared by all facilities to make enter Arabic characters, where OCR devices make a scanning of the page and to recognize the characters through matching the recognized characters with the stored characters in the database already trained, some of these devices has more ability to recognize characters without any beforehand training

The requirements of the output of Arabic characters in terms of the number of characters does not differ from the requirements of input Arabic characters where

reading is for written text, but writing is the opposite of the process of reading. However, there are many items related to the way of writing Arabic. In the beginnings of computer printed Arabic characters have the same shape with addition of decorations sometimes, and then some improvements happened for the shape of the character into word such in the beginning, in the middle, or in the end of the word but the width of all characters were the same.

By the improvements of the output devices (screens) each character has its own width depending on its place in the word as needed, the diacritic considered as characters and have the same width of characters either with kashida " ـ " or not, for example the fatha " ـَ " in the middle of word needed kashida " ـ " while the fatha " َ " on non-connected character from the left such as (واو) did not need kashida, this led to the need of two types of diacritic one with kashida and the other without. However, the evolution of information processing technology by changing the width of the characters enabled putting diacritic over the last character directly, also, enabled merging vocalization marks merging automatically on the same character such as, when there are shadda and fatha on the same characters they will be merged automatically,

In the beauty of Arabic script and with the evolution of technology printing and writing shown on the screen with higher accuracy, especially techniques appeared (Latin and others). One of these techniques so-called True Type Font and symbolizes by (TTF), this technique mainly based on storing the general characteristics of the character needed to be shown on the screen whatever the size of the character is, whereas the shape of the character is the same whatever the size is. This technique enables us to show adjacent characters in any shape needed. Recently many shapes of Arabic script modified in different kinds of Arabic fonts by Arab and Latin programming companies in Lebanon, Saudi Arabia, Egypt and others.

Upon what mention above, the thesis will discuss the recognition of Arabic words through computer specially the Artistic Arabic Fonts, where it is prefer to choose Thulth font for its artistic shape, trying to overcome some of the obstacles facing researchers in Arabic character recognition mainly overlapping and over-segmentation through segmentation based on colors.

1.3 Pattern Recognition

Pattern recognition in the meanwhile is more common and has more importance, where pattern recognition discuss the classification of objects and the ways of classification , also can be considered as a group of techniques have a main role in performing tasks instead performing these tasks manually by human by automatic performing of these tasks through computer. Definitions of pattern recognition varied since the sixties of 20^{th} century till now. Here it is mention the following definitions as examples:

In paper Pattern Recognition: An overview (Liu et Al, 2006), Gonzalez, and Thomas defined pattern recognition as "a classification of input data via extraction important features from a lot of noisy data". But Fukunaga defined pattern recognition as "A problem of estimating density functions in a high-dimensional space and dividing the space into the regions of categories of classes".

Many methods used in pattern recognition such as; statistical pattern recognition, data clustering, application of fuzzy sets, neural networks, structural pattern recognition, syntactic pattern recognition and approximate reasoning approach, the method will be used in the thesis will be discussed in details in the next chapters.

Pattern recognition techniques considered one of most important aspects for many applications, developed and used in many fields such as; artificial intelligence, computer engineering, space navigation, archaeology, geologic reconnoitering, medicine image analysis, nerve biology, armament technology and so on.

For Example, the pattern recognition is used in Optical Character Recognition (OCR), recognition of alphabet and numeric character based on image-input.

Character and numerals recognition procedure consist of many steps, starting by acquisition of the image by scanner or computer, preprocessing of the image such as binarisation, smoothing, and aligning , then segmentation of the image. The complexity of this procedure is the type of script based on either handwritten or printed, also, many other problems presented here mainly regarding the Arabic characters which are written

cursively where the problems of segmentation of the word into characters emerge, after segmentation there is a need for feature extraction for Arabic characters to enable the computer to recognize these characters by one of the recognition methods used to recognize the exact characters of the word.

The font thulth has been used in this thesis for many reasons, some of these reasons are; thulth is one of the most beautiful Arabic fonts specially when there is an overlapping in the text, many publishers use this font in printing the Holy Quran, using thulth in artistic tableaus and books covers, used for the decoration of mosques, used in writing Arabic manuscripts and the complexity of this fonts, these reasons lead to using thulth font to study it in details and to find solution can help in the recognition process of font.

1.4 Contribution

The process of recognition of Arabic characters is very important in many fields such as industrial, commercial, medical, technical and others, from here it was necessary for scientists to engage in this area, researchers have faced several problems in the process of segmentation in order to recognize characters, and that the accuracy of recognition based on the accuracy of segmentation, was working hard to find new ways to solve the problems faced by the segmentation of Arabic characters as the connected characters problem is most important problems facing the segmentation.

There are many obstacles facing researchers working on Arabic character recognition mainly in segmentation process such as; overlapping and over-segmentation.

The contributions of the thesis in this area can be summarized as follows:

- Finding a proper way to segment the artistic characters, which include two main points to solve the overlap exists in the text which are:
  A. Color algorithm development which used to recognize connection parts.
  B. Development of an algorithm to segment the line into connection parts depending on color algorithm where each connection part will be moved respectively from other connection parts, then Segmentation of the connection parts into characters.

1.5 Thesis Layout

The layout of this thesis is as follows:

- Chapter two discusses the overview of Arabic characters characteristics, types and stages of Arabic text recognition system, and literature review.
- Chapter three discusses the proposed algorithms. Segmentation and recognition methods are described in separate subsections.
- Chapter four represents implementation details and results.
- Chapter five discusses results of system and analysis.
- Chapter six includes conclusions and future works.

CHAPTER TWO
LITERATURE REVIEW

Chapter Two
Literature Review

2.1 Character recognition concepts.

The process of character recognition is part of the process of patterns recognition, passed by many researches in the last forty years. the processing and recognition of printed or handwritten characters has been studies, due to the desire to improve communication between machine and human, and because of the lack of commercial products for the process of character recognition on the market because of the diversity of fonts and sizes and different forms of characters. There are many theories proposed to deal with characters learning, classification and recognition such as Intendance-based learning, neural network, rule induction, decision tree, and others.

The process of character recognition relies heavily on the accuracy of the segmentation. If the segmentation process was correct, then the process of recognition will be correct, and vice versa, because of the presence of some problems in the process of segmentation such as the cursive characters in the Arabic language and the variation of forms at the beginning, middle and the end of the word as well as in some other languages. In addition to the mood of the writer and nature of writing, the difficulties of segmentation appeared, so many theories emerged to manipulate the segmentation of printed or handwritten characters such as segmentation methods based on vertical projection, segmentation methods based on the upper distance function, segmentation methods based on the thinned characters and others (Zeki, 2005).

► Overview of Arabic characters characteristics.

Because Arabic text have some properties that make it difficult to be recognized the Arabic characters. For this reason, Arabic characters characteristics will be described by the following points:

1. Arabic texts are cursive and are written from right to left (Mostafa,2004) Therefore, the recognition rate of Arabic characters is lower than that of disconnected characters such as printed English (Amin et al,1996).

2. Arabic character shape can be changed dramatically in different fonts (Mostafa, 2004).
3. An Arabic letter might have up to four different shapes, depending on its position in the word. Table (2-1) shows the variation of shapes of the Arabic characters according to their positions in the word (Cheung et al, 2001).
4. In addition, some Arabic characters have exactly the same shape, and are distinguished from each other only by addition of diacritics, namely: a dot, double dots, or triple dots. These dots may appear above or below the baseline. Or this can be above, below, or inside the character, but never in above and below at the same time (AMIN et al, 1996).
5. Arabic characters have the same shape and differ from each other only in some dots or zigzag bars.
6. It is worth noting that any erosion or deletion of these dots in the scanning process of the document image results in a wrong classification of the character.
7. Each word, machine-printed or handwritten, may consist of several separated sub-words. (A sub-word is either a single character or a set of connected characters).
8. The script consists of separated words which are aligned by a horizontal virtual line called "Baseline". The baseline is a

Table (2-1): The different shapes of the Arabic characters

Name	Isolated	First	Middle	Last
Alif	ا	ا	ا	ا
Baa	ب	بـ	ـبـ	ـب
Taa	ت	تـ	ـتـ	ـت
Thaa	ث	ثـ	ـثـ	ـث
Geem	ج	جـ	ـجـ	ـج
Hha	ح	حـ	ـحـ	ـح
Kha	خ	خـ	ـخـ	ـخ
Dal	د	د	ـد	ـد
Thal	ذ	ذ	ـذ	ـذ
Raa	ر	ر	ـر	ـر
Zain	ز	ز	ـز	ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Saad	ص	صـ	ـصـ	ـص
Dhad	ض	ضـ	ـضـ	ـض
Tta	ط	طـ	ـطـ	ـط
Zha	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Ghain	غ	غـ	ـغـ	ـغ
Faa	ف	فـ	ـفـ	ـف
Gaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Haa	ه	هـ	ـهـ	ـه
Waw	و	و	ـو	ـو
Yaa	ي	يـ	ـيـ	ـي

medium line in the Arabic word in which all the connections between the successive characters take place, Shown in Figure (2-1) (Mostafa, 2004).

9. Arabic words are formed by connecting some letters together to make a word. However, some Arabic characters are not connectable with the succeeding character. Therefore, if one or more of these characters exist in a word, the word is divided into two or more sub-words., as shown in Figure (2-1).
10. Another feature of the Arabic writing is the ligatures. A ligature is the combination of two characters to form a unit shape, as shown in Figure (2-1). Ligatures occur only in some fonts, for example the two characters Noon and Meem can be "□□□" in the "Simplified Arabic" or "□" in the "Traditional Arabic".

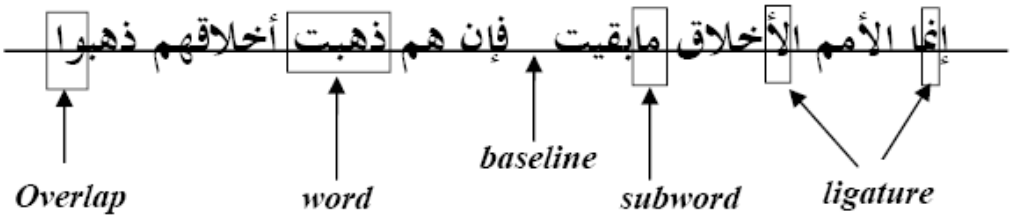


Figure (2-1): Examples of Arabic writing characteristics.

11. Also two letters can be overlapped to raise a problem for the segmentation using simple horizontal projection. Furthermore, characters of the same font have different sizes, i.e. characters may have different widths even though the two characters have the same font and point size.
12. Arabic words may horizontally overlap and characters may stack on others. These induce problems for both the word and the character segmentations. Figure (2-2) demonstrates an example of word overlapping; the dotted box is where the overlapping occurs.

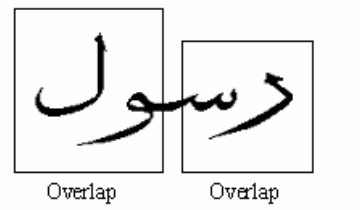


Figure (2-2): An example of overlapping Arabic word.

13. There are 28 characters in the Arabic alphabet. Due to the numerous shapes for character, there are 100 classes to be recognized (Cheung et al, 2001).

14. Many Arabic characters contain a loop, while others contain double loops.
15. Arabic scripts have various styles. Also, each writing style can contain new and compound forms of letters, for example, an unconstrained handwritten Arabic text, the number of separate classes that must be considered will be too many. This makes the recognition process very difficult (Safabakhsh and Adibi, 2005).
16. The distance between a character and the middle baseline or the vertical position of the character, depending on other characters of the word, can be different as shown in figure (2-3).

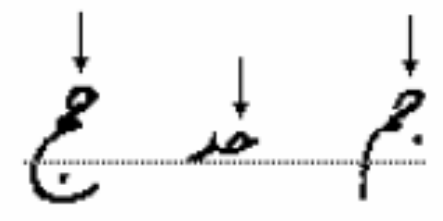


Figure (2-3): Distance between a character and the middle baseline

17. Some characters and compound forms rest on the baseline while others do not. In fact, some parts of words are written in an angle about 30 degrees to the baseline or other as shown in figure (2-4).

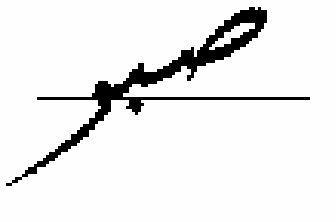


Figure (2-4): A word written in an angle

For these reasons, the difficult recognition of Arabic characters is a result of these characteristics.

► Arabic Text Recognition System

Arabic text recognition can be divided according to the Data Acquisition and the Writing Style as shown in figure (2-5).

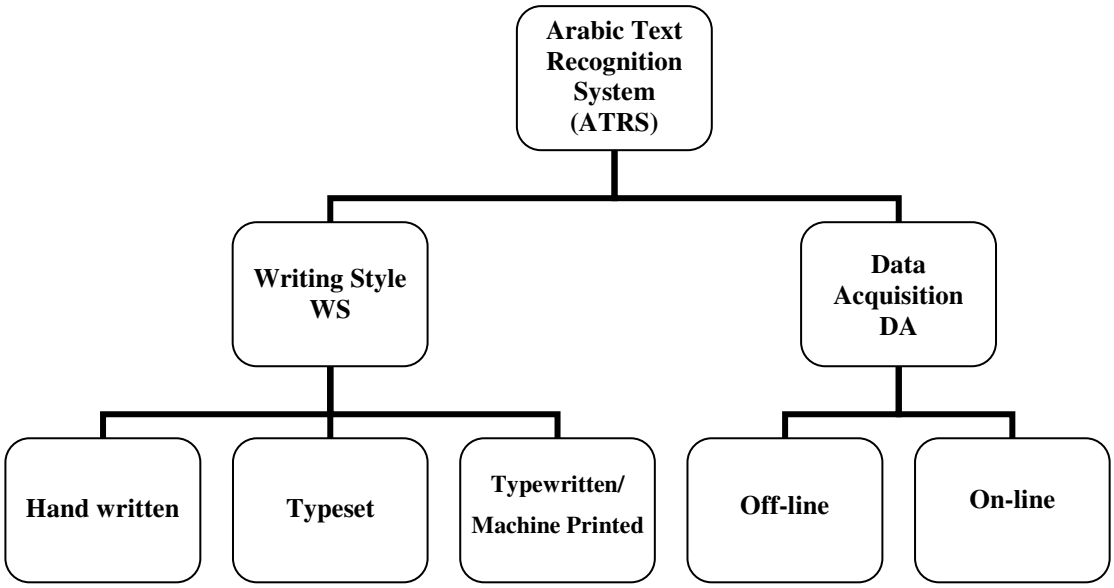


Figure (2-5): Arabic text recognition system

• Systems Classified According to the Data Acquisition

The different approaches covered under the general term character recognition systems is involved in two categories on-line and off-line (AMIN et al, 1996). Each have its own hardware and recognition algorithms.

► On-line Character Recognition Systems

On-line character recognition accepts (x,y) coordinate pairs from an electronic pen touching a pressure-sensitive digital tablet. On-line processing happens in real-time while the writing is taking place. Also, relationships between pixels and strokes are supplied due to the implicit sequencing of on-line systems that can assist in the recognition task.

On-line recognition has several interesting characteristics. First, recognition is performed on one-dimensional data rather than two-dimensional images as in the case of off-line recognition. The writing line is represented by a sequence of dots whose location is a function of time as shown in figure (2-6).

Figure (2-6): Writing character by a sequence of dots (on-line recognition)

The on-line recognition system has two major advantages: the high-recognition accuracy and the interaction; while the disadvantage is limited to recognizing handwritten text (Khorsheed, 2002). The writer requires a special equipment which is not as comfortable and natural to use as pen and paper, and punching is much faster and easier than handwriting for small size alphabet such as English or Arabic (Arica , 1998).

► Off-line Character Recognition Systems

Off-line character recognition takes a raster image from a scanner, digital camera or other digital input source, off-line recognition is performed after the writing or printing is completed, research interest is increasing in this field and some development has been made. However, the performance of even the best handwritten text recognition systems is as yet far from human reading ability (Safabakhsh and Adibi, 2005), the image is binaries using a threshold technique if it is color or gray-scale so that the image pixels are either on (1) or off (0).

Optical Character Recognition (OCR) deals with the recognition of optically processed characters rather than magnetically processed ones. In a typical OCR system, input characters are read and digitized by an optical scanner. Each character is then located and segmented and the resulting matrix is fed into a preprocessor for smoothing, noise reduction, and size normalization. Off-line recognition can be considered the most general case; no special device is required for writing and signal interpretation is independent of signal generation, as in human recognition.

The drawbacks of the off-line recognition are; off-line recognition usually requires costly and imperfect preprocessing techniques prior to feature extraction and recognition stages, and off-line recognition is not real-time recognition.

In general, the online problem is usually easier than the offline problem since more information is available, like the movement of the pen may be used as a feature of the character (Aburas and Rehiel, 2007).

The two domains (offline & online) can be further divided into two areas according to the character itself that is either hand written or printed character.

► Printed Character Recognition

The printed texts includes all the printed materials such as books, newspapers, magazines and documents, the results appear through typewriters, printers or plotters, printed characters can be classified into recognition of a specific font (Fixed font characters recognition), recognition of more than one font, (multi font character recognition), and recognition of any font (omni font character recognition).

► Hand Written Character Recognition

Hand written character recognition, based on the form of written communication, includes two classes, cursive script and hand printed character,

- **Systems Classified According to the writing style**

Writing style may be classified according to complexity into three categories (Khorsheed, 2002):-

1. Typewritten or machine-printed: this is a computer-generated style, and it is the simplest among all styles because of the uniformity in writing a word as shown in figure (2-7).

Figure (2-7): Example of Typewritten Arabic Style

2. Typeset: this is normally used to print newspapers and books. Typeset style is generally more difficult than the machine-printed style, because of the existence of overlaps and ligatures, which poses a challenging problem as shown in figure (2-8). Recently, some computer-generated fonts have imitated the typeset style in providing ligatures and overlaps.

Figure (2-8): Example of Typeset Arabic Writing.

3. Handwritten: this is assumed to be the most difficult style because of the variations in character shape even if it is rewritten by the same person as shown in figure (2-9).

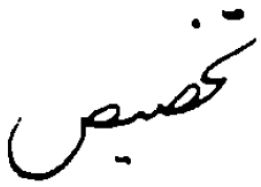


Figure (2-9): Example of Handwritten Arabic Writing.

Penman writing is more careful than a common person handwriting that represents the daily usage of Arabic alphabet by individuals. Few people are able to perform an exquisite handwritten script. A usual handwriting is certainly different from decorative handwriting, which is normally used for adornment purposes.

►Stages of Arabic Text Recognition System

The Arabic text recognition system consists of the following stages:

- **Image Acquisition**

This stage is considered the first step in the recognition system. The main goal is to acquire the text and transform it into a digitized raster image, in off-line, the scanner, can run at 200, 300, or 600 dots per inch (dpi). Lower resolution and poor binarisation can cause readability problem when essential features of characters are deleted or obscured in the image.

- **Preprocessing**

The preprocessing stage attempts to compensate for poor quality originals and/or poor quality scanning. This is achieved by reducing data variations and by reducing both noises "Signal-independent noise adds a random set of grey levels, statistically independent of the image data, to the pixels in the image and in Signal-dependent noise, the value at each point in the image is a function of the grey level there"(Khorsheed, 2002).

The preprocessing stage can be pass into many of stages such as the preprocessing stage can be execute many of processes such as Smoothing, Skew Detection and Correction, Document Decomposition, Slant Normalization, Thinning and Skeletonization.

Apply the smoothing stage to reduce the noise in an image using mathematical morphology operations. Two operations are mainly used, Opening (small gaps or spaces between touching objects in an image) and Closing (small gaps in an image). The Opening will break narrow isthmuses and eliminate small islands. The closing will eliminate small holes on the contour.

Skew Detection and Correction is using to scanning a document so that text lines are within about three degrees of the true horizontal is acceptable. Skew detection, is used to estimate the orientation angle, the skew angle, of the text lines. Skew correction is process of rotating the document with the skew angle, in the opposite direction.

Hough Transform is one of methods that used to estimate the skew angle, and from the other methods that use to estimating a skew angle is based on using bounding boxes of Connected Components.

A document image consists of blocks of text that are interspersed with tables and figures. The document decomposition and structural analysis task can be divided into three phases (Khorsheed, 2002).

- Phase one consist of block segmentation where the document is decomposed into several rectangular blocks. Each blocks containing a text, an image, a diagram or a table.
- Phase two consist of block classification. Each block is assigned a label (title, regular text, picture, table, etc).
- Phase three consists of a logical grouping and ordering of the blocks.

The classical method for identifying text lines in an Arabic text image is to use a fixed threshold to separate the pairs of consecutive lines, second approach is to use the horizontal projection and look for the pixel lines that have a density of zero, then consider that every text line is situated between two lines which contains white pixel density, another attempt at decomposing the Arabic script into words is based on the connected components of that script.

The character inclination that is normally found in cursive writing is called slant. Slant correction is an important step in the preprocessing stage of both handwritten words and

numeral strings recognition. The general purpose of slant correction is to reduce the variation of the script and specifically to improve the quality of the segmentation candidates of the words or numerals in a string, which in turn can yield higher recognition accuracy (Cheriet, 2007).

The thinning and Skeletonisation are operations that produce the skeleton. A skeleton is presumed to represent the shape of the object in a relatively small number of pixels, all of which are structural and have semi-equal distance from two or more contour points.

Thinning algorithms may be classified into parallel (on all pixels simultaneously) and sequential (examine pixels and transform them, depending on the preceding processed results).

The approach in both cases is to remove the boundary pixels of the character that are neither essential for preserving the connectivity of the pattern, nor for representing any significant geometrical feature of the pattern. There are many algorithms tackle thinning and Skeletonisation (Khorsheed, 2002).

- **Segmentation**

Segmentation is segmenting the word into sub-word (combination of two characters or more-connection parts), character, and strokes (part of characters), it is not easy to segment it directly into perfect characters by the computer, and these hardly are hesitating from ligature or destroyed characters.

Ahmed M. Zeki (Zeki, 2005) is divided Segmentation methods into categorized based on the techniques used.

1. Segmentation methods based on vertical projection

The aim of the projection method is to reducing two-dimensional information into one-dimension, to simplify a system of character recognition, it works better with printed documents, especially with fonts which do not form ligatures, and these methods are based on the fact that the connection stroke is always of less thickness than other parts of the words. In these methods the vertical and horizontal projections of the image are obtained.

The horizontal projection is useful in separating the lines and finding the text baseline, while the vertical one helps in segmenting the words, sub-words and characters.

The segmentation methods that use the vertical projection histogram depend greatly on the determination of the baseline. They are independent on the shape, size or font of characters as far as the font contains no overlapping. They are best suited for machine printed characters, while proved inadequate for segmenting overlapping characters or handwritten script because the connection points are not along the baseline due to such data frequently contain undulations and shifts in the baseline, baseline-skew variability and inter-line distance variability. Moreover, this approach will not work effectively for skewed images. However, not all sub-words can be separated by this method.

Special treatment in a later step is required to separate overlapped characters and to recombine the strokes resulted because of the over-segmentation.

2. **Segmentation methods based on the upper distance function**

The set of the highest points in each column called upper distance function. For each upper distance function, you can determine the baseline of each sub-word, then the distance between the baseline and the top of this sub-word is measured, then one of three tokens (up, middle and down) are given to each point. The tokens related to the vertical distances between the baseline and the top of this column of each point and the vertical distance of the previous point, by using a special grammar, and then you can find the connection points.

The researchers reported that the advantage of this method is that the character can be obtained completely in a single piece, hence the number of different shapes is minimal, and this facilitates the recognition, this method is insensitive to scale, it is very fast, and it is multi-font and multi-style.

3. **Segmentation methods based on the thinned characters**

These methods is used to extract skeletons, for example the baseline of the thinned word is found first, and then only those columns that have no pixels above or below the baseline are considered in finding the segmentation points. The segmentation point will be in the middle of the connection segment.

The disadvantages of this method, is that different thinning algorithms may produce different thinned characters, moreover, the thinning process might alter the shape of the character, especially in the case of poor quality characters, which in turn makes it difficult to be recognized. Some of the common problems encountered during the thinning process include the elimination of vertical notches in some characters and elimination or erosion of secondary characters. These modifications make the recognition of the thinned image a difficult task.

4. Segmentation methods based on contour tracing

Segmentation also achieved by tracing the outer contour of a given word. The segmentation method based on the outer contour of the main body of the words, according to outer contour determine (start , end) upper contour, lower right point and the lower left point of the contour, then a segmenting of the upper contour into parts. The researchers used different ways to get a better segmentation for Arabic characters.

The advantage of this method is avoiding all problems resulting from the thinning process because it analyzes the structural shape of characters as they have been scanned. However, in many cases, the contour needs to be smoothed first.

5. Segmentation methods based on line adjacency graph

This method is based on the topological relation between the baseline and the line adjacency graph representation of the text. The baseline is parameterized by two values; the base top and base bottom values, there values are set such that a certain percentage, of the black pixels in the text line is included between these two rows and the height of the baseline is minimized (Elgammal and Ismail (2001).

Through the determination of black pixel above the base line or under the base line and through some rules, you can define the overlapped words, dots, diacritics then connection parts can be determined.

6. Segmentation methods based on morphological operations

Morphological operations are used to understand the structure or form of an image. This usually means identifying objects or boundaries within an image. There are three primary morphological functions: erosion, dilation, and hit-or-miss.

In the morphological dilation and erosion operations, the state of any given pixel in the output image is determined by applying a rule to the corresponding pixel and its neighbors in the input image. The rule used to process the pixels defines the operation as dilation or erosion.

In Arabic handwriting, almost all the characters are connected by horizontal lines, so, applying Morphological operations such as erosion followed by dilation, segments the word into several segments. Each segment might represent a character, part of a character, or more than one character. Moreover, the beginning of the character preserves most of the significant information required to identify the character. Vertical or semi-vertical strokes that might represent the start, end, or a transition to another character (or sub word) are found by singularities. On the other hand, regularities contain the information required for connecting a character to the next character. Hence, these regularities are the candidates for segmentation (Motawa et al, 1997).

7. Recognition-based segmentation methods

These methods take the approach of splitting a word into segments to be recognized, using a segmentation algorithm. There have been various attempts to reliably segment characters but many have imperfections.

Humans can easily segment characters by first recognizing them. Whilst some methods try to combine segmentation with recognition, normally they are done as two separate stages. Therefore a computer has difficulty reliably segmenting characters. For example, many groups take the approach of over-segmenting a word and then using a separate stage to combine these fragments into whole characters. Whilst this two stage approach makes it easier for the computer to segment words, there is the possibility of error in both stages. If two segments which shouldn't be combined happen to look like two parts of one character, it is likely a segmentation error will result.

- **Feature Extraction**

During or after the segmentation procedure the feature set, which is used in the training and recognition stage, is extracted. Feature sets play one of the most important roles in a recognition system. A good feature set should represent characteristic of a class that

helps distinguish it from other classes, while remaining invariant to characteristic differences within the class.

By Nafiz Arica (Arica, 1998), a set of characteristics found in literature classified as shown in the table (2-2).

Table (2-2): Characteristics of feature extraction types

Type	Example of the type
Global Transformation and Series Expansion Features	• Fourier Transforms • Walsh=Hadamard Transform • Rapid transform • Hough Transform
Statistical Features	• Zoning • Characteristic Loci • Crossing and Distances
Geometrical and Topological Features	• Strokes • Stroke Directions and Bays • Chain-Codes

- **Training and Recognition Techniques**

There are many techniques used in automatic character recognition such as:

- Statistical Techniques.
- Syntactic Techniques.
- Neural Networks.

These techniques use either holistic or analytic strategies for the training and recognition stages:

- Holistic strategy employs top-down approaches for recognizing the full word, eliminating the segmentation problem due to the complexity introduced by the whole cursive word, but disadvantage is the recognition accuracy is decreased. The characteristics for this type are whole word recognition, limited vocabulary and no segmentation.

- Analytic strategies employ bottom-up approaches starting from stroke or character level and going towards producing a meaningful text. The characteristics for this type are sub word or letter recognition, unlimited vocabulary and requires explicit or implicit segmentation.

Statistical Techniques is concerned with statistical decision functions and a set of optimality criteria (Arica, 1998).

The major statistical approaches, which are applied in the character recognition field, are the Non-parametric Recognition, Parametric Recognition, Clustering Analysis, Hidden Markov Modeling, and Fuzzy Set Reasoning.

In Syntactic Techniques, the patterns are used to describe and classify the characters in Automatic Character Recognition systems (ACR). The characters are represented as the union of the structural primitives.

Syntactic Pattern Recognition methods are applied to the ACR problems such as Grammatical Methods, Graphical Methods, and matching method.

But a Neural Network is defined as a computing architecture that consists of massively parallel interconnection of simple "neural" processors, because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques, because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal, it is used in pattern recognition by defining nonlinear regions in the feature space. A neural network contains many nodes. The output from one node is fed to another one in the network and the final decision depends on the complex interaction of all nodes (Arica, 1998).

2.2 Literature Review

There are many researchers worked in this field; Hamami and Berkani (Hamami and Berkani, 2007) defined the segmentation as "based on the observation of the histograms of the lines and columns and takes into account the particular characteristics of Arabic

writing" this method included three stages: Locating text lines, locating connected parts and separating the connected parts into characters.

Horizontal histograms are used to detecting the text lines, to locate line, detected spaces between lines must be detected, to locate lines spaces between lines are detected which are line pixels with a zero or minimal histogram. In horizontal segmentation, horizontal sweeping is used to determine the beginning of the text line corresponding to the first line of the binary matrix, which has at least one black pixel. Then, the end of the text line corresponding to the line of the binary matrix is determined, which has no black pixel. The horizontal sweeping is done from right to left.

To detect connected parts in the line text, vertical projection of a line text on a horizontal axis is applied, to determining the beginning and the end of the connected parts the vertical sweeping is done from top to bottom. Proceeding by a vertical sweeping is used to determine the beginning of the connected part corresponding to the first column of the binary matrix, which contains at least one black pixel. To determine the end of the connected part corresponding to the first column, those contain no black pixel.

Connected parts is contain more than one character, to extract the character from connected parts, many of steps must be determined such as the junction line, top and bottom line of each column of the connected parts, and vertical histogram for each column of the connected parts.
There are many criteria put for determine the beginning and the end character, also there are many conditions put on final column, after applied the algorithm to extract the character from connected parts, the result is shown as the figure (2-10).

.

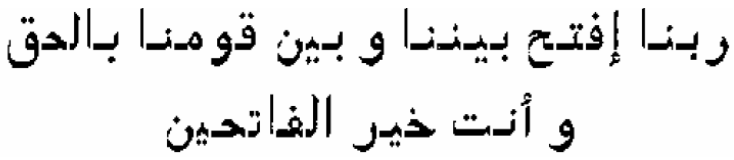


Figure (2-10): Segmentation in Characters

There are many criteria must be considered to character classification such as concavities and holes, to determine the concavities and holes, the principle proposed is clarify in the following :-

"After the character delimitation, a horizontal and vertical sweeping is performed on the whole surface containing the character. During this sweeping, the lines (respectively columns) coordinates are memorized, which represent four or more transitions, and at that time, the character is said to be crossed twice or more. Therefore, there is a possibility of having a hole or concavity" (Hamami and Berkani, 2007).

Secondary characteristics such as character form (square, elongated or upright), filling rate, Existence of punctuation, Position of punctuation, and number of points are used to recognize the character by compare all characters of the class from through the secondary characteristics for character.

This system can recognize multi-font of Arabic script. The advantages of this system are the problem of over-segmentation of some characters was solved such as character "س", The recognition stage uses a structural method, instead statistical method, therefore the system can be used for bilingual writing (Arabic/Latin) But the disadvantage in this system cannot solve Under-segmented problem such as □ □ □ □ □□□□ □ □□, and this character was considered as one character.

Optical character segmentation systems are used the baseline in segmentation process, but due to some of problems for using baseline segmentation such as combination of adjacent Arabic characters ,many of the Arabic characters have a middle region that is the same as the connection between two characters, and the distance between them is short ,therefore, placing the segmentation points is a difficult task, because of these problems, Mostafa G. Mostafa used T-junction to solving these problems, these method depend on "Arabic characters in a printed text start a T-junction at the baseline" (Mostafa,2004).

The character form for beginning the word may be start by T-junction such as "ب" , and may be end by T-junction such as "س", and some Arabic characters has more than one T-junction.

This approach based on more than one stage, preprocessing stage is used to help in determining location of T-junction by decomposing the image into text, noise removal

from image, thinning process, and finding the location of the baseline by vertical projection.

In scanning stage, character features is detect such as stroke, dots, loop, and cursive. Stroke recognition is based upon number from features such as "The stroke is always above the baseline, the stroke ends with an end point, the stroke does not contain any holes, the stroke width is always smaller than a threshold value, the stroke height is always smaller than a threshold value"(Mostafa, 2004). Those values are given by:

The width threshold value (WTV) = 1/4 W

The height threshold value (HTV) = 1/2 L

Where W is the average character width in the same line, and L is the average character height in the same line.

Analysis stage, extracted features are analysis to find the suitable location for character segmentation by rule "A character starts at a T-junction and ends before the next T-junction.", in this stage there is problem in the character that contains more than one T-junction such as ـشـ , ـسـ . Therefore, some rules are added for treating of such characters.

In the final stage the grouping stage the characters are segmented by grouping its components. Figure (2-11), illustrated Character segmentation occurs through three steps. The scanning step (a), the analysis step (b), and the grouping step (c).

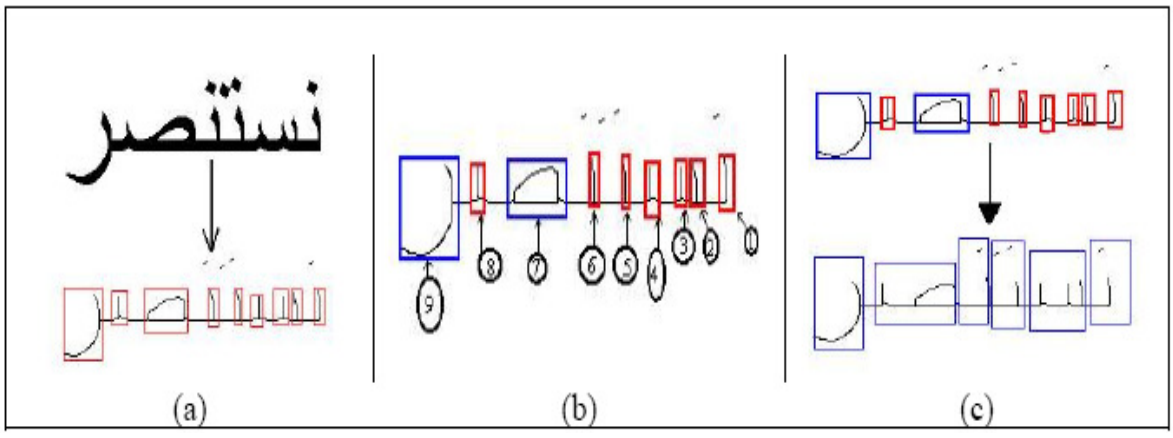


Figure (2-11): Character segmentation occurs through three steps.
(a) The scanning step, (b) the analysis step, (c) the grouping step.

In this technique, segmentation accuracy is 96.5% in segmenting a text printed with simple font, but some problems occur in the text thinning stage. These problems arise from some characters when written in some fonts and/or with small size such as " ع"and " م " characters.

Some of researchers proposed segmentation techniques based on the vertical histograms and some rules. Liying Zheng and others (Zheng et al, 2004) proposed such techniques, which is used structural characteristics between background regions and character components, and the characteristics of isolated Arabic characters, it used vertical histogram of each sub-word, and in addition to some rules such as use rules to check whether the sub-word includes only one character, scans the vertical histogram of the sub-word which may consist of more than one characters from right to left, also used the rules to decide whether the point is a real segmentation point, and it partition the sub-word at the segmentation points.

Segmentation process are divided into three levels, line segmentation (text into line), word and sub-word segmentation (line into word), and character segmentation (word into character).

Line segmentation method is used to segment the text image into lines image by horizontal histogram, through, draw the horizontal histogram of the text image, and consider:

- BaseLineflag = True.
- Then worked scan the horizontal histogram from top to bottom.
- If the histogram value changes from zero to nonzero and BaseLineflag is True, then the row is the start of a line and let BaseLineflag = False.
- If the value of the histogram change from nonzero to zero and the space is larger than a threshold, then the row is the end of the line and let BaseLineflag = True.
- For a word segmentation or sub-word; vertical histogram is used of each line scanned from right to left.

Character segmentation is divided into two categories, isolated character detection and character segmentation, isolated character does not need to be segmented, to check the character is isolated or not, some rules are use such as "the features extracted from background regions; zones that are covered by the sub-word; the ratio of height and width of the sub-word image; Number of vertical and horizontal cross points of the sub-word image" (Zheng et al, 2004).

Vertical histogram is used to character segmentation; after draw the vertical histogram is used rules to detect the character.

The correct segmentation rate is near from 94.8% for some character that its used; but drawback in over-segmentation rate that is happens, due to some of fonts (simplified Arabic and Arabic transparent) are alike; and no overlap between sub-words.

Omidyeganeh and others (Omidyeganeh et al ,2005), proposed technique, based on the conditional labeling of the up and down contour. Also uses up contour curvature and adaptive base line for each sub-word.

In the pre-processing stage, pen size is calculated (w) and a global base line is defined as a horizontal line, all across a text line whose width is equal to (w), and in this stage up and down contour is extracted by moving from right to black pixel clockwise through the contour, and from left down black pixel to right black pixel clockwise through contour.

The segmentation algorithm consists of four stage are contour labeling, contour curvature grouping, character segmentation and post preprocessing.

Contour labeling is based on the conditional labeling of the up and down contour of each sub-word, which is used 1,0 and -1 standing for up, middle and down of the base line, to labeling process used state diagram for determine up and down contour.

To improve the segmentation results, contour curvature is used, up and down contour of the sub-word traced counter clockwise are represented.

In character segmentation stage, a probable segmentation point is defined as "All segmentation points must be 1.5w apart from left and 2w apart from right end of the sub-word, and all segmentation points must be around base line segmentation within w/2 of it " where w is the most frequent pen thickness in columns that adopted as the pen size, (Omidyeganeh et al ,2005).

The advantages in this paper are applied on multi-font (18 several font) , the segmentation rate of about 97% , but there is found the segmentation errors were mainly due to skewed text lines.

Mansoor Al-A'ali and Jamil Ahmad (Al-A'ali ,2007) are used the horizontal projection to split the text in the image matrix into lines of text and vertical projection technique into segments.

By vertical and horizontal projection on the text, can segment the word into sub-word or character, by scanning the columns, if found density of pixel equal zero, it is indicates the beginning of a segment, then searching another density of pixel equal zero, if existed, that is indicate the end of a segment, figure (2-12) illustrate to word segmentation into its sub-words.

Figure (2-12): Word segmentation into sub-words

After determine the bottom left point and top right point of the enclosing rectangle, horizontal, vertical and intersection markings are done by using middle of the distance method for each segment, to determine the stroke in the segment, horizontal and vertical intersection of location of the segment is used, there is a relationship between cursor sized and number of pixel which is employee to determine the stroke, these relation is:

$$\text{Number of pixel} = ((\text{size} * 2) + 1)^{\wedge}2$$

Where size is the size of the cursor (0, 1, 2), the movement of the cursor in an enclosing box is governed by the group rules.

Features of a segment are defined as a group of strokes and each stroke is represented by "A starting point, a string of directions, an ending point and an array of cursor sizes that correspond to each direction in the direction string" (Al-A'ali ,2007).

To character classification a knowledge base is built to contain information about character classes. Such as class for closed loop, class for all characters that have a half open circle and so on, then the strokes assemble to form valid characters, and assign a suitable character class, then chose related ASCII value.

The character classification module was not implemented due to problems in the feature extraction module such as types of strokes extracted varied as two constants were varied.

There are some disadvantages in this paper such as assumed that there is no horizontal or vertical overlapping between characters in the text, and there should also be no slanting of text and the quality of text should be good.

Mohamed Fakir and M. M. Hassani (Fakir, 2000) are divided the segmentation stage into three levels: line, word, and character segmentation.

The line segmentation based on the analysis of the horizontal projection profile of the text, which search on the black pixel on the full row, if no black pixel, this row is used to separate between lines.

Vertical scan is used for word segmentation, when scan columns, if not found black pixel in the column, denoted the beginning word segmentation, and search on column subsequently that not found black pixel, to denote the end of word segmentation.

Character segmentation involves the building of vertical projection profile of the middle zone of the word, which used a fixed threshold for segmentation a word into character; figure (2-13) illustrate the result of segmentation.

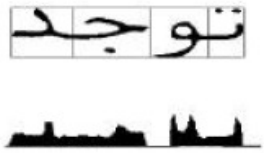


Figure (2-13): Result of segmentation, (this word is segmented into four characters)

In this paper used Hough transform (HT) to feature extraction mechanism, which is a linear transform originally developed for line detection in digital pictures.

The classification process consists of two steps (Fakir, 2000):

1. The character main body is classified using features selected in the HT space and dynamic programming (DP) matching technique.
2. Simple topological features extracted from the geometry of the secondary parts are used by the topological classifier to recognize the character completely.

The recognition rate was about 95% but there are some of errors such as substitution errors are usually occur because of thinning problems, and rejection errors are usually caused by bad printing.

This method overcomes the problem of noise sensitivity in the local approach, and also the problem of time being consumed in the global approach. That used in feature extraction.

Conclusion

There are many obstacles facing researchers working on Arabic character recognition mainly in segmentation process such as overlapping and over-segmentation, these obstacles solved by developing color algorithm which used to recognize connection parts, and Developing an algorithm to segment the line into connection parts depending on color algorithm where each connection part will be moved respectively from other connection parts, then Segmentation of the connection parts into characters.

CHAPTER THREE
THE PROPOSED SYSTEM

Chapter Three
The Proposed System

3.1 Introduction

The aim of the thesis is to develop a system that is able to recognize a typewritten Artistic Arabic script. The purpose is to produce a system, which recognizes a given input as a bitmap format picture of Arabic text. The picture will be segmented into lines, words and characters. After segmentation process, the segmented character is compared with the ones which we have considered to be an identical character in the library. So the segmented character is recognized as one of these characters. This process is repeated for all segmented characters in the picture. As a result of the mentioned processes the picture of the Arabic text will be converted to a text.

This chapter presents a discussion of research methods that help to reach a satisfactory outcome to recognize printed Arabic letters. And because the presence of a wide varieties of Arabic fonts with different sizes and specifications, the thulth font with specifications; size 36, bold, resolution of 81x81 dots per inch is selected in this thesis for scientific research purposes.

In an overview to proposed system it is noticeable that it is composed of the following basic components:

- Image Acquisition.
- Preprocessing.
- Segmentation.
- Classification and recognition.

Figure (3-1) illustrates the components of the proposed system.

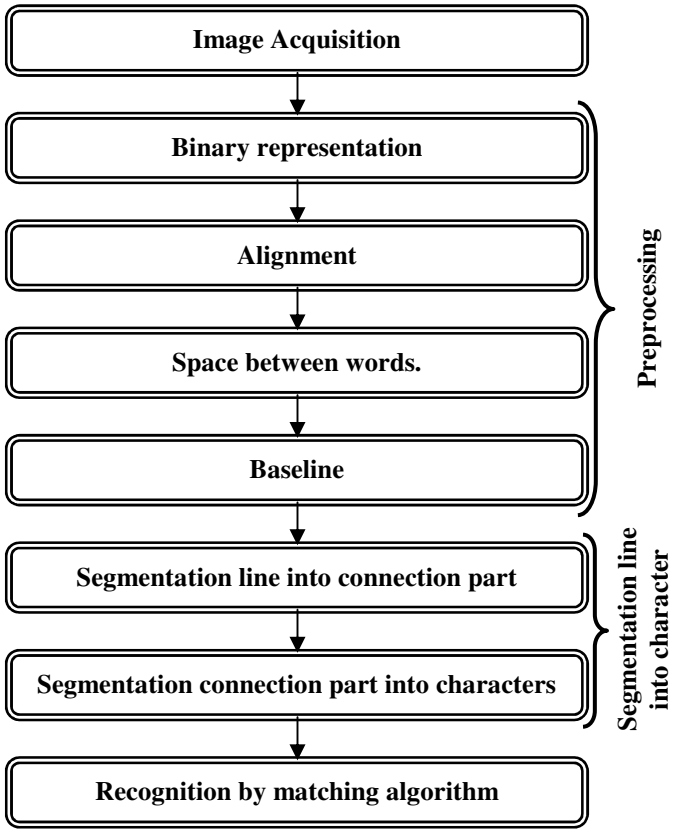


Figure (3-1): Proposed system components

3.2 Image Acquisition

The image of Thulth font printed characters, with a size of 36 and resolution of 81x81 dots per inch was taken from one of the paint applications, such as paint or Photoshop and others, the image was stored in the form of bitmap format (bmp) because this style is support up to 16777216 colors and it was not compressed, not allowed any loss of data.

3.3 Preprocessing

The aim of preprocessing the image is to improve the quality of the images of Arabic characters, and this will help us to segment and recognize Arabic words accurately.

There are many techniques used in the preprocessing of the printed and handwritten text image, such as the following:

- Binary representation
- Smoothing
- Alignment
- Determination of the space between words
- Baseline

Some of these techniques will be discussed here and others will be used during segmentation.

3.3.1 Binary representation

In general, the outputs of Image Acquisition devices such as television and camera are analog signals, accordingly to treat these images by computer, these images must be converted to a proper form to be suitable for preprocessing, this form is called digital signal.

The image of the characters is a binary color in nature, so image foreground contains patterns from a single group of similar characteristics, it is natural that all are one color, for this the general method to recognize Arabic characters is representing the image through binary system, where the binary image characterized by a small size and can be processed faster and easier than other kinds of images.

Binary image is that image which contains two gray levels, "1" for the white pixels which form the image background and "0" for the black pixels which form the image foreground.

Binary image for characters is used in this research in a form of binary matrix because of the easiness of it to be programmed. Where the matrix consist of rows and columns, the intersection of row and column form a pixel, which is the smallest part in the image.

The number of pixels in the image significantly affects the image, whereas, number of pixels increases, image contrast and image size increases too.

In the binary representation algorithm the color of the pixel obtained which is represented by a number, the three main colors red, blue and green extracted, after that the average of these numbers calculated. If the average is greater than the specified threshold , the color of the pixel will be white else it will be black, this operation will be repeated for all the pixels in the image, algorithm (1) in figure (3-2) will illustrates binary representation.

Algorithm Name: - Binary Representation
Input: - Original picture saved as bitmap format
Output: - Binarised picture
Step -1:- Find scale width for original picture
Step -2:- Find scale height for original picture
Step -3:- For each column i = 0 to scale width do
For each row j = 0 to scale height do
Step -3.1:- Read point (j, i)
Step -3.2:- Find color number for red, green and blue colors of the point
Step -3.3:- Calculate average of color numbers for colors in step 3.2
Step -3.4:- If average greater than specify threshold then set point (j, i)
1 Else Set Point (j, i) 0.
Step -4:- End

Figure (3-2) algorithm (1): Binary representation

3.3.2 Alignment

Alignment stage is considered one of the preprocessing stages of the image. It is implemented before segmentation and recognition of the image, and will lead to the alignment of the text lines to the upper left corner of the image area by deleting all empty upper and lower rows, in addition of the deleting of the empty side columns. In another word deleting all rows and columns not contain any black pixel, the algorithm can be explained by figure (3-3).

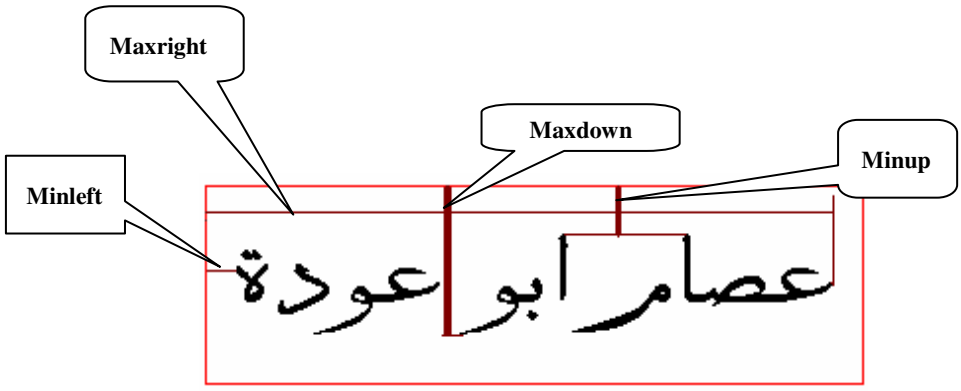


Figure (3-3) : Original picture with alignment parameters.

Where:

Maxright: is the longest horizontal distance carrying at the end a black pixel.

Minleft: is the shortest horizontal distance carrying at the beginning a black pixel.

Maxdown: is the longest vertical distance carrying at the end a black pixel.

Minup: is the shortest vertical distance carrying at the beginning a black pixel.

Figure (3-4) algorithm (2) illustrates the alignment.

Algorithm Name: - Alignment.

Input: - Binarized picture.

Output: - Aligned picture (move binarised text picture into upper left corner).

Maxright: is the longest horizontal distance carrying at the end a black pixel.

Minleft: is the shortest horizontal distance carrying at the beginning a black pixel.

Maxdown: is the longest vertical distance carrying at the end a black pixel.

Minup: is the shortest vertical distance carrying at the beginning a black pixel.

Step -1:- Calculate scale width and scale height for binarisation picture as the following equation:-

Scale width = Maxright – Minleft.

Scale height = Maxdown – Minup.

continue

Step -2:- Create empty picture (alignment picture) according to the calculated height and width in step 1.
Step -3:- For each column x = 0 to scale width
L=x + minleft
For each row y = 0 to scale height
Step -3.1:-Read the point (x, y) from binarisation picture according to the equation: P = Point(L , y + minup).
Step -3.2:- Copy the point (P) in alignment picture in point (x, y).
Step -4:- End

Figure (3-4) algorithm (2) : Alignment.

3.3.3 Determination of the space between words

This thesis deals with one line, where the line is segmented into connection parts, then the connection parts segmented into characters.

Before segmentation and recognition all the empty spaces between words should be determined to be able to segment each word alone, then each word can be segmented into characters.

To determine spaces between words, the first step is to locate the pixels that have a white color along the column, the number of this column considered as a start point of the space separating two words, then locate another column that contains white pixels along the column and define its number, if the absolute difference between these two column's numbers is greater than specified values (N1) then this distance considered a space separating two words, but if the absolute difference is not between a specified values, then another column that matches the criteria should be located. Algorithm (3) in figure (3-5) illustrates determination of the space between words.

Algorithm Name: - : Determination of the space between words.
Input: - Aligned picture.
Output: - A picture with spaces between words.
Old x: - Value of x when no move of x
New x: - Value of x when move of x
M: - Scale width for aligned picture

continue

```
N: - Scale height for aligned picture
New color: - Color of pixel.
N1:- value is (9)
Old x = New x = -1
Step -1:- For each column x = 0 to M do
          New color = 1
        Step -1.1:- For each row y = 0 to N do
                    If point (x, y) = black then New color = black
                    Next y
        Step -1.2:- If New color = 1 then
        Step -1.3:- If Old x = -1 then Old x = x
                    New x = x
                    Else
                If old x>-1 and new x > - 1 then
                    If absolute (Old x – New x) > N1 Then
                    draw symbol to refer tospace
                    End if
                End if
                    Old x = New x = -1
       End if
Step -2:- Next x
Step -3:- End
```

Figure (3-5) algorithm (3) : Determination of the space between words

3.3.4 Determination of the baseline

Baseline is the imaginary line, to write the letters along this line, where some characters or part of these characters can be written upper the baseline called ascender, or under the baseline then called descender.

Determination of the baseline often necessary step precedes recognition, to extract features and to facilitate the segmentation process.

Projection histogram method is used in this thesis, which transacts with horizontal lines of the image. The idea of the baseline is the line that have the largest total of black

pixels. The black pixels gathered along each row, the row that have the largest total of black pixels determined as baseline.

Due to stretching of the majority of black pixels along the baseline, projection histogram method is better for long words which considered as an advantage of the method, while the disadvantage of this method that it gives wrong suggestions for the short isolated words. Figure (3-6) shows wrong baseline as in the word "راج" and right baseline as in the sentence "راج العلم في البلاد".

A

B

Figure (3-6) : (A) Wrong baseline (B) Right baseline

Figure (3-7) algorithm (4) illustrates determination of the baseline.

Algorithm Name: - : Determination of the baseline.
Input: - Aligned picture.
Output: - Picture with determined baseline.
No_of_bline: row number that is considered as the baseline.
M: - scale height for aligned picture.
N: - scale width for aligned picture.
counter: - number of black pixel in the row.
old counter: used to save of the largest number of pixels in the row. ..
Bline (No_of_bline): to determine the baseline row.
Step -1:- No_of_bline= 0
Step -2:- For each row y = 0 to M
Counter = 0 continue

```
        Step -2.1:- For each column x = 0 to N
                    If point (x, y) = black then counter = counter +1
                    Next x
         Step -2.2:- If counter > old counter then
                     Old counter = counter
                     Bline (No_of_bline) = y
                     End if
        Step -2.4:- Next y
Step -3:- Recognize the base line by color differ for writing color
Step -4:- End
```

Figure (3-7) algorithm (4) : Determination of the baseline

3.4 Segmentation

Character recognition is mainly depending on character segmentation, as the accuracy of segmentation increases the accuracy of character recognition increases.

Word segmentation is the main problem in character recognition, there is disagreement between researchers on segmentation methods to achieve the most possible accuracy of recognition

Segmentation method used in this thesis consist of three stages, the first stage is coloring text, the second stage is segmentation of the colored text to connection parts, and the third stage is segmentation of the connections parts into characters.

Segmentation of the page passes through the following stages:

- Determination of text lines locations (text lines location stage).
- Segmentation of text lines into connection parts (Connection parts location stage).
- Segmentation of connection parts into characters (Segmentation stage).

3.4.1 Text lines location stage

This stage includes the determination of lines of the text, that means segmentation of the text page into lines, then determination of the location of each line in the text page, through calculating the space between two lines and comparison of the calculated space

with a specified value, if the calculated value is greater than or equal the specified value then the next line considered as separate line.

Latifa Hamami and Daoud Berkani (Hamami and Berkani, 2007) used in their research titled "Recognition System For Printed Multi-Font And Multi-Size Arabic Characters" horizontal segmentation method to segment the text page into lines, this method will not be discussed in my thesis because it was discussed in many researches.

3.4.2 Connection parts location stage

Determination of connection parts widely helps in segmentation word into characters. Many researchers used vertical segmentation method in the process of determining connection parts, through the procedure "vertical projection of a line text on a horizontal axis, the obtained histogram will have some zero value columns, these columns are used to delimit the connected parts; it consists of determining the beginning and the end of the connected parts, each text line obtained in the horizontal segmentation is sub-divided into connected parts" (Hamami and Berkani, 2007).

This method (vertical segmentation) is suitable for fonts which have no overlap or ligatures, where two problems emerged under segmentation and over segmentation, such as the word "محمد" cannot be segmented in some fonts like thulth, as will as the letter "س" segmented into three parts.

Thulth as artistic Arabic font faces in recognition many problems, ligatures and overlapping are the most important problems and need big efforts to solve due to the complexity of these problems, this thesis is a try to solve these problems.

Segmentation of the line into connection parts passes through two stages:

- ► Coloring line according to connection parts with different colors (coloring algorithm).
- ► Segmentation of the colored line into connection parts according colors (connection part algorithm)

► Coloring algorithm

At this stage each connection part is colored by a specified color differ than the other connection parts from the start to the end of the text line, it must be notable that in this

stage black color is not used because the color of the original text will be black after the binary representation as referred to it in section 3.3.1. Figure (3-8) algorithm (5) illustrates this stage.

In this algorithm, the height and width of the image needed to be segmented into connection parts should be determined, then the treatment of the image performed pixel by pixel, if the value of the pixel equal zero (black pixel) another color rather than black will be selected to color this pixel and the attached black pixels, in the next black pixel a different color will be selected rather than the black and previously used colors to color it, this procedure repeated until all pixels in the image are treated, using this method will give us a colored connection parts.

► Connection part algorithm

The colored text segmented into connection parts according to colors which determined in figure (3-8) algorithm (5), where each color refers to a specified connection part.

Name: - Text line coloring.
Input: - Picture after determination of the space between words..
Output: - Colored text line picture.
M:- scale width for colored picture
N:- scale height for colored picture
Step -1:- Save the space picture to the empty picture which called colored picture
Step -2:- For each column x = 0 to M
Step -2.1:- For each row y = 0 to N
Step -2.1.1:- Read point (x,y) and save the point in p variable.
Step -2.1.2:- If p = black then
- Selecting fill color, not black color
- Fill black pixel and connected black pixels by selected color (fill color)

Step -2.2:- Next y
Step -3:- Next x
Step -4:- End

Figure (3-8) algorithm (5) : Text line coloring

Determination of connection parts passes through the following stages:

- ► Determination colors of colored picture stage.
- ► Determination of the connection parts without dots stage.
- ► Smoothing and moving dots into connection parts stage.

► Determination colors of colored picture stage:

This stage includes reading the pixels presented in the colored picture according colored picture width and the baseline determined in algorithm (4) section 3.3.4. if the color neither white nor black and the color is new, then the new color will be saved in an empty location in the matrix, else the process will be repeated for a new pixel. Figure (3-9) algorithm (6) illustrates this stage.

```
Algorithm Name: - Text line color determination in the colored picture.
Input: - coloring picture.
Output: - matrix contains all color for text line coloring picture.
M: - scale width for colored picture.
y: - is base line.
Color: - is point color.
Ma: - is matrix used to save colors for text line coloring picture  .
c: - is location number in matrix Ma.
D: - is a value does not represent a color such as –No .
Step -1:- For x = 0 to M
        Step -1.1:- Color = point (x, y)
        Step -1.2:- If Color is black or white then goto step 3
Step -2:- if ma (c) = color then goto step 3
            If Ma (c) = D   then ma (c) = color: goto step 3
            c= c +1: goto step 2
Step -3:- next x
Step -4:- End
```

Figure (3-9) algorithm (6) : Text line color determination in the colored picture

From algorithm (6), we notice that it saves all colors in the text present in colored picture rather than white and black in a matrix to be used later.

► Determination of the connection parts without dots stage:

This stage includes extraction of connection parts from the text according to the color read, and saved in a new picture differ than the colored picture, after saving in the new picture it is erased from the colored picture by coloring it with white color, it should be notable that dots are not treated in this stage, before explaining the algorithm an example will be shown in figure (3-10) to clarify this stage.

In this example the sentence "فيلادلفيا" will be transmitted to its own connection parts, with a remark that colors existed in the image have been saved in a matrix as shown previously in algorithm (6).

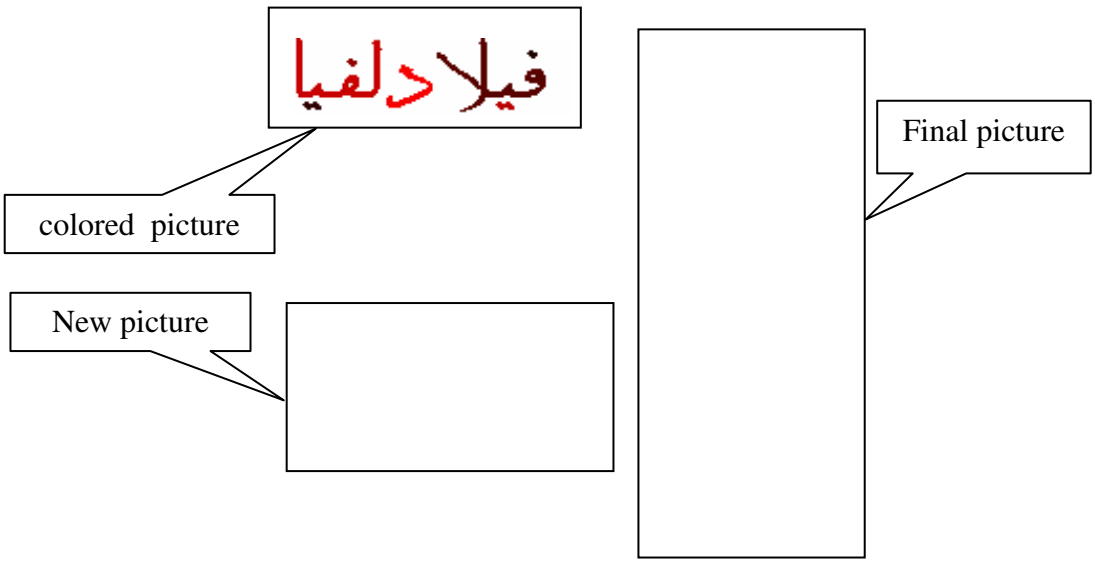


Figure (3-10): Determination of the connection parts without dots.

The program start form the point (x,y) with suggested zero value, where its value indicates the color of the pixel, if the color of the pixel neither white nor black this pixel will be put in the same location of the new picture with a black color, and saves the coordinates of the start and the end of the part, the procedure continuous until finishing all the pixel that have the same color, it is remarkable that if the number of black pixels of the part processed is less than a specific value and not exist on the baseline then it will be considered as a dot, in this case this dot will be erased from the new picture and will remain in the colored picture. The result can be shown in figure (3-11).

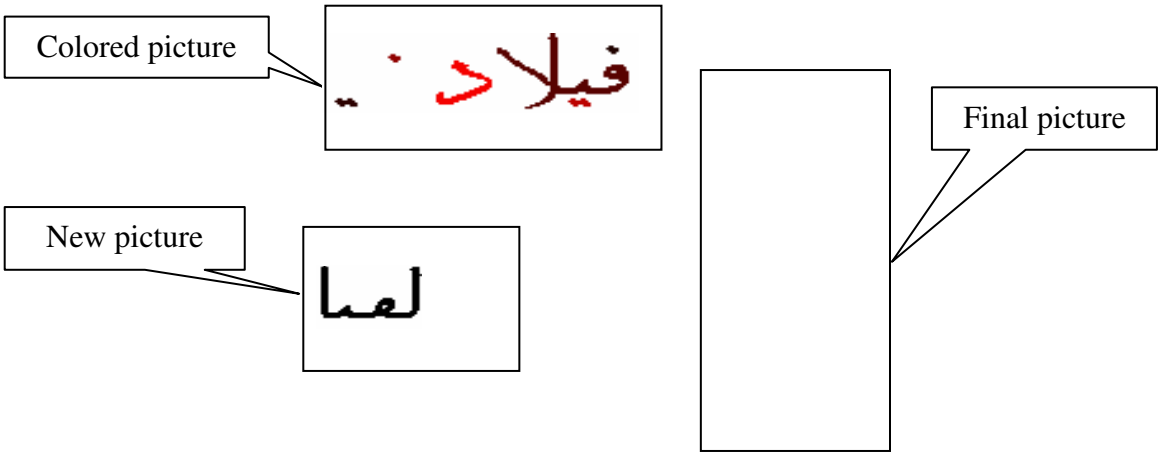


Figure (3-11): Transmission of the connection part into new picture without dots

When there is a different color, the program will transmit the part exist in the new picture to its location in the final picture and deals with it as a dependant area has a start point and end point, and erase this part of picture from the new picture to get another part of another color with a new coordinates, the result can be shown in figure (3-12).

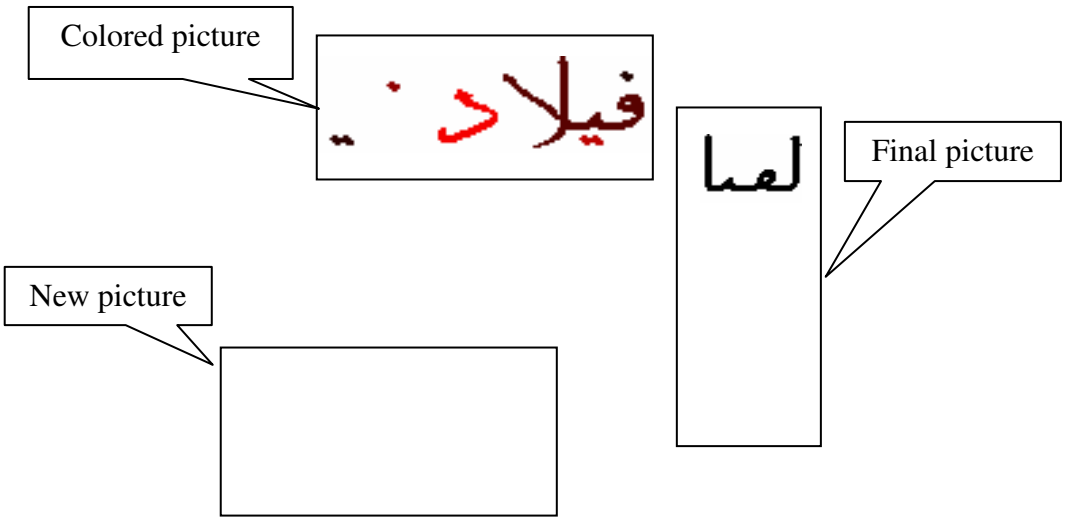


Figure (3-12): Transmission of the connection part from new picture into the final picture

The process continues until finalizing all connection parts; figure (3-13) shows the process. And figure (3-14) Algorithm (7) illustrates determination of the connection parts without dots.

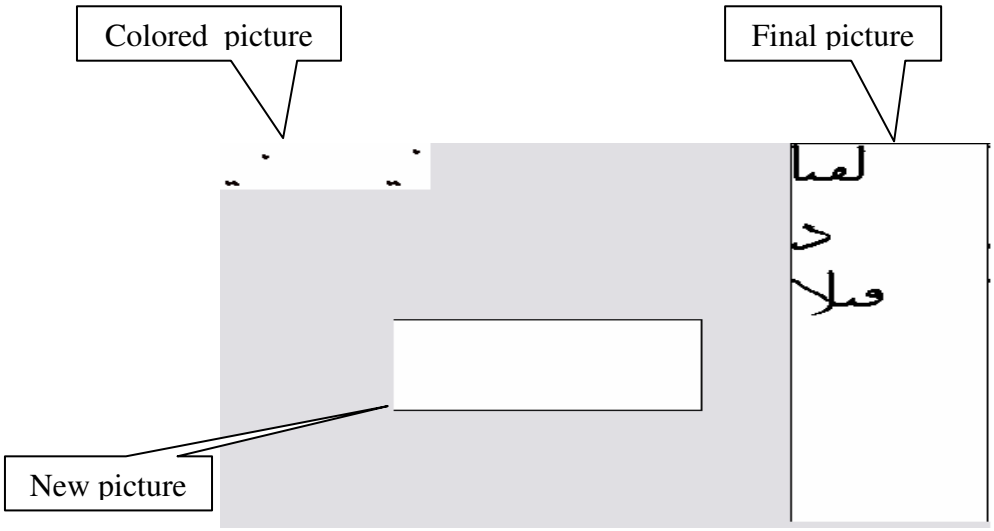


Figure (3-13): transmission of all connection parts from colored picture into final picture without dots

Algorithm Name: - determine the connection parts without dots

Input: - coloring picture.

Output: Determination of the connection part without dots

X,Y:- are pixels in text line coloring picture, where the beginning X,Y is zero

D: - is a value does not represent color such as -10.

x0,y0 :- is start point for rectangle form which is called new picture

x1,y1 :- is end point for rectangle form which is called new picture

dt :- is number of pixel for dot

n: - the largest number of pixel can represent the dot (45 pixels)

np number of connection part

ybase :- 0 to deal with dots , 1 to deal with characters.

Step -1:- x-loop

Step -2:- y-loop

Step -2.1:- new color = point (X,Y)

Step -2.1.1:- if new color is black or white then goto step 3.

Step -2.1.2:- if old color= D then

save start point coordinates for x0=x1=X.y0=y1=Y

old color – new color **continue**

Replace X, Y with white color in the coloring picture
Replace X, Y with black color in new picture
Else
If old color = new color then
If Y is exist on base line then ybase = 1
- Save start point coordinates for x0,y0 and end point coordinates for x1,y1
- replace X,Y with white color in the coloring picture
- Replace X,Y with black color in new picture

End if
Step -3:- Y= Y+1
Step -4:- If Y < scale height for coloring picture then goto step 2
Step -5:- If old color is used then
Step -5.1:- If dt < n and ybase = 0 then
Erase new picture
Else
Put point(X, Y) in the coloring picture white color
Step -5.2 -copy connection part from new picture to final picture
erase new picture
np =np +1
save x0,x1,y0,y1 in matrix
End if
End if
Step -6:- x= x+1
If x < scale width for coloring picture then goto step 1
Step -7:- End

Figure (3-14) algorithm (7): Determination of the connection part without dots

► Smoothing and moving dots into connection parts stage:

This stage includes decreasing noise around dots according specific criteria to help in word and connection part segmentation properly, and then transmitting these dots to its

suitable connection parts, figure (3-15) algorithm (8) illustrates movement of dots into suitable connection part.

Algorithm Name: - Movement of dots into suitable connection part
Input: - coloring picture.
Output: Movement of dots into suitable connection part
N: - scale width for colored picture
M: - scale height for colored picture
Op :- number of probable dots
Step -1:- For x = 0 to N
Step -1.1:- For y = 0 to M
- If the point (x,y) not exist on baseline and its size less than specific number of pixel (45 pixel), the point is dot then
- Smoothing the dots for specific criteria

End if
Step -1.2:-Next y
Step -1.3:- Next x
Step -2:- For I = 0 to N
Step -2.1:-For j = 0 to op
- If loc (0,j) = I then
- Copy the dot from colored picture to final picture in suitable location
- Remove the dot from colored picture

End if
Step -2.2:- Next j
Step -2.3:- Next I
Step -3:- End

Figure (3-15) algorithm (8): Movement of dots into suitable connection part

Segmentation of the text line into connection parts process includes transmitting the colored text from the colored picture to its connection parts in the final picture, taking in consideration that the space between words considered as a connection part; figure (3-16) illustrates segmentation of the text line into connection parts.

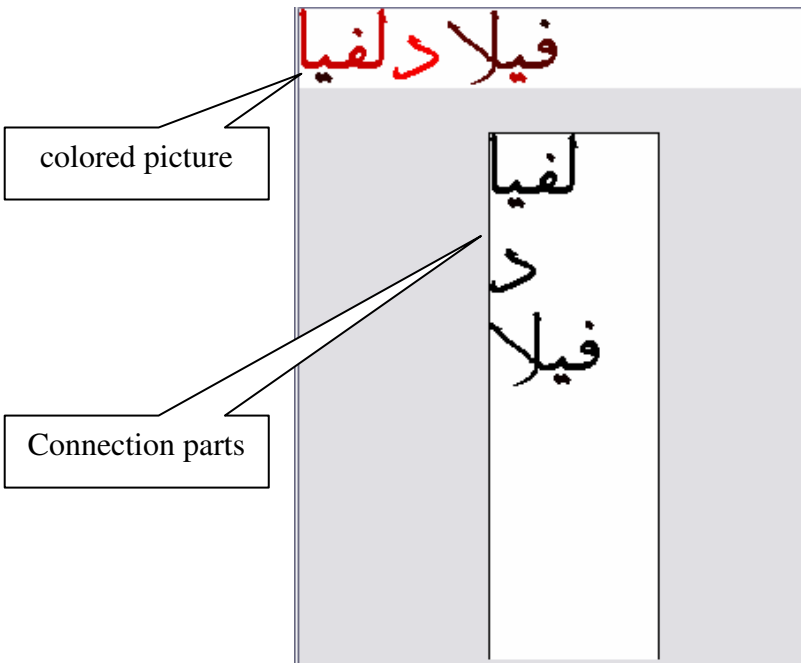


Figure (3-16) : Segmentation of the text line into connection parts

At the end of this stage, we notice that the overlapping problem was solved through the segmentation of words into connection parts.

3.4.3 Segmentation stage

In segmentation stage connection parts are segmented into characters, where two kinds of segmentation can be distinguish, isolated characters or cursive character. There are two conditions to perform segmentation into characters, the first is that the pixel color should be black and the second condition is that the pixel should be located on the baseline, where the baseline considered as three vertical points (upper=baseline, middle=baseline+1, lower=baseline+2), Figure (3-17) illustrates constructing of the baseline.

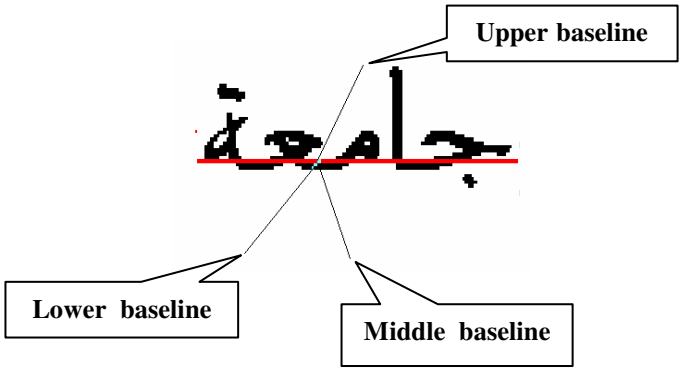


Figure (3-17): Construction of the baseline

Segmentation means segmenting each connection part into its characters, figure (3-18) illustrates segmentation.

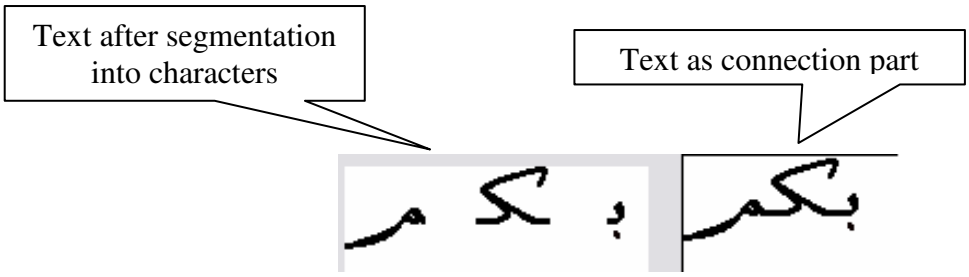


Figure (3-18): Segmentation of the connection part into characters

For segmentation algorithm is used as illustrated in figure (3-19) algorithm (9) Segmentation of connection part into characters.

Algorithm Name: - Segmentation of connection part into characters.
Input: - connection parts of text line in final picture (connection part)
Output: segmented characters of the connection part.
Np:- is number of connection parts
R :- pixel number where segmentation will be performed
M:- scale width of the final picture (connection part)
N:- scale height of the final picture (connection part)
D := pixel is out of the picture range such as -No.

```
Step -1:- For h = 0 to Np
        R = 0
        Step -1.1 :- For x = 0 to M
                Sum = 0
                Step -1.1.1:- For y = 0 to N
                        New color  = point (x,y+(h*M)
                        If new color = black then
                            If y is exist on the base line then
                            Sum = sum+1
                            Else
                            Sum =sum + value greater than or equal 10
                            End if
                        End if
```

continue

Next y

Step -1.1.2 If sum = 0 then new sum = 0

If sum > 0 and sum <10 then new sum = 1 else new sum = 2

If st =D then st=x: old sum = new sum: goto step 1.2

If old sum = new sum then goto step 1.2

R= R+1

If R < specific threshold then goto step 1.2

R=0

Save coordinates for start point of connection part segmentation

Step -1.2:- Next x

If st = D then goto step 2

Save coordinates for end point of connection part segmentation

Step -2:- Next h

Step -3:- Copy the characters from final picture to segmentation picture according to the beginning and the end of segmentation coordinate

Step -4:- End

Figure (3-19) algorithm (9): Segmentation of connection part into characters

There are some characters connected vertically called ligatures such as: لـا ، لـأ ، لـإ ، مـحـ ، حـ ، لآ ، these ligatures did not segmented by the system but considered as a one character image.

3.5 Matching

Two main methods used to recognize characters in OCR, Matrix Matching and Feature Extraction, where matrix matching is the easier and most common used , if the font type and font size are fixed.

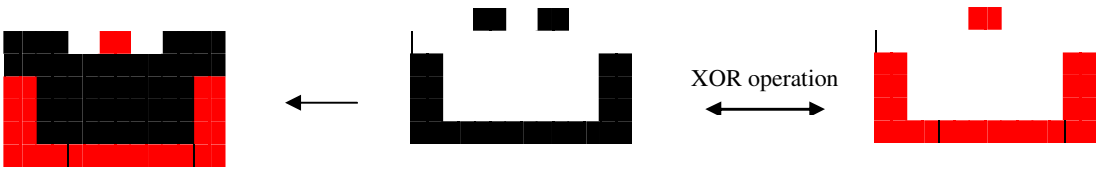
Matrix Matching compare the scanned picture of the character with the matrix library or pattern characters, if the scanned picture of the character match one of the pictures in the matrix library, the program consider that picture as identical ASCII character.

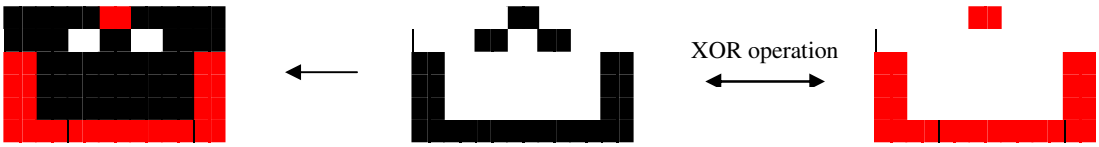
The following example illustrates the character recognition, suppose the segmented character "ن" which is represented in figure (3-20 a), and we need to make matching for this character with all characters " ث ، ت ، ن" represented in figure (3-20 b), the first step is inserting the segmented character in an empty matrix consist of 7x6 pixels -for example- using And operation, where number 1 refers to white color, while number 0 refers to black color which is converted to red color as shown in (3-20 c), then a matching using XOR operation performed between the red segmented character with the character in (3-20 b) as shown in figure (3-20 d)

A

B

C





D

Figure (3-20) : Example discusses recognition characters by matching matrix

(a) Segmented character (b) Matching characters

(c) Segmented character with empty matrix using AND operation

(d) Matching segmented character with matching characters using XOR operation

The program counts the pixels which are not black or red after matching process, the results was for the matching of the segmented character "ن" with the matching characters "ن", the sum of pixels equal zero, while with "ت، ث" the results were two pixels for both characters, the smaller sum of pixels means that the character is the matching character for the segmented character, and in our example the character "ن" is the identical character, figure (3-21) algorithm (10) illustrates Recognition of characters using matrix matching.

Algorithm Name: - Recognition of characters using matrix matching

Input: - Segmented characters picture

Output: character recognition for segmented character

M: - Scale width for picture recognition

N: - Scale height for picture recognition

Step -1:- Select the segmented character to match

Step -2:- Insert the segmented character into empty picture called "mat" using AND operation and convert the black color into red color.

Step -3:-Match the segmented character with matching characters using XOR operation.

Step -3.1:- For x = 0 to M

Step -3.2:- For y = 0 to N

If point(x, y) in picture recognition black or red then count = count Else count = count +1

Next y

Next x

Step -4:- If count is lowest value, then the character has the result is the identical character

Step -5:- End

Figure (3-21) algorithm (10): Recognition of characters using matrix matching

Through the stages mention previously all stages of character recognition explained for artistic Arabic characters starting from Image acquisition, preprocessing, segmentation and recognition.

CHAPTER FOUR
IMPLEMENTATION

CHAPTER FOUR
IMPLEMENTATION

4.1 Introduction

A full implementation is done for the Arabic OCR system described previously and provided a friendly graphical window-based user-interface. The developed system has been implemented on visual basic 6 language, the implementation and its results discussed in this chapter.

In chapter three, the developed system consists of image Acquisition stage, preprocessing stage, segmentation stage and recognition stage. The implementation of these stages will be discussed in this chapter.

4.2 Image Acquisition stage implementation

An image has been taken using one of the drawing software, where a thulth of a size 36, bold with a resolution 81x81 dot per inch. The image stored as a bitmap format (bmp). Where, the picture is represented as rectangular array of pixels. It is stored completely as digitally encoded images. Figure (4-1) shows some of the images used in this system.

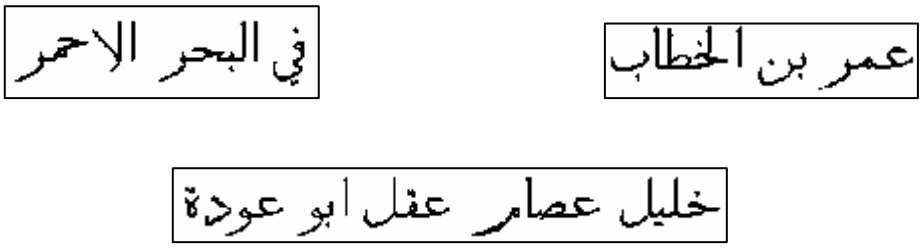


Figure (4-1): Examples of images used in the system

4.3 Preprocessing stage implementation

Preprocessing is an important stage because it affects all the next stages, such as segmentation and recognition, preprocessing implementation passes through the following steps:

► Binarisation step implementation

In this step matrix pixels are converted into two gray scale black or white, the color consists mainly from three basic colors red, blue and green, each pixel color represented

by a number, mathematically this pixel number is analyzed to its basic colors and the average is calculated, the color black or white is determined after a comparison of the average with a specified threshold.

The figure (4-2) illustrates the implementation result of algorithm (1) binarisation mentioned in section 3.3.1

A

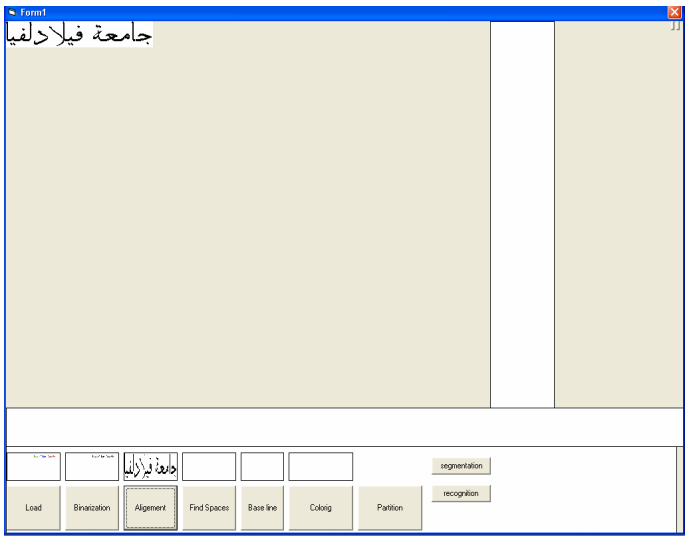
B

Figure (4-2): Binarisation step implementation (A) original image (B) binarised image

► Alignment step implementation

The text image is moved to the upper left corner, where all upper and lower empty rows and side empty columns are erased; figure (4-3) illustrates the implementation result of algorithm (2) alignment as mentioned in section 3.3.2.

A



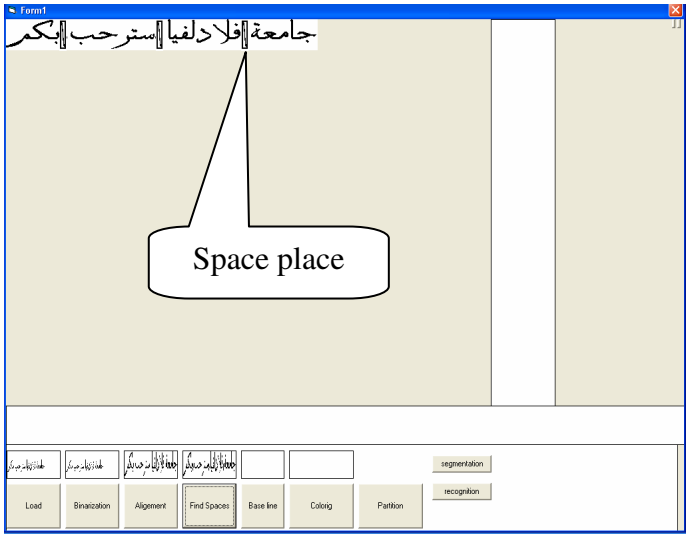
B

Figure (4-3): Alignment step implementation (A) image before alignment
(B) Image after alignment

► Find space between words step implementation

The location of the spaces between words should be determined to facilitate segmentation of the words and recognition of these words. Figure (4-4) illustrates the implementation result of algorithm (3) referred to in section 3.3.3.

A

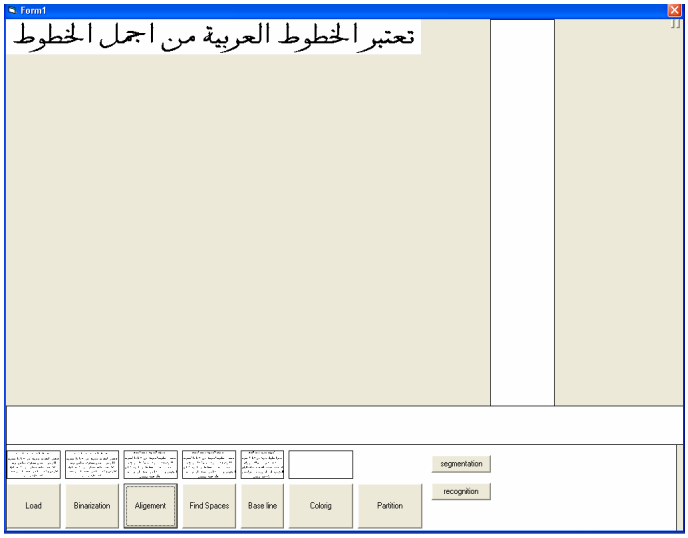


B

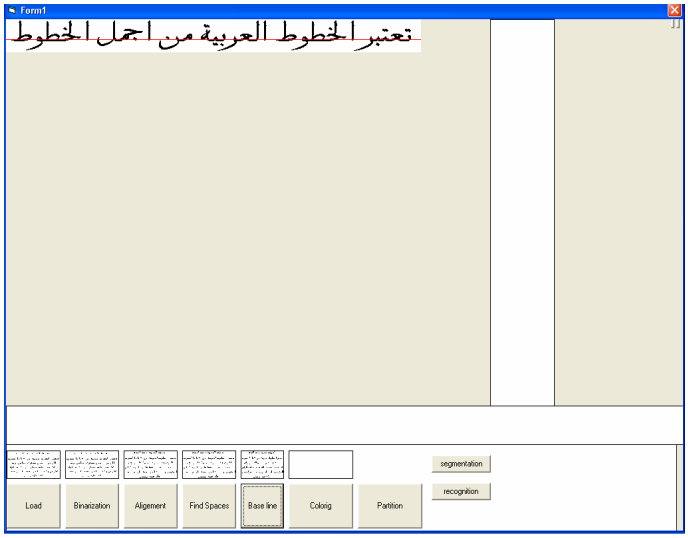
Figure (4-4): Find space between words step implementation
(a) Image before finding spaces (b) image after finding spaces

► Baseline step implementation

In this step an imaginary line determined to write on it called baseline, baseline is considered an important line to determine segmentation locations as well as in recognition, a projection histogram is used to allocate baseline. Figure (4-5) illustrates the implementation result of algorithm (4) referred to in section 3.3.4.



A



B

Figure (4-5): Baseline step implementation (A) image before baseline determination (B) Image after baseline determination

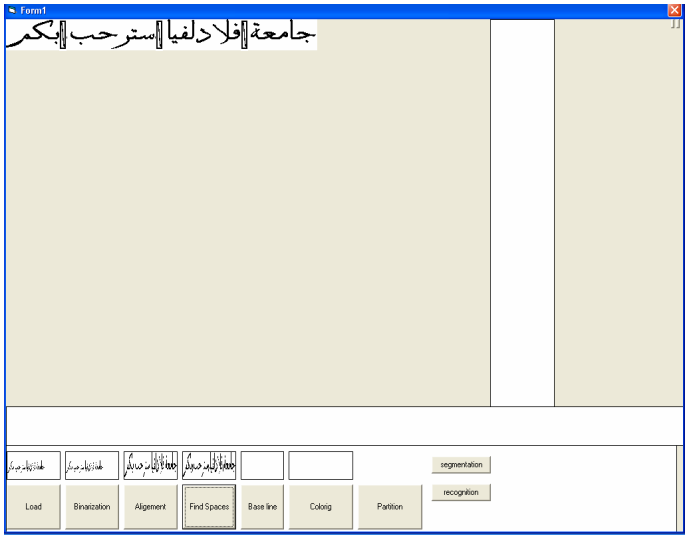
Baseline has an advantage on the level of long words that have different characters, while it has disadvantage on the level of short words, and may affect at the segmentation stage and later on its recognition.

4.4 Segmentation stage implementation

Segmentation is the most important stage in character recognition stage due to the complexity and the overlapping of Arabic characters, as referred to in section 3.4 segmentation passes through three steps; text lines location, connection parts location and segmentation.

► **Connection part location step implementation**

After determination of text line locations in the text image, connection parts determined in each line through two main steps coloring and transmitting text line into connection parts, in the coloring step, figure (4-6) illustrates implementation results of algorithm (5) referred to in section 3.4.2.



A

B

Figure (4-6): Coloring step implementation (A) image before coloring connection parts (B) Image after coloring connection parts

It is remarkable from figure (4-6) that each connection part colored by a color differ than another connection parts, taking into consideration that the space considered as a connection part, also white and black colors excluded.

In transmitting coloring text line into connection parts step the colored text is transmitted into its connection part. The implementation of this step performed through algorithm (6) figure (3-9), algorithm (7) figure (3-14) and algorithm (8) figure (3-15). The implementation of the three mentioned algorithms is illustrated in figure (4-7).

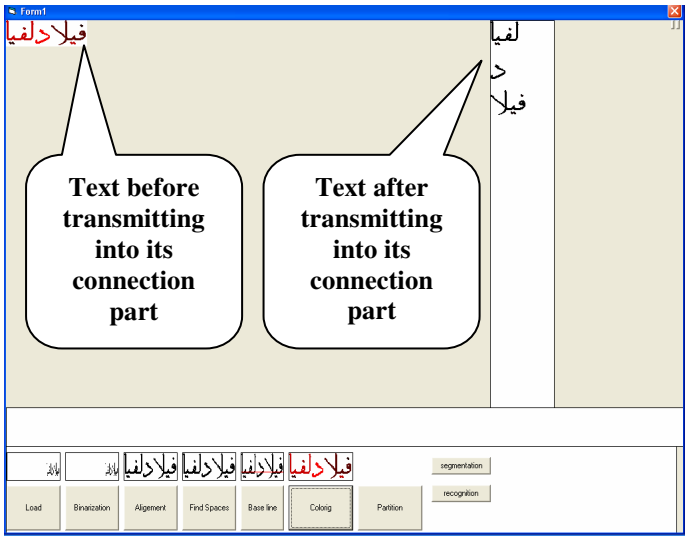


Figure (4-7): Transmitting colored text into its connection parts

► **Segmentation of connection parts into characters step implementation**

After transmitting colored text into its connection parts, each connection part segmented into its characters as illustrated in algorithm (9) section 3.4.3 with a notice that segmentation implemented according to baseline which consist of three vertical points, figure (4-8) illustrates implementation result of algorithm (9).

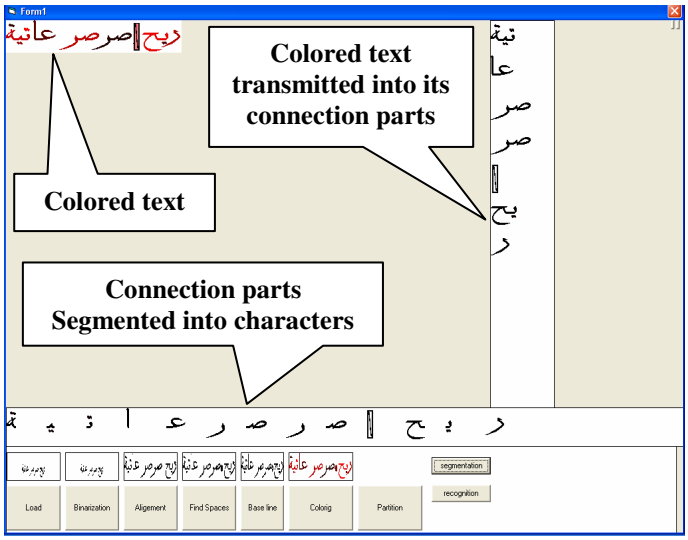


Figure (4-8): Segmentation of connection parts into characters

4.5 Matching stage implementation

Recognition of the segmented characters implemented through the matching of the segmented character with the saved matching characters in the program assigned for matching using matrix matching method, the matching characters were saved in library as illustrated in figure (4-9).

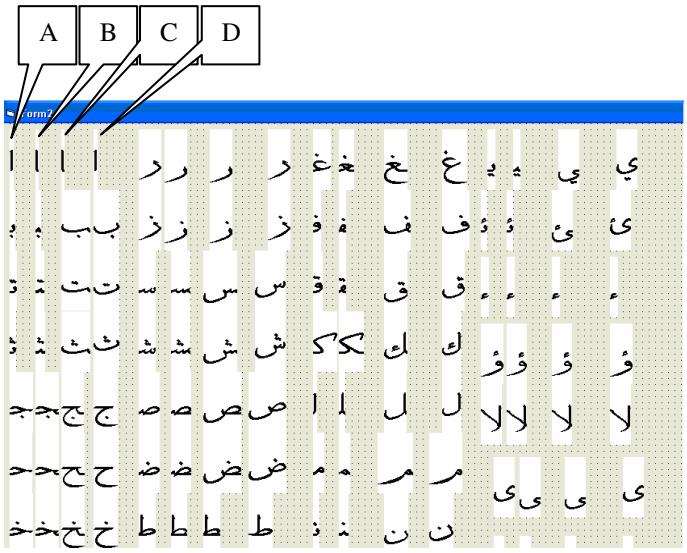


Figure (4-9): A group of matching characters

(A) Beginning characters

(B) Middle characters

(C) End characters

(D) Isolated characters

In this stage a matching is implemented for each segmented character image with matching characters images where the absolute difference between the width of the matching characters and the segmented character did not exceed 5 pixels, matching will start as shown in algorithm (10) figure (3-21), the identical image is that image that has the lowest difference in pixels of the segmented character image with a matching characters images, but when there is no differences in pixels between the segmented character image and the matching characters images then the matching character image is the identical image of the segmented character. Figure (4-10) illustrates the matching process character by character.

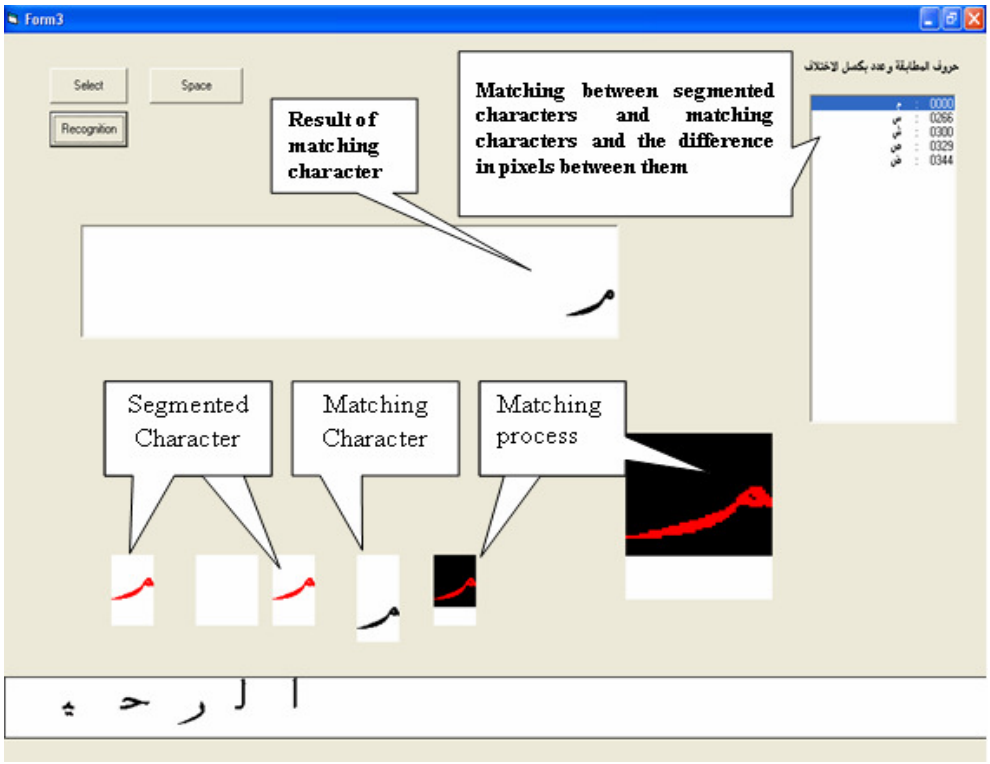


Figure (4-10): Matching process character by character

After finalizing all the mentioned stages, the final result of the implementation of the program is illustrated in figure (4-11)

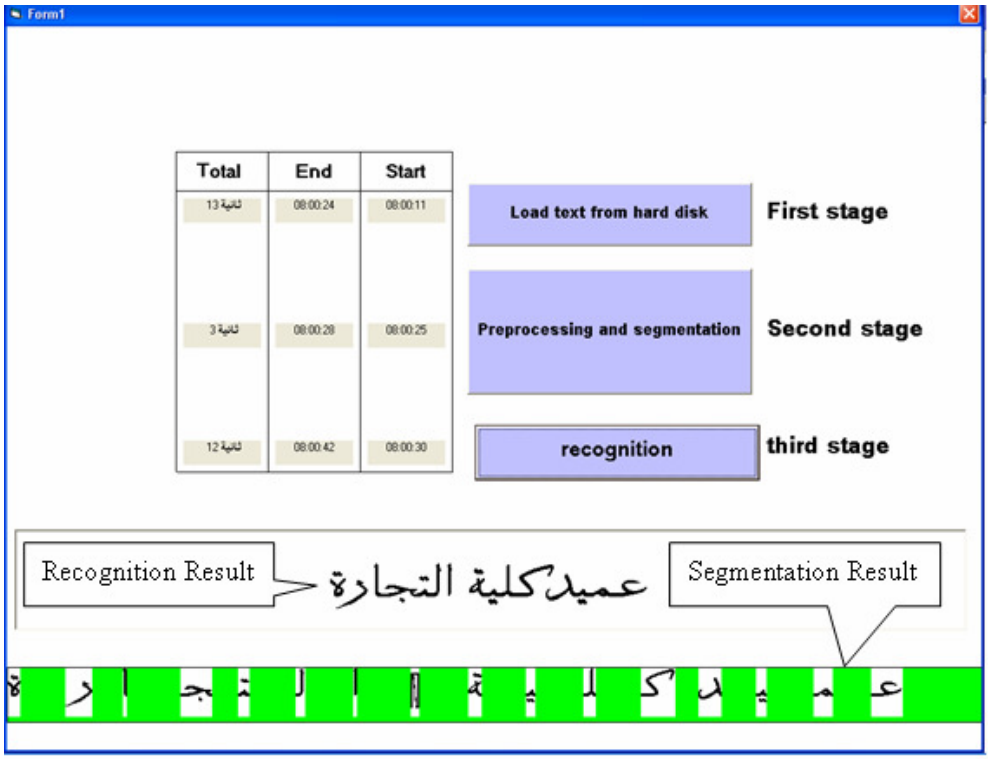


Figure (4-11): Final result of program implementation

CHAPTER FIVE
SYSTEM TESTING AND RESULT COMPARISON

CHAPTER FIVE
SYSTEM TESTING AND RESULT COMPARISON

5.1 Results of system and analysis

Section 2.5 in chapter two summarized some of the previous researches works in the field of Arabic character recognition, in each of these works efforts, methodologies, results, advantages and disadvantages explained for the systems developed.

The thesis focused on Arabic text recognition due to the unique characteristics of Arabic language such as cursive characters, overlapping and ligatures, where overlapping and ligatures considered two of the main problems faces researchers in Arabic text recognition. In addition to the two main ways of Arabic text writing which are handwritten and printed, each of which can be written normally or in artistic way.

Many researches specialized in printed and handwritten Arabic text with normal not artistic styles, few of these researches work on artistic Arabic characters, so this thesis concentrated on artistic Arabic characters, thulth font was selected from different types of artistic Arabic fonts such as Kufi, Diwani, Naskh, Farsi, Ruqaa, and for the purposes of scientific researching.

Testing of the proposed system implemented on five groups of sample; one character called ch1, two character words called ch2, three character words called ch3, four character words called ch4, mixed words called ch5. Segmentation and recognition accuracy calculated by the following equations-:

$$\text{Segmentation accuracy (\%)} = \frac{\text{number of characters correctly segmented}}{\text{total number of characters}}$$

$$\text{Recognition accuracy (\%)} = \frac{\text{number of characters correctly recognized}}{\text{total number of characters}}$$

The following table (5-1) represents the accuracy percentage of segmentation and recognition of the proposed system.

Table (5-1): Segmentation and recognition accuracy

Word type	Segmentation accuracy (%)	Recognition accuracy (%)
Ch1	**91.12**	**70.2**
Ch2	**95.58**	**95.38**
Ch3	**87.9**	**95.75**
Ch4	**86.3**	**90.15**
Ch5	**91.6**	**94.49**
Average	**90.39**	**89.19**

Through the table above, segmentation and recognition accuracy was satisfactory. In a general view on the table above it is noticeable that there was a decrease in the accuracy of the recognition in the level of one character (ch1), the reason for this decrease was that segmentation and recognition depended on the baseline which differ than the baseline for long words and sentences, where the recognition of the short words was low, while for the long words was high, as referred to in section 3.3.4.

The accuracy of the recognition of one character word (Ch1) was relatively low due to that the baseline of one character differ than the baseline of a sentence where sentence baseline was accredited for the recognition. And because of the rarity of one character word in Arabic language, one character word (ch1) can be excluded from the table above, which will lead to an increase in the accuracy of recognition. Figure (5-1) shows a chart of segmentation and recognition accuracy according different type of words.

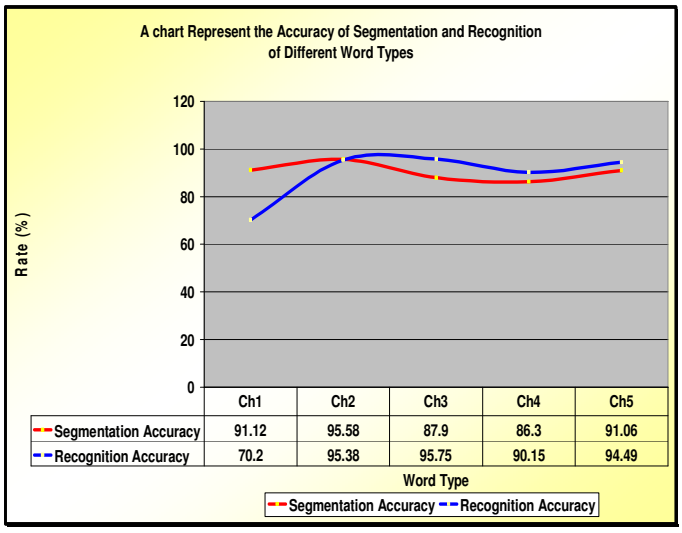


Figure (5-1): Segmentation and recognition accuracy

The problems that have had a negative impact on the process of segmentation and recognition are the following : –

Segmentation of connection parts into characters using algorithm (9) figure (4-8), succeeded in most of the Arabic characters, except of some characters (ب،ت،ث،ف) as an isolated characters, not segmented in a right way due to its location on the baseline which not achieve the required conditions, where part of the character is found away from the baseline, in addition to another problem with the character (ســ) which is segmented into two parts, these problems resolved programmatically. Figure (5-2) illustrates these problems.

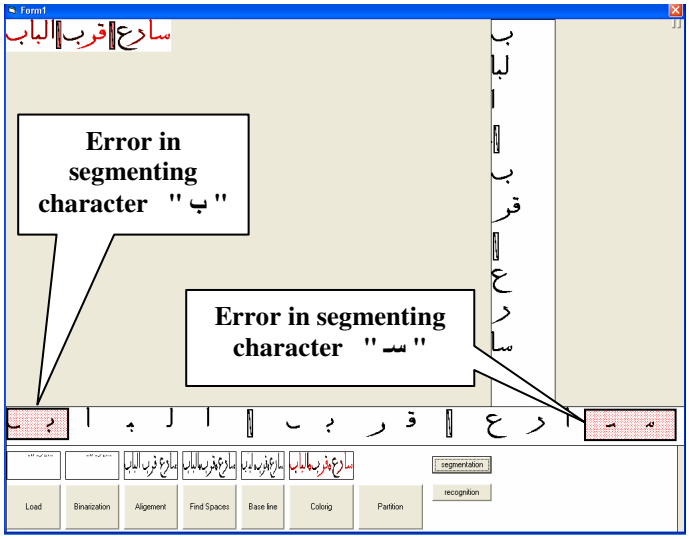


Figure (5-2): Problems in segmentation of the characters " ب ، ســ"

Due to that segmentation process in the proposed system of Arabic text start from left to right, one of the un-resolved problems is the segmentation of some words when the character (ا) come separately after another characters such as (ر،ز،و) , for example the word (ابراهيم) , because the lower left corner of the character (ر) comes before the middle character (ا) then the character (ر) will be segmented before the middle character (ا) so the result of the segmentation of the word will be error "ا ا ب ر ه ي م", as shown in figure (5-3).

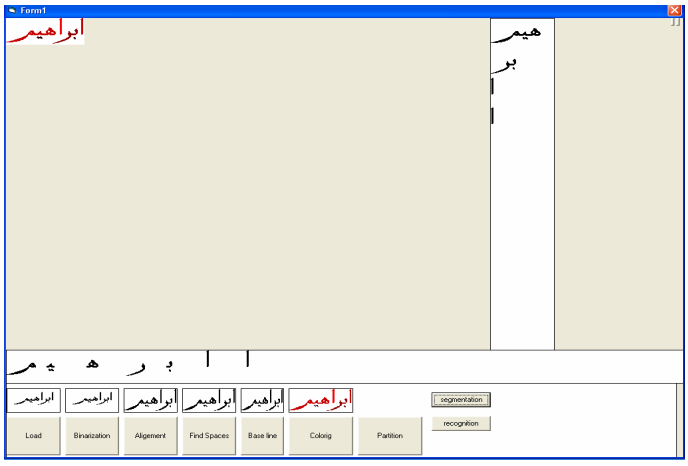


Figure (5-3): Problem in segmentation the word "ابراهيم"

During recognition process all kashidas " ـ " within segmented characters erased to get the final results of the characters same as its original shape, and regarding the space complexity the system deals with it as one space between words as shown in figure (5-4).

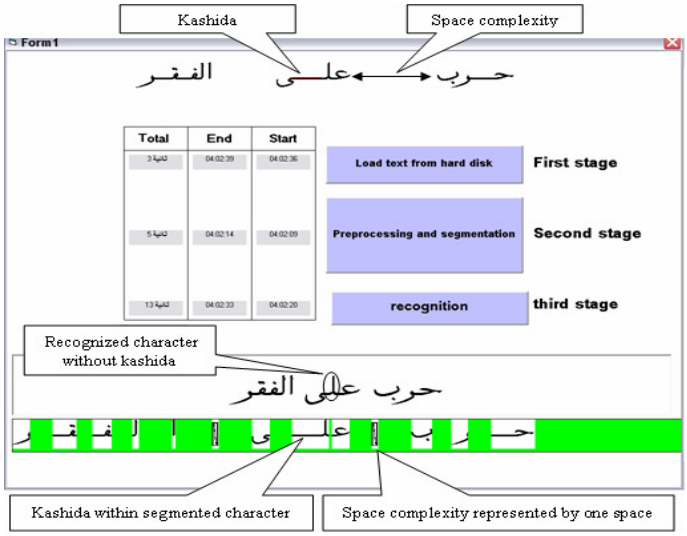


Figure (5-4): kashida and space complexity

Regarding segmentation and recognition execution time, it was different from word type to another; table (5-2) illustrates segmentation and recognition execution time.

Table (5-2): Segmentation and recognition execution time

Word Type	**Segmentation** Execution **Time (ch/s)**	**Recognition** Execution **Time (ch/s)**
Ch1	3.8	0.96
Ch2	5.9	1.03
Ch3	5.3	1.05
Ch4	4.7	1.13
Ch5	4.7	1.08
Average	4.9	1.05

The recognition execution time was not acceptable due to the followings:

- the size of the font used in the system was thulth 36, bold which is considered large if compared with another sizes of the same font type such as sizes 12 or 14, and that reflected on the number of pixel of a character, such as the number of pixels of the

character "ح" with a size 36 equals 756 pixels, while the number of pixels of the same character with a size 12 equals 81 pixels which is equal ninth the "ح" with a size 36, and so on for all characters with different ratios. That was a reasonable justification for the slight increase in the execution time of recognition.

- Matching in the system done by comparing the segmented character with matching characters under the condition that the difference between the segmented character and matching characters did not exceed 5 pixels, which led to take longer time for comparing due to the existence of many matching characters conformed the condition, for example there are 64 matching character have the same difference with the segmented character "ح" that means equal or less than 5 pixels, while there are 46 matching character have the same difference with the segmented character "د". The effect of this reason was the increase of the execution time of recognition. To avoid this problem, the difference between the segmented character and matching characters should be decreased to one or two pixels to decrease the execution time of recognition.

5.2 Results comparison

The researcher M.S. Khorsheed worked in his paper titled " Offline recognition of omnifont Arabic text using the HMM Toolkit (HTK)" on the recognition of different types of Arabic fonts Thuluth, Naskh, Simplified Arabic, Traditional Arabic, Tahoma and Andalus. The results of his paper was 87.6%, 86%, 88%, 89.5%, 92.1% and 92.4% respectively (Khorsheed, 2007).

The researcher Al-Qahtani and others worked in their paper titled "Recognizing Cursive Arabic Script Using HMMs" on the recognition of three types of Arabic fonts Thuluth, Simplified Arabic and Naskh, the results of their paper was 87.81%, 94.65% and 87.1% respectively (Al-Qahtani at el, 2004).

The result of the proposed system recognition accuracy percentage for the Thulth font used in the thesis found (89.19%) which is considered relatively high.

CHAPTER SIX
CONCLUSION AND FUTURE WORKS

CHAPTER SIX
CONCLUSION AND FUTURE WORKS

6.1 Conclusion

The main aim of the thesis is to build a system with an ability to recognize characters of the artistic Arabic text, using Thulth style, and that by selecting appropriate techniques in all the processes of recognition in this thesis that lead to the purpose of segmentation the Thulth words into characters and recognizing these characters properly.

By examining the system, it was able to segment line into words then characters, followed by the successful recognition of each character of the word.

Many algorithms used in the system starting from preprocessing and ending with recognition, coloring algorithm and connection part algorithm was the most important algorithms in the system. Where coloring algorithm function is to color each part in the word with a color differ than the other parts, to facilitate segmenting the word into parts. Connection part algorithm function is to segment colored word into connection parts according the colors given in the coloring algorithm. The following points describe the most important results obtained:

- Overlap problem is solved in this thesis.
- There are some characters connected vertically called ligatures such as: لـ.ـهـ.ـلـ لـ.ـهـ .ـهـ.ـلا , these ligatures did not segmented by the system but considered as a one character image.
- Some obstacles occurred in the implementation of the system, especially during segmentation of connection parts into characters specifically in the characters "ت ، ث ، ب ، س" where this problem overcome in the recognition stage.
- Because the segmentation in the system starts from left to right, another obstacle appeared when the isolated character "ا" followed the characters "ر ، و، ز" in the same word, where the character "ر ، و، ز " segmented before the character "ا", this obstacle did not overcome in the thesis.
- The final results achieved after examination and analyzing of the system was satisfactory, where the segmentation accuracy reached 90.39%, recognition accuracy reached 89.19%.

- The segmentation execution time was 4.9 ch/s. and recognition execution time reached 1.05 ch/s. it seems from the mentioned results that the recognition execution time was slightly long due to two reasons, firstly the size of the font used and secondly the difference in pixels between the segmented character and the matching characters.

6.2 Future Works

- The need to develop the segmentation methods of Arabic characters to overcome ligatures.
- The need to expand the researches in this field to include a wide variety of artistic Arabic font with different sizes.
- The need to develop algorithms to increase its flexibility during managing variables in the recognition system.
- The need to expand researches to find the logic solution for the complicated overlapping and ligatures.

REFERENCES

[Abuhaiba] I. S. I. Abuhaiba, S. A. Mahmoud, and R. J. Green, (1994). "Recognition of Handwritten Cursive Arabic Characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 6, pp. 664-672.

[Aburas] Abdurazzag Ali Aburas and Salem M. A. Rehiel ,(2007). "Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression", ARISER Vol. 3 No. 4, pp. 123-135

[Al-A'ali] Mansoor Al-A'ali and Jamil Ahmad, (2007). "Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach", Journal of Computer Science 3 (7), ISSN 1549-3636, pp 549-555.

[Al-Qahtani] S. A. Al-Qahtani, M. S. Khorsheed and M. A. Al-Suliman, (2004). "Recognizing Cursive Arabic Script Using HMMs", 17th national conference for computer science, King Abdel Aziz University.

[Amin] Adnan Amin, Humoud Al-Sadouni and Stephen Fischer, (1996). "Hand-Printed Arabic Character Recognition System Using An Artificial Network", Pattern Recognition, Vol. 29, No. 4, pp. 663-675.

[Arica] Nafiz Arica, (1998). "An Off-Line Character Recognition System for Free Style Handwriting", Msc. Thesis, Middle East Technical University,

[Cheriet] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, and Ching Y. Suen, (2007). "Character Recognition Systems A Guide For Students and Practitioners ", Published by John Wiley & Sons, Inc., Hoboken, New Jersey

[Cheung] A. Cheung, M. Bennamoun, N.W. Bergmann,(2001). "An Arabic Optical Character Recognition System Using Recognition-Based Segmentation", Pattern Recognition 34, pp. 215-233

[Elgammal] Ahmed M. Elgammal and Mohamed A. Ismail,(2001). "A Graph-Based Segmentation and Feature Extraction Framework for Arabic Text Recognition" , Document Analysis and Recognition, Sixth International Conference pp. 622 - 626

[Fakir], Mohamed Fakir and M. M. Hassani, (2000). "On The Recognition of Arabic Characters Using Hough Transform Technique", Malaysian Journal of Computer Science, Vol. 13 No. 2, pp. 39-47

[Hamami], Latifa Hamami and Daoud Berkani, (2007). "Recognition System for Printed Multi-Font and Multi-Size Arabic Characters", the Arabian Journal for Science and Engineering, Volume 27, Number 1B.

[Khorsheed] M. S. Khorsheed, (2002). "Off-Line Arabic Character Recognition – A Review", Springer, Pattern Analysis & Applications 5, pp. 31–45.

[Khorsheed] M.S. Khorsheed, (2007). "Offline recognition of omnifont Arabic text using the HMM Toolkit (HTK)", Pattern Recognition Letters 28 pp. 1563–1571.

[Liu] Jie Liu, Jigui Sun, Shengsheng Wang, (2006). "Pattern Recognition: An overview", IJCSNS International Journal of Computer Science and Network Security, Vol. 6 No. 6.

[Mostafa], Mostafa G. Mostafa, (2004). "An Adaptive Algorithm for the Automatic segmentation of Printed Arabic Text", 7th National conference of computer automation, King Abdul Aziz University, Al-Madinah Al-Munawwarah, pp. 437-444.

[Motasswa] Deya Motawa, Adnan Amin and Robert Sabourin, (1997). "Segmentation of Arabic Cursive Script", IEEE, pp. 625-628

[Omidyeganeh] M. Omidyeganeh, K. Nayebi, R. Azmi and A. Javadtalab, (2005). "A New Segmentation Technique for Multi Font Farsi/Arabic Texts", IEEE, pp. 757-760.

[Safabakhsh] Reza Safabakhsh and Peyman Adibi, (2005). "Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM", the Arabian Journal for Science and Engineering, Vol. 30, No. 1B.

[Schmieg] Sebastian Schmieg, (2007). "History of Arabic Type Evolution from the 1930's till present". http://29letters.wordpress.com.

[Wikipedia] http://en.wikipedia.org/wiki/Template_matching

[Zeki] Ahmed M. Zeki, (2005). "The Segmentation Problem in Arabic Character Recognition the State Of The Art", IEEE, pp. 11-26.

[Zheng] Liying Zheng, Abbas H. Hassin and Xianglong Tang, (2004). "A new algorithm for machine printed Arabic character segmentation", Pattern Recognition, pp. 1723–1729.