OFF-LINE HANDWRITTEN ARABIC CHARACTERS SEGMENTATION USING SLANT-TOLERANT SEGMENT FEATURES (STSF)

by

SHUBAIR A. ABDULLAH

Thesis Submitted in Fulfillment of the

Requirements for the Degree of Master of Science

April 2007

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Rosalina Abdul Salam, for her help, guidance, and advice throughout this research.

And also I would like to thank:

- *Dr. Amer Al Nassiri, for his guidance.*
- *My family, for their fully support and encouragement.*

TABLE OF CONTENTS

CHAPTER 4 – THE ARABIC HANDWRITTEN DATABASE – AHD/USM

CHAPTER 5 – USING SLANT-TOLERANT SEGMENT FEATURES (STSF) FOR OFF-LINE HANDWRITTEN ARABIC CHARACTER SEGMENTATION

CHAPTER 6 – EXPERIMENTAL RESULTS AND DISCUSSIONS

CHAPTER 7 – CONCLUSION

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AHD/USM	: Arabic Handwritten Database/Universiti Sains Malaysia
AHDB	: Arabic Handwritten Database
CEDAR	: Center of Excellence for Document Analysis and Recognition
CR	: Character Recognition
DSTSF	: Decisive Slant-Tolerant Segment Feature
FSP	: Final Segmentation Point
JSA	: Junction-Seeking Approach
MSE	: Mean Squared Error
NIST	: National Institute of Standards and Technology
NSP	: Nominated Segmentation Point
PSTSF	: Primary Slant-Tolerant Segment Feature
STSF	: Slant-Tolerant Segment Feature
RMSE	: Root Mean Squared Error
RSA	: Recognize-Segment Approach

PENSEGMENAN AKSARA TULISAN ARAB LUAR TALIAN MENGGUNAKAN CIRI SEGMEN TOLERATE-KECONDONGAN (STSF)

ABSTRAK

Tema utama bagi kajian ini ialah pensegmenan aksara tulisan Arab luar talian. Suatu sistem pengecaman aksara tulisan Arab yang baik mampu meningkatkan kesalingtindakan antara manusia dengan komputer. Pembangunan suatu sistem pengecaman aksara Arab yang baik tidak dapat dilakukan tanpa menyelesaikan dahulu masalah pensegmenan. Salah satu masalah ialah kecondongan sesuatu kata ketidaksamaan panjang dan kecerunan garis simpang antara aksara tulisan. Cabaran di sini adalah untuk mendapatkan ciri-ciri morfologi dan mencari garis simpang. Sumbangan terpenting kajian ini ialah penghasilan satu algoritma pensegmenan baru untuk aksara tulisan Arab menggunakan Ciri-ciri Segmen Tolerate-Kecondongan (Slant-Tolerant Segment Feature – STSF). STSF merupakan satu rentetan simbol '+' atau '-' dan mewakili lejang aksara kata. Ujian terhadap algoritma STSF telah dijalankan menggunakan pangkalan data yang baru dibina. Pangkalan data yang baru ini dinamakan Pangkalan Data Tulisan Arab/Universiti Sains Malaysia (AHD/USM). Pangkalan Data tulisan Arab mengandungi 12300 kata yang telah ditulis oleh 82 penulis berbeza yang berusia daripada 5 tahun hingga yang berusia melewati 45 tahun. Kajian ini melaporkan kadar pembetulan sebanyak 90.12%. Kajian ini juga telah menyumbang kepada pemudahan kajian masalah pensegmenan aksara Arab dengan mewujudkan sistem kategorian baru untuk kaedah-kaedah pensegmenan. Ia mengkategorikan kaedah-kaedah pensegmenan kepada dua pendekatan: Pendekatan Pencarian Simpang (JSA) dan Pendekatan Mengecam Segmen (RSA).

OFF-LINE HANDWRITTEN ARABIC CHARACTERS SEGMENTATION USING SLANT-TOLERANT SEGMENT FEATURES (STSF)

ABSTRACT

The main theme of this research is the off-line handwritten Arabic characters segmentation. A successful handwritten Arabic character recognition system improves interactivity between the human and the computers. Building successful Arabic character recognition system cannot be fulfilled without solving the segmentation problem. One of the major problems is the slant of the words and the dissimilarity in length and slope of the junction line between the handwritten characters. The challenge is to extract the morphological features and to find the junction line. The foremost contribution of this research is the new segmentation algorithm for handwritten Arabic characters using the Slant-Tolerant Segments Feature (STSF). STSF is a string of '+' or '-' signs and represents the strokes of the word characters. The testing of the STSF algorithm was conducted using a newly created database. The new database is called Arabic Handwritten Database/Universiti Sains Malaysia (AHD/USM). The Arabic handwritten database consists of 12300 words written by 82 different writers aged from 5 to more than 45 years. The experiments reported 90.12% correctness rate. The research also contributes in facilitating the study of the Arabic characters segmentation problem by creating new categorization system for the segmentation methods. It categorizes the segmentation methods into two approaches: Junction-Seeking Approach (JSA) and Recognize-Segment Approach (RSA).

Chapter 1

INTRODUCTION

1.1 CHARACTER RECOGNITION (CR)

Character Recognition (CR) automation occupies a big and intensive research zone of the pattern recognition research area. It has attracted the attention of researchers during the last five decades. CR automation means translating images of characters into an editable text; in other words, it represents an attempt to simulate the human reading process.

The CR can be an on-line or off-line type. If the CR system has the ability to trace the points generated by moving a special pen on a special screen, the system belongs to the on-line type, while the system belongs to the off-line type when it accepts only the pre-scanned text images to perform the recognition function. The off-line type deals with printed and handwritten texts, while the on-line type deals with handwritten texts only. See Figure 1.1.

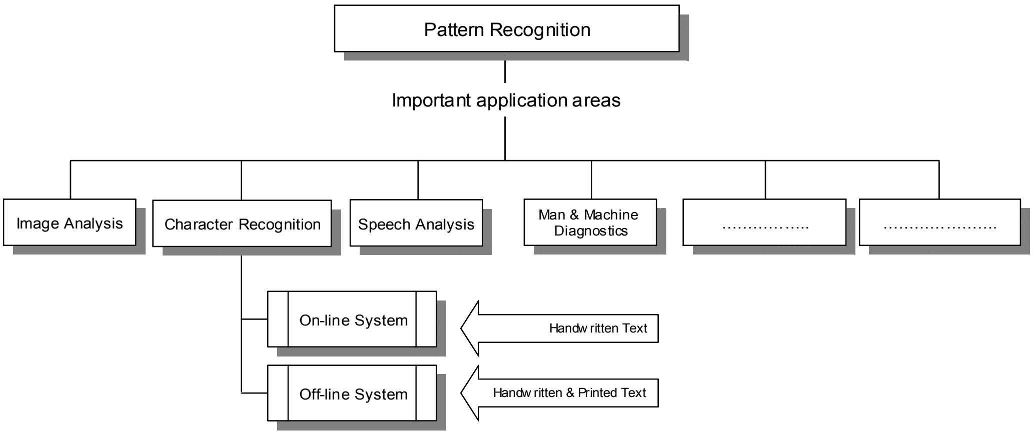


Figure 1.1 The pattern recognition and the character recognition

The off-line CR system is divided into predefined processes which yield a recognized text. There are four main processes which apply equally importantly to any off-line CR system: Preprocessing, Segmentation, Features Extraction, and Recognition.

The number of these processes is standard even if it is different in some off-line CR systems. Figure 1.2 illustrates the standard processes of the off-line CR system. The first process of Preprocessing takes the raw data which is a text image scanned by data acquisition hardware, then it performs functions such as noise reduction and normalization to produce a clean data. This clean data is analyzed to subcomponents: lines, words, and characters by the Segmentation process which is the subject of this research. Upon finishing of the Segmentation process, the third and fourth processes are performed. The more accurate subcomponents are produced by the Segmentation process the higher recognition rate obtained from the off-line CR system.

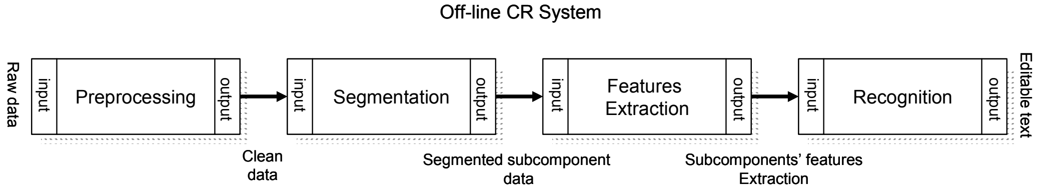


Figure 1.2 The off-line CR system, four standard processes (Preprocessing, Segmentation, Features Extraction, and Recognition)

The earliest attempts to automate CR can be traced to the middle of the last century (Amin A., 2001) and (Richard G. et al, 1996). The data acquisition development had an impact upon the automation of CR and paved the way for the researchers to express their theories in the CR automation. Although the first colored scanner was invented in 1948 by Kodak as in (Hoskins S., [online] accessed 12th July 2006), the data acquisition deployment correlated with the deployment of the microcomputer. A lot of research on the CR was done after the explosion in the hardware development especially in the computer hardware and data acquisition hardware. Regarding the automation of Arabic CR, the emergence of researches started in 1980s (Bushofa B. M. F. et al, 1997) a long time after the first attempts of Latin CR automation, this delay is due to the following:

1. The microcomputer and data acquisition became available everywhere in the 1980s and beyond.

2. The cursive nature of the Arabic script which calls for solving the segmentation problem.
3. The lack of adequate databases of Arabic words which can be used as a test bench similar to NIST and CEDAR databases of Latin words.

1.2 STATEMENT OF THE RESEARCH AND OBJECTIVES

This research focuses on the problem of segmenting the Arabic characters to produce successful off-line CR system. The difficulty of the off-line CR system increases when the system deals with handwritten cursive scripts like the handwritten Arabic scripts where the extraction of morphological features is a complicated task. The segmentation algorithm, in such case, tries to find and remove the junctions between the characters. Although the spoken Arabic is somewhat different from a country to another country in the Arab world, the writing is a standard version used by all the Arab people for their communication. The handwritten Arabic CR system is considered by more than 18 nations who are using the Arabic characters in their writing (Zeki A., 2006). Moreover, a successful handwritten Arabic CR system improves the interactivity between the human and the computers through bank checks validation, digital archiving manuscripts, reading and processing invoices and receipts. Therefore a successful handwritten Arabic CR system is extremely beneficial, and its successfulness cannot be fulfilled without overcoming the difficult problem of the segmentation stage. This research seeks to fulfill the following objectives:

1. To overcome the problems of the segmentation such as overlapping, and slanting.
2. To advance the segmentation and recognition of the handwritten Arabic script by providing a wide range of handwriting styles in a database to be used as a test bench.
3. To facilitate the study of the off-line Arabic characters segmentation problem by categorizing and analyzing the segmentation methods.

1.2.1 CONTRIBUTIONS

The main contribution of the research is the new algorithm for the off-line handwritten Arabic characters segmentation using Slant-Tolerant Segment Feature (STSF). The research also contributes to the field by introducing the following:

1. An adequate Arabic handwritten database for the experiments.
2. A categorization system for the segmentation methods in which the segmentation methods are categorized into two approaches: Junction-Seeking Approach (JSA) and Recognize-Segment Approach (RSA).

1.2.2 METHODOLOGY

The idea of the STSF algorithm hypothesizes that every direction rolled to write one character (see Figure 1.3) represents an area. It divides the upper contour of the word into areas according to the strokes.

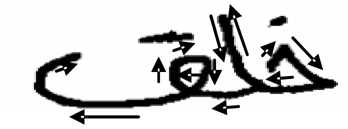


Figure 1.3 The direction rolled in the course of writing

The algorithm is categorized as Junction-Seeking Approach (JSA). The following steps summarizes the junction seeking process: (Figure 1.4)

1- Smoothing the contour of the image by eliminating any point created improperly.

2- Dividing the word or sub-word into areas according to its strokes.

3- Setting apart of the areas that may contain the segmentation point.

4- Preparing a list of nominated segmentation points (NSP).

5- Finding the final segmentation points (FSP) from the nominated list.

The algorithm has been tested by using a new Arabic handwritten database.

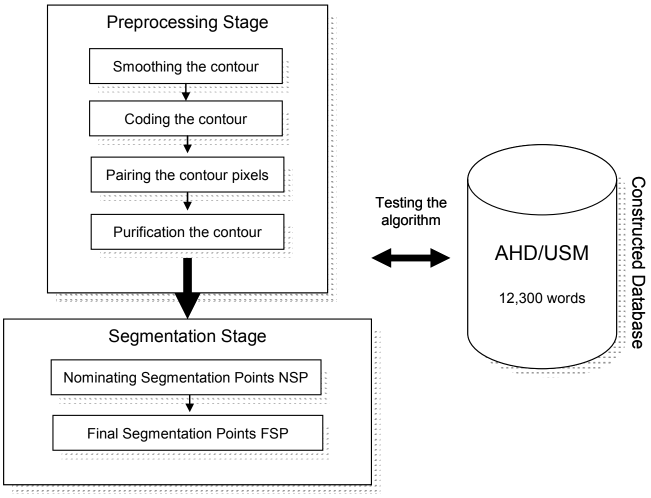


Figure 1.4 The central concepts of the research

1.3 THESIS OUTLINE

Chapter 2 gives the background of the Arabic script features and describes the character segmentation and techniques used along with the segmentation process. It also includes the definition of Junction-Seeking and Recognize-Segment approaches.

Chapter 3 reviews the literature and categorizes the method into Junction-Seeking and Recognize-Segment approaches. The segmentation of Latin and Chinese character is also investigated.

Chapter 4 explores the current databases for the Arabic words and introduces the AHD/USM database.

Chapter 5 explains the stages of the STSF algorithm. It also summarizes the features and the implementation of the STSF algorithm.

Chapter 6 discusses the experimental results and shows the evaluation of the STSF algorithm.

Chapter 7 summarizes the work covered by citing the contribution to the field of Arabic character recognition, and suggests some possible future works.

Chapter 2

CHARACTER SEGMENTATION

2.1 THE ARABIC SCRIPT

The Arabic script consists of 28 letters. Each letter has two or more shapes depending on the adjacent letter in the word as shown in Table 2.1.

Each letter has at least one stroke. The strokes of most of the letters lie in the upper part of the letter. The position of the strokes causes the Character Recognition (CR) system to employ the strokes in segmenting the word.

Table 2.1 Arabic characters' shapes

No	Character	Character's Shapes								
1	Alif	ا	ـا	ئ	ئـ	ء	أ	ؤ		
2	Baa	ب	بـ	ـب	ـبـ					
3	Taa	ت	تـ	ـت	ـتـ	ة	ـة			
4	Thaa	ث	ثـ	ـث	ـثـ					
5	Jeem	ج	جـ	ـج	ـجـ					
6	Hhaa	ح	حـ	ـح	ـحـ					
7	Khaa	خ	خـ	ـخ	ـخـ					
8	Dal	د	ـد							
9	Thal	ذ	ـذ							
10	Raa	ر	ـر							
11	Zay	ز	ـز							
12	Seen	س	سـ	ـس	ـسـ					
13	Sheen	ش	شـ	ـش	ـشـ					
14	Sad	ص	صـ	ـص	ـصـ					
15	Dhad	ض	ضـ	ـض	ـضـ					
16	Tta	ط	طـ	ـط	ـطـ					
17	Thaa	ظ	ظـ	ـظ	ـظـ					
18	Ain	ع	عـ	ـع	ـعـ					
19	Ghain	غ	غـ	ـغ	ـغـ					
20	Faa	ف	فـ	ـف	ـفـ					
21	Qaf	ق	قـ	ـق	ـقـ					
22	Kaf	ك	كـ	ـك	ـكـ					
23	Lam	ل	لـ	ـل	ـلـ					
24	Meem	م	مـ	ـم	ـمـ					
25	Noon	ن	نـ	ـن	ـنـ					
26	Haa	ه	هـ	ـه	ـهـ					
27	Waw	و	ـو							
28	Yaa	ي	يـ	ـي	ـيـ					

The strokes of a letter may resemble the strokes of other letters; therefore the dot is used to discriminate between letters, as shown in Figure 2.1. The position of the dot is either upper or under the letter. The strokes in some letters form a cavity, which may be considered as one of the letter features. The number of cavities in all letters does not exceed 2 cavities, for example the letter "ـهـ" has 2 cavities, and the letters "ـف", "ص", "م" and "و" have one cavity. There is a small number of letters which have the same shape in any position like 'و' and 'ر'.

A- two letters (Tta) and (Thaa) with the same strokes B- two letters (Seen) and (Sheen) with the same strokes

Figure 2.1 The dot used to discriminate between two letters with the same strokes

The Arabic script is a right to left cursive script. The words in the line are separated by spaces and the sub-word is generated when containing unconnectable letters like "ر", "ز", and "و". The letters in one word or sub-word are connected to each other by a junction line of points. The junction lines pieces formalize the baseline of the word. Some letters are located above the baseline ("ـف" and "ك") and some are located above and under the baseline ("و" and "ح") as shown in Figure 2.2.

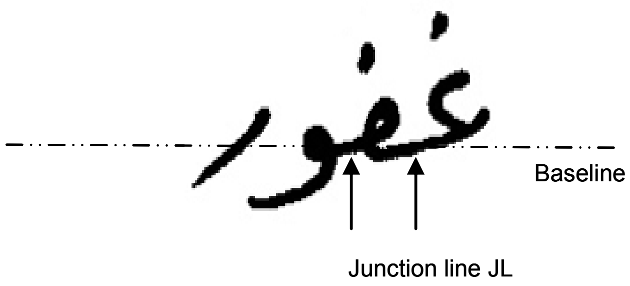


Figure 2.2 Letter's position according to the baseline

The Arabic letter may have diacritics. A diacritic is a stroke written above or below the letter and it is important in the pronunciation and meaning of the word. Table 2.2 shows the diacritics of the Arabic language. The letter in the Arabic scripts is identified by a unique features bundle which contains the following features:

1. Number and shape of the strokes.
2. Number and position of the dot.

3. Number of cavities.

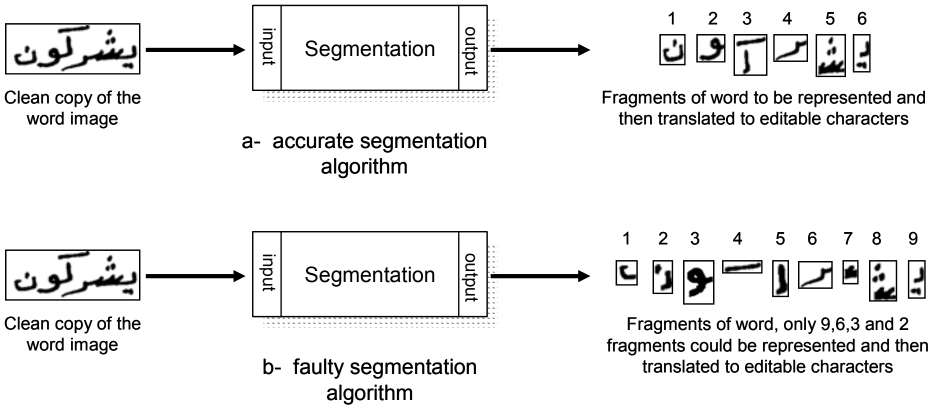
4. Connectable or not connectable.

Table 2.2 Arabic diacritics

S	Diacritic	Example	Position
1	Fatha	كَسَرَ	Above
2	Kasra	إِبِل	Below
3	Dhamma	قُدُس	Above
4	Sukun	نوْر	Above
5	Shadda	مرّ	Above
6	Madda	رآه	Above
7	Tanween - Double Fatha - Double Kasra - Double Damma	 باسماً باسمٍ باسمٌ	 Above Below Above

2.2 THE SEGMENTATION PROCESS

The segmentation stage represents the main obstacle in the off-line CR automation task. Character segmentation process seeks to analyze the image of text and decompose it into small sub-images of separated fragments. An accurate segmentation algorithm leads to correct recognition (Zhao X et al, 1994). It produces the images of separated fragments of the word to be represented in a right manner and then translated to editable characters. Figure 2.3 shows the importance of the segmentation process role by giving two instances of the off-line CR system segmentation algorithms, accurate and faulty algorithms. The accurate segmentation algorithm (Figure 2.3-a) produces images of characters while the faulty segmentation algorithm (Figure 2.3-b) produces error images of fragments of characters which may cause omitting a portion of the word in the final output of the off-line CR system.



Clean copy of the word image
input
Segmentation
output
1 2 3 4 5 6
Fragments of word to be represented and then translated to editable characters
Clean copy of the word image
input
Segmentation
output
1 2 3 4 5 6 7 8 9
Fragments of word, only 9,6,3 and 2 fragments could be represented and then translated to editable characters

a- accurate segmentation algorithm

b- faulty segmentation algorithm

Figure 2.3 Two instances of segmentation algorithms, accurate and faulty algorithms

2.3 THE PREPROCESSING OF THE SEGMENTATION

The techniques of preprocessing stage are divided into two stages according to their function:

1. The techniques that produce clean raw data such as: noise reduction, normalization, and smoothing. These techniques are considered as preprocessing for the off-line Arabic character recognition system.
2. The techniques that prepare the data image to be segmented such as: vertical and horizontal projection, contour tracing, and skeleton extraction). These techniques represent the preprocessing of the segmentation process.

Separation between the two sub-lists is necessary to study the segmentation issue. The segmentation approach defines its preprocessing that prepares a concise representation of the image of word in order to be segmented. The following techniques are common preprocessing of the segmentation process.

2.3.1 Vertical and Horizontal Projection

The vertical projection method helps in detecting the white spaces and the junction lines between the adjacent characters by counting the black pixels in each column of the word. Performing the vertical projection transforms the junction lines to a low-thickness vertical lines, Figure 2.4 shows the word and its vertical projection analysis. The overlapping occurs when a letter extends beyond its succeeding letter with or without touching. Although this method is not efficient in the handwritten script due to the overlapping problem (Zahour A. et al, 2001), it has been used with the horizontal projection to analyze the word into lines, words, and characters since the beginning of the CR automation attempts (Mori S. et al, 1999).

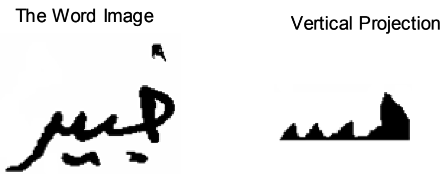


Figure 2.4 Vertical Projection

2.3.2 Thinning (skeleton extraction)

The thinning operation means producing the skeleton of the image. A skeleton is a one pixel width produced by highlighting the centerline of the word. It helps in restoring the essential information about the word (Abuhaiba S. et al, 1994). Figure 2.5 shows a word and its skeleton. Two methods are basically used for thinning, pixel wise and non-pixel wise (Lam L. et al, 1992). The pixel wise method iteratively removes the outer pixels and only the central line remains. This method is highly noise-influenced. On the other hand, the non-pixel wise method is more noise tolerant and is not iteration-based (Peng J., 2000).

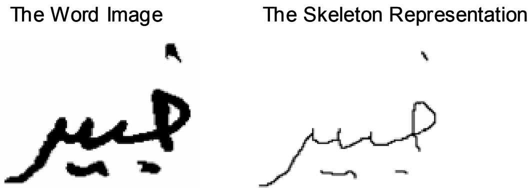


Figure 2.5 The word "خبير" and its skeleton

2.3.3 Contour Tracing

The tracing of the contour aims at transforming the border of the word into a string of codes to extract the features of the image. The coding scheme starts by identifying the position of an initial pixel and continues identifying the relative positions of the successive pixels on the contour until reaching the starting pixel. The Freeman chain code is widely used as a scheme for features extraction which to be employed either for the segmentation process (Romeo-Pakker K. et al, 1995) or for the recognition process (Abu Zitar R., 2005). The Figure 2.6 shows the result of tracing the border of the character "ب". The tracing of the contour is efficient in solving the overlapping problem (Al-Nassiri A., 2006) but in most cases it needs to smooth the border of the word.

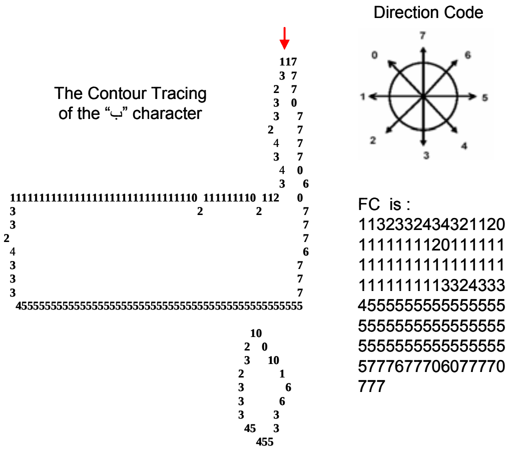


Figure 2.6 The Freeman chain code for the "ب" character

2.3.4 Baseline Detection

The baseline is an imaginary horizontal line to connect the characters of the word. The baseline in the segmentation stage is usually detected. It helps in distinguishing the strokes of the characters. Several methods have been published for detecting the baseline. Kanai in (Kanai J. et al, 1998) utilized the projection profile technique to detect the fiducial points by decoding the lowest resolution layer of the image. Pechwitz in (Pechwitz M. et al, 2003) implemented a method completely based on polygonally approximated skeleton processing. Detecting the baseline is a common step in many off-line handwritten Arabic CR systems and it is often an important step before the segmentation and the feature extraction steps.

2.4 SEGMENTATION METHODS

Several methods have been created to overcome the obstacles of the segmentation in off-line handwritten Arabic CR system. Some methods are derived from the nature of the Arabic scripts and some others are modified and carried out from the Latin cursive scripts. The modification is necessary as the methods of segmentation of Latin scripts, in general, are insufficient for segmenting Arabic characters (Mostafa M. 2004). This insufficiency because of the Arabic script has characteristics that highly influence the segmentation process such as the disparity of Arabic character length (for example "ك" and "أ" and the plurality of strokes in Arabic characters.

2.4.1 CURRENT CATEGORIZATIONS:

The methods of segmentation in CR system have been classified in many ways. Arica in (Arica N. et al, 2001) stated that there are two types of segmentation: external and internal segmentation. In internal type there are three strategies: implicit, explicit, and mixed strategies. The External Segmentation is the isolation of paragraphs, sentences, and words, while the Internal Segmentation is the isolation of letters especially in cursive script. Amin in (Amin A., 1998) said that there are two approaches have been applied in off-line handwritten Arabic CR system: Analytical approach and Global approach. In the Analytical approach, the words are segmented into characters while the recognizer tries to recognize the whole representation of the words in the Global approach. Khorsheed in (Khorsheed M., 2002) believed that the off-line handwritten Arabic CR systems fall into two categories: segmentation-based and segmentation-free systems relative to their approach in tackling word segmentation. He provided four categories for the segmentation-based recognition systems which are: pre-segmented characters, segmenting a word into characters, segmenting a word into primitives, and integrating of recognition and segmentation. Zeki in (Zeki A., 2006), on the other hand, gave somewhat differences in his categorizing. He categorized the segmentation of the Arabic characters based on nine techniques.

2.4.2 PROPOSED CATEGORIZATION

After reviewing the segmentation strategies and techniques that have been applied in the off-line handwritten Arabic CR system during last 20 years, this research is introducing a new categorization system in which the segmentation approaches fall into two types:

1. Junction-Seeking Approach (JSA): in which the system seeks the segments that lead to produce significant fragments of characters.
2. Recognize-Segment Approach (RSA): in which the system seeks any segment that matches the morphological features of the characters.

Accordingly, after completion of the preprocessing stage in the off-line handwritten Arabic CR system, the system either seeks the junctions between characters or tries to find the character's representation based on predefined criteria.

The proposed categorization system classifies the Arabic character segmentation methods into two groups by following an analytical approach. The segmentation method analysis is simplified to a domain of four components:

1. Prototyped criteria
2. Preprocessing techniques
3. Output expected
4. Methodology used to search for the criteria to produce the output.

Although the Arabic character segmentation techniques have different approaches, they might share the first three of these components. The proposed categorization system is considered as a focus of attention in how to search for the prototyped criteria to produce the output expected.

2.4.3 JUNCTION-SEEKING APPROACH (JSA)

This approach was adopted by (Parhami B. et al, 1981). The handwritten Arabic CR system seeks the segments that lead to produce significant fragments of the word. The handwritten Arabic CR system provides features of the pattern points that lie on the junction between two characters and an algorithm to seek the features provided, these features might be: Horizontal stroke-less pattern points, low thickness line, small distance to the baseline. However, the features must be studied precisely to avoid passing a meaningless fragment to the recognition stage. Figure 2.7 shows the flowchart of the JSA program.

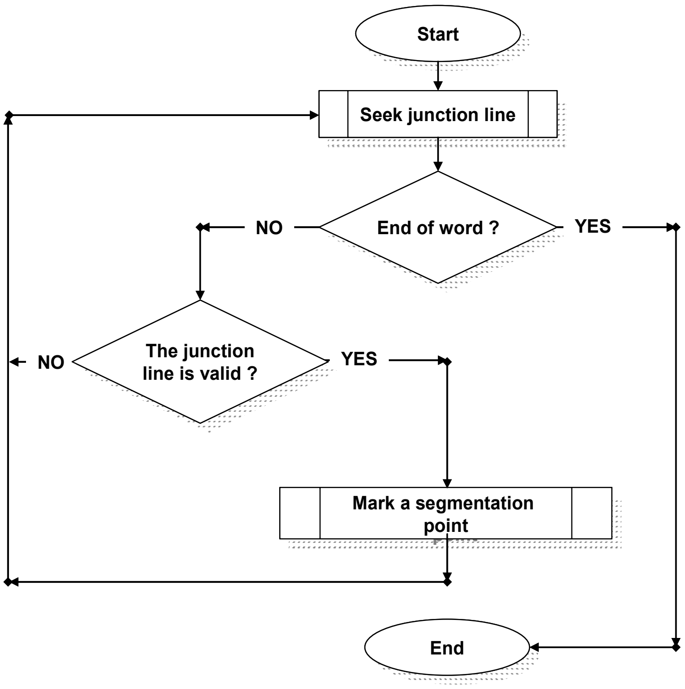


Figure 2.7 Junction-Seeking program

Various researches adopted this approach to perform the segmentation task (as mentioned in next chapter). The variety is latent in the *(Seek junction line)* and the *(Mark Segmentation Point)* functions. The preprocessing of JSA can be Vertical/Horizontal Projection, Image Thinning, or Contour Tracing, while the program output might be meaningful or meaningless fragments. The output depends on the segmentation preprocessing and the features of the junction line provided. These two factors have an impact upon the accuracy of the program.

2.4.4 RECOGNIZE-SEGMENT APPROACH (RSA)

This approach of segmentation was adopted by (Ramses R. et al, 1988). The Recognize-Segment Approach (RSA) seeks any segment that matches the criteria of the characters' shapes. The RSA scans the word from right to left and collects the morphological features. The stops are determined by matching number and nature of the collected features with features that are standardized. Figure 2.8 shows the program flow of the recognize-segment approach.

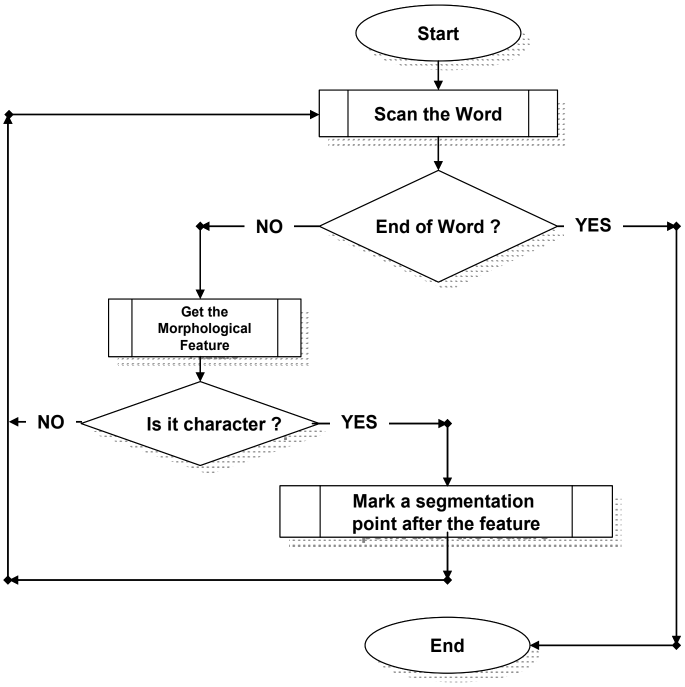

Start
Scan the Word
End of Word ?
NO
YES
Get the Morphological Feature
Is it character ?
NO
YES
Mark a segmentation point after the feature
End

Figure 2.8 Recognize-Segment program

The RSA prototypes the segmentation points for all characters before scanning the image. Translating the feature scanned into a segmentation point is, therefore, a crucial step in the RSA program. After translating the feature, a decision function is performed to decide whether the scanning pointer reaches a segmentation point or not. All the segmentation points are marked to pass the fragments to the recognition process. The scanning is usually performed by moving a window or by utilizing a pre-produced word-image representation to estimate the boundaries of the characters. The preprocessing of the RSA can be: Vertical/Horizontal Projection, Image Thinning, or Contour Tracing, while the program output is a primitive character which may be recognized by the recognition process and may not. In some systems, a feedback is generated by the recognizer and passed to the segmentation processing.

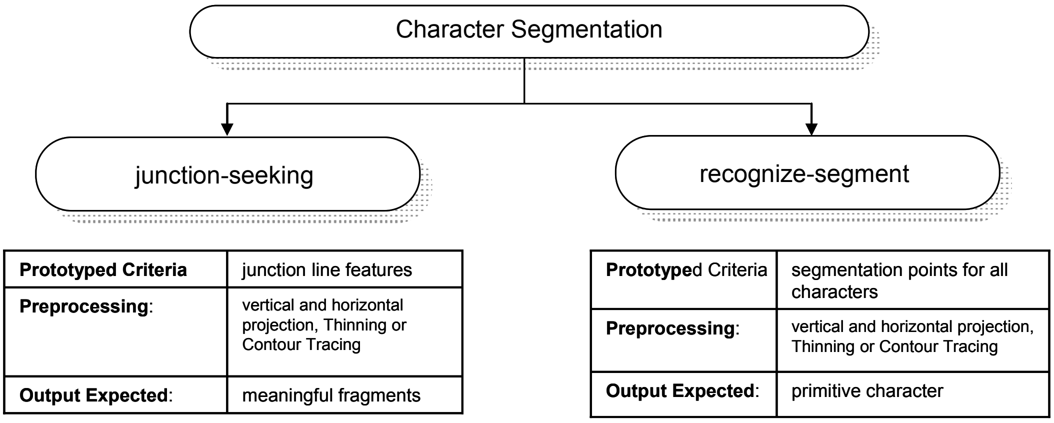
2.5 CHAPTER SUMMARY

Character segmentation process introduces the most serious problem in the development of off-line handwritten Arabic character recognition system. It aims at decomposing the text into lines, words, sub-words, and characters.

The techniques of the preprocessing stage can be divided into two stages according to their ultimate goal:

1. Techniques that produce clean raw data such as normalization.
2. Techniques that prepare the word representation to be segmented such as skeleton extraction.

The segmentation methods can be categorized into two approaches: Junction-Seeking and Recognize-Segment Approach. Each approach includes a predefined and dedicated preprocessing, prototyped criteria, and an expected output as shown in Figure 2.9. This new categorization system defines the similar components of the segmentation methods and focuses on how to find the prototyped criteria in order to produce the output expected.



Prototyped Criteria	junction line features
Preprocessing:	vertical and horizontal projection, Thinning or Contour Tracing
Output Expected:	meaningful fragments

Prototyped Criteria	segmentation points for all characters
Preprocessing:	vertical and horizontal projection, Thinning or Contour Tracing
Output Expected:	primitive character

Figure 2.9 The segmentation approaches

Chapter 3

THE LITERATURE REVIEW

3.1 THE ARABIC CHARACTER SEGMENTATION

A significant progress has been made in Arabic character recognition automation in the last two decades. Since the beginning of 1980s the obstacles of segmenting the Arabic characters was experienced. Different algorithms are published involved a lot of efforts to solve the isolation of the Arabic characters. However, the situation is still far from being matching the ambitions. Table 3.1 summarizes the literature of the Arabic character segmentation in last 20 years. In this chapter, the segmentation methods that have been published by the Arabic characters segmentation community in last 20 years are reviewed and classified as Junction-Seeking Approach or Recognize-Segment Approach. The revision is based on the preprocessing technique used. The knowledge of other languages character segmentation problems increases the understanding of the segmentation issue. Based on that, this chapter looks into the character segmentation of Latin and Chinese script.

Table 3.1 Summary of the literature review

S	Year	Researchers	Writing Mode	Approach	Preprocessing
1	1987	(Almuallim H. et al, 1987)	Handwritten	RSA	1- Skeleton extraction 2- Baseline detection
2	1988	(Sheikh T. et al, 1988)	Printed	JSA	1- Baseline detection 2- Tracing the Contour
3	1989	(Amin A. et al, 1989)	Printed	JSA	V projection
4	1990	(El-Khaly F. et al, 1990)	Printed	JSA	1- Skeleton extraction 2- Baseline detection
5	1992	(Amin A. et al, 1992)	Printed	JSA	1- Skeleton extraction 2- Tracing the skeleton
6	1992	(Goraine H. et al, 1992)	Handwritten	JSA	Skeleton extraction
7	1992	(Margner V., 1992)	Printed	JSA	Tracing the Contour
8	1994	(Altuwaijri M. et al, 1994)	Printed	JSA	1- V/H projection 2- Baseline detection
9	1995	(Ben Amara N. et al, 1995)	Printed	RSA	V/H projection
10	1995	(Romeo-Pakker K. et al, 1995)	Handwritten	JSA & RSA 2 approaches	1- Ver projection 2- Baseline detection
11	1996	(Olivier C. et al, 1996)	Handwritten	JSA	Tracing the Contour
12	1997	(Motawa D. et al, 1997)	Handwritten	JSA	Baseline detection
13	1997	(Bushofa B. et al, 1997)	Printed	RSA	1- baseline detection
14	1999	(Mostafa K. et al, 1999)	Handwritten	RSA	Tracing the Contour

S	Year	Researchers	Writing Mode	Approach	Preprocessing
15	2000	(Fakir M. et al, 2000)	Printed	JSA	H and V projection
16	2001	(Elgammal A. M. et al, 2001)	Printed	RSA	1- Baseline detection 2- H projection
17	2002	(Sari T. et al, 2002)	Handwritten	RSA	1- Tracing the contour 2- Baseline detection
18	2003	(Nawaz S. N. et al, 2003)	Printed	JSA	H and V projection
19	2004	(Zheng L. et al, 2004)	Printed	JSA	1- Baseline detection 2- H and V Projection
20	2005	(Zidouri A. et al, 2005)	Printed	JSA	1- H and V projection 2- baseline detection 3- Skeleton detection
21	2005	(Lorigo L. et al, 2005)	Handwritten	RSA	Optimize the baseline the IFN/ENIT images

3.1.1 JUNCTION-SEEKING APPROACHES (JSA) REVIEW

A. Baseline Detection

Some algorithms were highly dependent on the baseline in searching the word for the junctions. The algorithm in (Sheikh T. et al, 1988) found the baseline and traced the contour of the word to calculate the distance between the extreme points of the intersection of the contour with a vertical line. The junctions must be on the baseline and no strokes above or below them. El-Khaly in his method in (El-Khaly F. et al, 1990) introduced rather difference. The method scanned the skeleton representation for the columns that have no black pixels above or below the baseline. Such adjacent columns formed the junctions and the segmentation points will be in the middle of the junction. The algorithm in (Motawa D. et al, 1997) tried to find these horizontal lines (baseline) after the stage of slant correction. The Singularities were obtained by applying an opening to the word. The Regularities which were the candidates for segmentations were found by subtracting the Singularities. The Regularities were classified to short or long and were examined to find the junctions.

B. Vertical/Horizontal Projection

The reason of adopting the V/H projection is fact that the junction lines are less thickness than other strokes of the word. In (Amin A. et al, 1989), the junctions showed the least sum of the average (AV) in the vertical projection, where:

$$AV = \frac{\sum_{i=1}^{N_c} C_j}{N_c}$$

Where N_c is the number of columns and
X_i is the number of black points of the i^{th} columns.

All the formulas applied are based on examining Arabic characters; they found that the distance between peaks does not exceed 1/3 of the width of Arabic character. The segmentation algorithm of (Altuwaijri M. et al, 1994) scanned the vertical projection representation from right to left to mark the part of the word with a small projection value and allocated near the baseline as a potential junction. Then, these parts are examined to determine the final segmentation points. This method is similar to the one that published by (Fakir M. et al, 2000) whereas the vertical projection was used to separate the words and to isolate the characters. In the segmentation of the characters, a fixed threshold was used in searching for the junctions between characters. In (Nawaz S. N. et al, 2003) the lines were segmented into words by the horizontal projection and divided into three zones: Upper Zone, Middle Zone, and Lower Zone. The vertical projection of the middle zones was created and the word was scanned from right to left, whenever the value of the vertical profile of the middle zone was less than two thirds of the baseline thickness, the area was considered as a junction between two characters. The segmentation process of (Zheng L. et al, 2004) scanned the vertical projection from right to left to find the point at which the histogram value changes from low to high and near the height of the baseline. If such point was found, the scanning continued to find the point at which the value of histogram changes from high to low and near the baseline. If the largest histogram value between the 1st point and the 2nd point was larger than 1.5×height of the baseline the 1st point was marked as a potential segmentation point.

C. Skeleton Extraction

The skeleton extraction has been applied by (Zidouri A. et al, 2005) where the skeleton of the word is scanned for a band of horizontal pixels having length less than or equal to the length of smallest character. A guide was drawn vertically to mark the potential segmentation area. Five features for each band were extracted (width, distance from predecessor, distance from successor, position with respect to the baseline, and midpoint of the band) and tested to select or reject the vertical band as a junction. The skeleton of the word was traced in (Amin A. et al, 1992) using a 3×3 window to describe the features of the word. The Freeman coding scheme is used in the describing. The structure of the word is described by a binary tree. Each node of the tree described the corresponding parts of the sub-word. The tree is divided into several sub-trees to represent each character. One of the rules that were followed in the dividing was searching for the proper junctions. The segmentation process in (Goraine H. et al, 1992) divided the skeleton into principal strokes and secondary strokes. The strokes were classified into three types of strokes: connection, feature, and stroke between two features. A 3×3 window was used in searching the skeleton for a stroke of connection type which was a horizontal string of pixels.

D. Contour Tracing

Other JSAs in the literature include tracing the contour technique. The segmentation method in (Margner V., 1992) was based on the information extracted by tracing the contour of the word. The contour was divided into a series of curves with a distinctive sign. Starting from a particular sign (like positive); whenever the contour sign is changed (from a positive to a negative curvature) a segmentation point was marked to segment a character. Two algorithm levels were forming the method in (Olivier C. et al, 1996). In level 1, useful signals were extracted by using two functions to represent the variations of the image coordinates. Then, the different changes of states were analyzed to detect the local minima and note the Primary Segmentation Points (PSP). The choice of Decisive Segmentation Points (DSP) was done provided that: no loops below the PSP and the line thickness must be smaller than a threshold.

3.1.2 RECOGNIZE-SEGMENT APPROACHES (RSA) REVIEW

A. Baseline Detection

The baseline is very important in feature extraction phase. Bushofa in (Bushofa B. et al, 1997) believed that there is an angle on the baseline to form the joining of any two characters. They proposed an algorithm to search for the occurrences of such angle.

The searching was from right to left over the baseline by using 7×7 window. The standardized matching criterion was an angle that formed the joining of two characters. The central pixel of the window was considered as a candidate segmentation pixel when the matching occurs. A graph-based method was proposed in (Elgammal A. M. et al, 2001). Each text line was represented by a line adjacency graph, and the line adjacency graph nodes represented a dot, diacritic, and an isolated character or sub-word that intersects the baseline. The line adjacency graph of the sub-word was transformed into another homomorphic line adjacency graph containing minimum number of nodes labeled according to their relation with the baseline. The segmentation points were detected by recognizing and computing the horizontal and the vertical gradients in the baseline strip in (Lorigo L. et al, 2005). Any point has more vertical gradients than horizontal and does not cross the upper border of the strip was added to the list of candidates.

B. Vertical/Horizontal Projection

The characters generally form a thick part in the vertical projection representation. The thickness and the length of the characters were the main criteria in the segmentation process. The segmentation process in (Ben Amara N. et al, 1995) started locating the characters and roughly recognizing by keeping only the peaks in the vertical projection representation of the word. The algorithm then proceeded at extraction of the features for each primitive character and an error check was performed for detecting any error. Two methods of segmentation were introduced by (Romeo-Pakker K. et al, 1995). In the first one, the thickness of the stroke in the vertical projection representation was calculated and set as a threshold in the searching for junctions. The second method is the enhancement of the first method. In which, the upper contour was detected to recognize the strokes and determine the location of primary segmentation points.

C. Skeleton Extraction

In character recognition, the morphological information of a word is stored in its skeleton. Based on that, the segmentation algorithms tried to roughly recognize the

Arabic characters in order to isolate them. The word are segmented in (Almuallim H. et al, 1987) after the recognition of the strokes shapes on the skeleton by finding the stroke starting point, and then following the curve to a point which was inferred to be the stroke endpoint. According to geometrical and topological properties, the strokes are combined to form the character.

D. Contour Tracing

The RSAs are also achieved by tracing the outer contour of the word. The algorithm of (Mostafa K. et al, 1999) searched for the potential letter boundaries which are the local maxima points along the lower contour and the local minima points along the upper contour of the words. The potential letter boundaries can be considered as the criteria of stopping the search. Based on the nature of Arabic script, a set of rules applied to eliminate improper potential letter boundaries. The algorithm produced labeled primitives of the word to be processed along with the diacritics and dots. The Arabic character segmentation algorithm in (Sari T. et al, 2002) separated the sub-words by tracing the outer contour. The authors standardized the morphological features from the shapes of the Arabic characters to compare them with the features that extracted from the sub-words image. The algorithm detected the Local Minima (segmentation points) and applied the morphological rules to accept or reject them as Valid Segmentation Points.

3.2 CURRENT SITUATION

The Arabic script can be written by using several font styles such as Ruq'a, Naskh, Dewani, Thuluth, Kufi. Unlike typing by computer, which allows one style of font, the handwritten script usually produces characters that belong to different font styles. This variety of the font styles in the handwritten text increases the difficulty of automating the text recognition.

It has been noted that the readability level of the experimented words or the published databases has not been considered in the literature. The stabilization of the words difficulty level maintains the focus on the main problems of segmentation.

Although the research on the Arabic character recognition started more than 25 years ago, the main difficulty of the segmentation is still situated especially in the handwritten scripts (Zeki A., 2006). The unconstrained nature of the handwritten Arabic scripts requires numerous of experiments to guarantee the efficiency of the algorithm. Several segmentation algorithms have not been tested by numerous database words as in (Almuallim H. et al, 1987), (Sari T. et al, 2002), and (Zidouri A. et al, 2005).

Usually the Arabic people avoid writing the diacritics in their handwriting, but some people write the diacritics in manner bears resemblance to the characters which makes the algorithm tries to segment them. Therefore, it is better to ignore the diacritics. In (Mostafa K. et al, 1999) and (Elgammal A. M. et al, 2001) more processes and classifiers were added to process the diacritics.

Tracing the contour solves the overlapping problem in printed and handwritten scripts. However, the contour must be smoothed first. Tracing the contour process is ideal in extracting the morphological features and in detecting the strokes to recognize roughly the characters and then isolate them. Most RSAs use the technique of tracing the contour like: (Olivier C. et al, 1996), (Mostafa K. et al, 1999), and (Sari T. et al, 2002).

The vertical and horizontal projection techniques are not fully successful in the handwritten Arabic text. It is difficult to detect the baseline of the handwritten words by the horizontal projection and to detect the junctions between the characters by the vertical projection. However, it helps in detecting the junctions of the printed text and has been employed in isolation of the printed characters in (Amin A. et al, 1989), (Altuwaijri M. et al, 1994), (Ben Amara N. et al, 1995), (Fakir M. et al, 2000), (Nawaz S. N. et al, 2003), (Zheng L. et al, 2004), and (Zidouri A. et al, 2005).

In characters "س", "ش", and "ص" there are two junctions and it is difficult to distinct the correct one. This problem appears in JSA, so it is better to verify the isolated characters before passing them to the recognition process as in (Olivier C. et al, 1996). In (Margner V., 1992), the author traced the contour in order to assign a distinctive sign for each curve. The problem of the characters "س", "ش", "ص", "ض", "ن", and "ي" is: when lying at the end of the word, there will be a curve in the terminal valley. This problem causes over-segmentation problem.

3.3 ARABIC CHARACTERS RECOGNITION WITHOUT SEGMENTATION

Due to the difficulties of the segmentation, the research community was trying to avoid the segmentation stage either by recognizing the whole representation of the word or by focusing in their research on the recognition of isolated characters.

The recognition of the whole word (the holistic strategy) recognizes an entire word as a unit. The main drawback of the holistic strategy is that there is usually a predefined lexicon since it is not possible to train all the words in the Arabic language. On the other hand, this drawback makes the holistic strategy suitable for application where the lexicon is limited such as bank check recognition (Richard G. et al, 1996). The strategy of recognizing without segmentation depends heavily on the feature extraction which has been treated elaborately and crossed through successfully in the literature of Latin and Arabic character recognition. Therefore, most of the published researches proved to be powerful in performance and the recognition rate. For example the recognition rate reported by (Amin A. et al, 1997) was 98% and recognition rate reported by R. El-Hajj et. al. was 87%. In most of the holistic and the isolated characters strategies published, the recognition was by using the Neural Network and Hidden Markov Models. Table 3.2 classified some researches according to weather they introduced holistic methods or focused on recognition of isolated character.