Arabic Handwriting Recognition: Challenges and Solutions

Mohamed E. Gumah [a] and Dr. E. Schneider [b]

Department of Information Technology,
University Technology PETRONAS
E-mail:[1] alrjele2004@yahoo.com
[2] etienne@ira.uka.de

ABSTRACT

Optical Characters Recognition (OCR) is one of the active subjects of research since the early days of computer science. Even if Arabic characters are used by more than a half a billion people; Arabic characters recognition has not received enough interests by the researchers. Little research progress has been achieved comparing to what has been done with Latin and Chinese. The cursive nature of the Arabic characters makes it more difficult to achieve a high accuracy in character recognition since even printed Arabic characters are in cursive form. This paper presents the main challenges (difficulties) researchers are facing and up to dated solutions (the common methods) are used for Arabic text recognition.

1.0 INTRODUCTION

Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input. It also can be defined as the task of transforming text represented in the spatial form of graphical marks into its symbolic representation. It can be used in many fields such as automatic mail sorting or cheque processing. In automated mail sorting, letters are directed to the correct location by recognition of the handwritten address. Similarly, cheque processing involves recognizing the words making up the cheque amount (Burrow, 2004).

There are two approaches in handwriting recognition systems; on-line and off-line systems. With on-line system, the machine recognizes the symbols as they are drawn, and there is no need for contour extraction. This property makes the recognition stage of the on-line systems easier than it is in the off-line systems. With off-line systems, input text is read and digitized by an optical scanner. Then, each character will be located and segmented. The resulting array is fed into a pre-processor for smoothing, elimination of noise, size normalization and other operations, to facilitate the extraction of features in the subsequent stage (Zeki &Zakaria, 2004). There are two approaches to deal with input image; segmentation-based methods and free segmentation methods. In segmentation-based method the word will be split into segments to be recognized, using segmentation algorithm. In contrast, segmentation-free method use features of the whole word image. A new structure of off line OCR system by using the technique similar to that is used in wavelet compression. It benefited from the wavelet image compression wavelet image compression 40x40 bitmap images as input to produce a decomposition vector for each character. The result was considerably high in terms of accuracy and recognition rate (Reaches 97.9% for some letters at average 80%) (Aburas & Rehiel, 2007).

.2 Main steps in OCR systems

There are three main steps which must be done in order to achieve any optical characters recognition:

A. Pre-processing:

This step should cover all those functions carried out prior to feature extraction to produce a cleaned up version of the original image so that it can be used directly and efficiently by the feature extraction components of the Optical Characters Recognition (OCR) system.

B. Features extraction:

According to (Lippman,1989), "Features should contain information required to distinguish between classes, be insensitive to irrelevant variability in the input, and also be limited in number to permit efficient computation of discriminate functions and to limit the amount of training data required". This step involves measuring the features of the input character which are related to classification step. After features extraction, the character will be represented by a set of extracted features (Ben Amor &Essoukri, 2005).

C. Classification:

Classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (Wikipedia) in this stage, algorithms should be used for pattern recognition such as Quadratic classifiers, k-nearest neighbor and Boosting

1.3 Nature of Arabic Characters

Arabic characters are used in several languages, like Arabic, Persian, Urdu, Jawi and Pishtu, (more than a half of a billion people), in addition to that, most of Muslims (almost ¼ of the people on Earth) can read Arabic because it is the language of Al-Quran, the holy book of Muslims. Even though, Arabic characters recognition has not received enough interests by the researchers. Little research progress have been achieved comparing to Latin and Chinese. The main difficulty in Arabic character recognition is due to the cursive nature of Arabic writing which doesn't allow direct application of many algorithms designed for other languages. Arabic characters are connected on an imaginary line called baseline (Figure). Unavailability of sources such as Arabic text database also makes it more difficult to develop Arabic characters recognition systems (Zakaria &Zeki , 2004).

Figure 1. The baseline

1.3.1 Arabic character features

Arabic writing like English in term of using letters, numbers, and punctuation but there are number of characteristics which make Arabic cursive writing is unique compared to Latin, Chinese and Japanese. These characteristics explain why there are not enough researches , which are done **concerning the** Arabic writing recognition. These characteristics can be summarized as follow:

- Arabic is written from right to left in both printed and handwritten forms (Figure1). No upper or lower case exists in Arabic.
- Arabic is always written cursively and words are separated by spaces. Only six characters can be connected from the right, these are: ?, ?, ?, ?, Q, ?(Table 1).
- The shape of the character is different according to its position in the word. Each character has either two or four different forms. This will increase the number of classes to be recognized from 28 to 84.
- Fifteen characters have dots with the character. They can be above or below the primary part (Table 1).
- Some characters share the same primary shape. Only dots make difference between them. (Table 1).
- If the following six characters (?, ?, ?, ?, Q, ?), appearing in a word, will cause the word to be divided into two or more sub-words separated by spaces, but usually shorter than the space between words. This must be considered to avoid segmenting a word into two words. Examples of words in which all characters are connected: ???, ?????, ??? .

  Examples of words consist of sub-words: ??????, ???? , ????
- Some characters contain closed loop (Table 1). Loop is an important feature to describe a character. Character ? contains two loops. The open portion of characters ?, ? and ? sometimes, if written by hand, is closed to form a triangle. The loop of character ?, ? and ? sometimes becomes too small that the internal opening part is disappeared.

Table 1: Printed Arabic characters style and shapes

No	Name	Isolated	Connected		
			Beginning	Middle	End
1	**Alif**	?	??	?-	??
2	**Baa**	?	??	???	? ?
3	**Taa**	?	??	??	? ?
	Thaa	?	??	???	? ?
	Jeem	?	??	???	??
	Haa	?	?S	???	??
	Khaa	?	??	???	F?
	Daal	O	O	??	??
	Thaal	?	?	??	??
	Raa	?	?	??	??
11	**Zaay**	?	?	??	??
12	**Seen**	a	??	?d?	ß ?
13	**Sheen**	e	??	???	? ?
14	**Saad**	?	??	?µ ?	? ?
15	**Shaad**	?	??	?þ ?	? ?
	Ttaa	?	?s	?t ?	??
17	**Dthaa**	?	??	???	f ?
18	**Ain**	?	??	???	??
19	**Ghen**	?	??	???	??
20	**Faa**	?	??	???	? ?
21	**Qaf**	?	??	???	??
	Kaf	?	??	???	? ?
23	**Lam**	?	??	???	??
24	**Mem**	?	??	???	??
25	**Noon**	?	??	???	??
26	**Haa**	?	??	???	??
27	**Wow**	?	?	??	??
28	**Yaa**	?	??	???	??

2.0 PERVIOUS RELATED RESEARCH WORK

Beside the main goal of any OCR system which is simulating human's reading capability, the accuracy and time consuming are very important issues in this aspect. Based on the latest survey which is published in December 2007 by A. Aburas and S. Reheil all covered papers presented their proposals seeking high accuracy and less time. Each one has treated the issue from different angle of view. Recalling the typical construction of OCR system and some characteristics of Arabic letters that are mentioned above we can classify their work into three main categories: preprocessing problems, features

extraction problems, and recognition (discrimination) problems. Many researchers have used skeletonization in their proposed preprocessing stages like for instant (Sari,Souici & Sellami , 2002 ; Pavlidis , 1981 ; Rosenfeld & Kak , 1982 ; Amir,Karim & Abolfazl , 2002 ; Amin & Al-Sadoun , 1994; Jang, Chin B.K. & R.T. Chin , 1992; Cheung & Bennamoun , 1998). Another method which is known as "contour" is the Freeman chain code of the character's border. In this method chain code stores the absolute position of the first pixel and the relative positions of successive pixels along the character's border. However, skeletonization suffers from mislocalization of feature and ambiguities particular to each thinning algorithm whereas the contour approach avoids these problems since no shape information is lost. For the features extraction stage, where the main role is played and mostly the accuracy of recognition depends on the information passed from this stage to the classifier (recognizer). These information can be structural features such as loops, branch-points, endpoints, and dots or statistical which includes but is not limited to, pixel densities, histograms of chain code directions, moments, and Fourier descriptors. Histogram of slopes along contour is used. Fourier descriptors have also been used in this stage. Artificial Neural Networks (ANNs) are the common seed of most if not all classifiers "recognition or discrimination stage". Many variations have been used in order to overcome the main disadvantage of ANNs which is time consuming. Parallel system is designed for back-propagation ANN approach in order to accelerate the computation process (Abdurazzag &Salem, 2007).

3.0 DIFFICULTIES IN RECOGNIZING ARABIC SCRIP TEXT

The written form of the Arabic language presents many challenges to the OCR developer that can be summarized as following:

3.1 Letter connectivity

In Arabic text, a script alphabet in which consecutive letters within a word are joined to one another by a baseline is used. To accommodate the baseline, the Arabic alphabet has four forms of each letter: an isolated form, an initial form, a medial form, and a final form. (Table 1) Several letters in the alphabet do not follow this rule and have different shapes for the medial and final forms. When one of these non-joining characters is involved in a word, the previous letter assumes its final (or isolated) form, and the non-joiner assumes its initial (or isolated) form. This feature of connectivity will cause a level of difficulty in the segmentation stage. In contrast of other languages, printed and handwritten Arabic text is always cursive. Thus, the rate of recognition of Arabic characters is lower compared to those languages.

3.2 Position-dependent letter shaping

Most Arabic letters contain dots in addition to the letter body, such as e which consists of letter body and three dots above it as illustrated in Table 1. In addition to dots, there are strokes that can be attached to a letter body creating new letters such as ?. Some of the dots and strokes might be missed in the Pre-processing stage where the text should be cleaned up to be used directly and efficiently by the feature extraction components of the OCR.

3.3 Different writing style

In general Arabic writing style can be classified into typewritten (Nuskah), handwritten (Ruqqah) and few others usually used for decorative calligraphy such as (Kofi, Thuluth and Diwani). Except professionals, normal people do not follow the handwritten style rules. That will cause more difficulties for recognition and make the database of the system even larger.

4.0 OCR AND ARABIC TEXT

Since the mid-40s many papers ware published on character recognition. In early 1950s, first commercial machine was introduced. In the early systems, the recognition logic was mainly on hardware technique rather than software. Only in early 70s, systems utilize software as recognition logic ware developed such as OCR-A and OCR-B. Nowadays, available systems can recognize different writing styles like handwritten, printed and character sets for Latin, Japanese, Chinese, Korean and Arabic (Stallings , 1976).

The first work in Arabic character recognition was written by Nazif in 1975 who developed a printed Arabic character recognition system based on stroke extraction. Several papers were published about the printed Arabic characters recognition. In (Lam,Lee,S.W & Suen; Lippmann, 1989)dealt with handwritten Arabic characters. However, there is a lack in Arabic character recognition researches compared with those for other languages. That lack might be caused be the following reasons:

(a) The specialty of Arabic characteristics.
(b) The shortage of relevant journals and books.

(c) The lack of conferences which lead to lack of interaction between researchers with the same interests.

Recently, OCR system for Arabic characters started to appear commercially. For example "TextPert Arabic " which was advertised by Connecticut Technology Associates(CTA) as a printed text reader that runs on Apple Macintosh with maximum recognition rate of 99%. IQRAA, which was developed by Arab Scientific Software and Engineering Technology runs on IBM-PC and can recognize several fonts with recognition rate of 99% and speed of 1500 characters per minute (Mohamed , 2005)

5.0 MAIN METHODS AND TECHNIQUES

For preprocessing step, the most popular methods followed are:

5.1 Binarization

This process is to convert a gray scale image into bi-level image by computing the histogram of the gray values of the image and then finding a cut-off point. Then we consider the values which are darker than cut-off point as 1 and the points which is lighter as 0.

5.2 Filtering and smoothing

In this process, unwanted variations should be removed from the input image. Some researchers have studied a conditioning step for removing stray ending strokes joining two strokes if they are so close to each other. Others suggest using a 3x3 window to traverse pixel-by-pixel 2-D image as a method for noise reduction (Al-Badr & Haralick, 1994).

5.3 Thinning

Thinning is the process of minimizing the width of a line, in the input image, from many pixels wide to only one pixel (Lam,Lee,S.W.&Suen). In this process, two kind of algorithm are used, sequential algorithm, and parallel algorithm, the main difference between these two types is that sequential algorithm operates on one pixel at a time where the operation is depending on preceding processed result, while the parallel algorithm operates on all pixels simultaneously. The most common thinning algorithm is based on an edge erosion technique where a window is moved over the image a set of rules applied to the contents of the window.

For the segmentation step, which is a crucial step in Arabic text especially handwriting, horizontal projection histogram is the most used technique for text image. With this technique, each gap (minimum value) corresponds to a line-break in the text, but this technique is not suitable in cases where characters are overlapping especially in Arabic handwriting. To solve this problem, the different sub-words should be identified by tracing their contours and then shift them apart and inserts a blank column between them (Tolba & Sheddad,1990).

For word segmentation, most of the researches have been done for Latin handwriting. Although some algorithms designed for Latin cursive word segmentation might be used for Arabic handwriting, they are not adequate for that task (Zeki & Zakaria, 2004; Lam,Lee,S.W & Suen) that's why researchers trying to modify their segmentation algorithms to be more suitable for Arabic handwriting.

In feature extraction step, where we use the character being produced in segmentation stage is to extract some features for the classification stage, some researchers use template matching between the template of the radicals and character image, others match the histograms of the input characters to those of the templates. For classification stage, where the features extracted from the primitive is compared to those of the model set, three methods are used, structural, statistical and using of neural networks(Mohamed,2005). In general, there are six main systems approaches have been proposed for Arabic text recognition which are all off-line recognition systems :

1) Based on pre-segmented characters
2) Based on segmenting words into primitives
3) Based on segmenting words into characters
4) Based on recognition of words prior to segmentation
5) Based on recognition without segmentation (Mohamed,2005)
6) Based on Image Compression. No preprocessing (no feature extraction and ANN classifications) New approach (Aburas & Rehiel, 2007).

6.0 CONCLUSION

The typical Optical Character Recognition (OCR) systems, regardless the character's nature, are depending on three stages, preprocessing, features extraction and classification. Each stage has its own problems and effects on the system efficiency which is the time consuming and the recognition errors.

In this paper, an overview is presented about Arabic character characteristics as introduction to discuss the main difficulties that face researchers whom are working in Arabic text recognition field such as letter connectivity, position-dependent letter shaping and different writing style. Fully covered and discussed existing solutions and approaches that are used to deal with Arabic text in optical recognition systems. Since none of the solution is accurate nor fast. New and intelligent system of building OCR for Arabic recognition is need. This may lead to introduce the HMM as part of the recognition engine.

REFERENCES

Abdurazzag A.A and Salem A.R., Comprehensive Review for Arabic Handwriting and Printed Characters Recognition, *International Conference on Intelligent and Advanced Systems ICIAS2007, 25-28November 2007, Kuala Lumpur, Malaysia*

Aburas, A A. and Rehiel, S M A. Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression. *ARISER* Vol. 3 No. 4 (2007) 123-135.

Al-Badr, B. & Haralick, R. 1994. Symbol recognition without prior segmentation. *Proc. IS&SPIE symp. On electronic imaging Sc*. 2181.pp.303-314

Amin, A. Al-Sadoun, H.B., "Hand printed Arabic character recognition system", Pattern Recognition, 1994., Volume: 2 , page(s): 536 - 539 vol.2

Amir Mowlaei, Karim Faes and Abolfazl T. " Feature Extraction with Wavelet Transform for Recognition of Isolated Handwritten Farsi/Arabic Characters and Numerals" Digital Signal Processing, 2002. On page(s): 923 - 926 vol.2, Digital Object Identifier: 10.1109/ICDSP.2002.1028240

Ben Amor N. & Essoukri N. (2005) Multifont Arabic Character Recognition Using HoughTransform and Hidden Markov Models. Proc. *4th International Symposium on Image and Signal Processing and Analysis , 285-288, 2005*.

Burrow, P. Arabic Handwriting Recognition. Master of Science School of Informatics.*University of Edinburgh.2004.*

Cheung, A. Bennamoun, M. Bergmann, N.W, "A recognition-based Arabic optical character recognition system", Systems, Man, and Cybernetics, 1998. Page: 4189 - 4194 vol.5

Jang & Chin Jang, B.K. and R.T. Chin "One-pass Parallel Thinning: Analysis, Properties, and Quantitative Evaluation", IEEE Trans. *Pattern Analysis and Machine Intelligence, Vol. 14, No. 11, pp. 1129-1140,1992*.

Lam, L. Lee, S.W. & Suen, C.Y. Thinning methodologies-A comprehensive survey, *IEEE Trans. On Pattern Analysis Machine Intelligence*. 146:89-885.

Mohamed A. Ali 2005. Arabic Handwriting Recognition. *University Kebangsaan Malaysia*.

Pavlidis T. "Algorithms for graphics and image processing", *Murray Hill ed., New Jersey, Sep. 1981*.

R. Lippmann, "Pattern Classification using Neural Networks", *IEEE Communications Magazine, p. 48, November 1989.*

Rosenfeld A. and A.C. Kak, "Digital image processing",*2nd Edn. pp. 347-349, Addison Wesley, London 1982.*

Sari, T. Souici, L. Sellami, M. "Off-line handwritten Arabic character segmentation algorithm: ACSA", Frontiers in Handwriting Recognition, 2002., page(s): 452 – 457

Stallings, W. 1976. Approaches to Chinese character recognition. *Pattern Recognition* 8:87-98.

Tolba, M. F. & Shaddad, E. 1990. On the automatic reading of printed Arabic characters, *Proc. IEEE Inter. Conf. on systems*, pp. 496-498.

Wikipedia website 1/12/2007, http://en.wikipedia.org/wiki/Statistical_classification. Accessed January 2008

Zeki, A. M. and Zakaria, M. S. Challenges in Recognizing Arabic Characters. *The national conference for computer. Abu-al-Aziz king University. Arabia Saudi .April 2004*