

A Recognition-Based Arabic Optical Character Recognition System

A. Cheung
Space Centre for Satellite
Navigation
Queensland University of
Technology
GPO Box 2434, Brisbane,
Qld 4001, Australia

M. Bennamoun
Space Centre for Satellite
Navigation
Queensland University of
Technology
GPO Box 2434, Brisbane,
Qld 4001, Australia

N. W. Bergmann
Space Centre for Satellite
Navigation
Queensland University of
Technology
GPO Box 2434, Brisbane,
Qld 4001, Australia

ABSTRACT

Optical character recognition systems improve human-machine interaction and are widely used in many government and commercial departments. After forty years of intensive research, OCR systems for most scripts are well developed. However, not for Arabic script. Since Arabic is a popular script, Arabic OCR systems should have great commercial value. Thus a recognition-based Arabic OCR system is proposed in this paper. It consists of the image acquisition, preprocessing, segmentation, character fragmentation, combination of character fragments, feature extraction, and classification. A signal is fed back to improve and determine the segmentation/recognition result. The system has been implemented and it has 90% recognition accuracy with a 20 chars/sec recognition rate.

1. INTRODUCTION

Optical character recognition (OCR) is the process of converting a raster image representation of a document, e.g. a machine printed or handwritten text scanned by a document scanner, into a computer processable format, such as ASCII code.

The origin of character recognition system was found in 1870 as an aid for the visually handicapped. In the 1940's, digital computers were invented and since then many engineers and scientists have started their research on OCRs. In the 1950's, the first commercial OCR system was available [1,2]. This subject has attracted an immense research interest not only because of the very challenging nature of this problem, but also because it improves human-machine interaction in many applications. Example appliances include office automation, cheque verification, and a large variety of banking, business and data entry applications. Thus, after forty years of intensive research, a lot of techniques and methods were developed for many scripts. Moreover, many OCR systems are commercially available nowadays.

The typed Latin, Chinese and Japanese scripts are widely used around the world. Their characters are separated from one another which makes their OCR techniques easier to develop. These are the reasons why OCR systems for these three scripts are well developed and most commercial available OCR systems

recognize either of these three scripts. Arabic is another popular script. It is estimated that there are over one billion Arabic script users. However, because of the technical difficulties induced by the cursive nature of the Arabic script, its OCR techniques have not been well developed yet. If OCR systems are available for Arabic characters, they will be very useful and have a great commercial value. Therefore, a recognition-based Arabic Optical Character Recognition system is proposed in this paper. Some background knowledge is given in Section 2 first. Then the detail structure of the proposed method is described in Section 3. The system performance and discussions are then presented in Section 4. Finally, this paper is concluded in Section 5.

2. BACKGROUND

2.1 The Characteristics of Arabic Script

This section illustrates some problems which are faced when developing an Arabic OCR system.

As each Arabic character has two to four different forms, this extends the classes to be recognized from 28 to 100. Fig 1 shows the character set of Arabic script which clearly illustrates that the appearance of Arabic character varies according to its position in a word or sub-word [3, 4].

Both typed and hand-written Arabic are cursive and are read from right to left. Fig 2 demonstrates the formation of an Arabic word and illustrates the variation of Arabic characters' shape in a word. Due to the cursive nature of the script, we can either recognize a word at a time or segment a word into characters and then recognize the characters. The first case seems to be impossible and not feasible due to the numerous numbers of words in a language. However, if the second case is used, research has been practically proved that the segmentation of a cursive word is a very difficult problem. However the segmentation is a crucial step in Arabic OCR systems [5].

We have also noticed that some Arabic words may be horizontally overlapped with others in a document. An example is given in Fig 3. This feature causes the traditional

segmentation method using projection profile not applicable in this situation and it brings out the word segmentation problem.

Name	EF	MF	BF	IF	Name	EF	MF	BF	IF
DĀD	ض	ض	ض	ض	ALIF	ا			أ
TĀ	ط	ط	ط	ط	BĀ	ب	ب	ب	ب
ZĀ	ظ	ظ	ظ	ظ	TĀ	ت	ت	ت	ت
'AYN	ع	ع	ع	ع	THĀ	ث	ث	ث	ث
GRAYN	غ	غ	غ	غ	JĪM	ج	ج	ج	ج
FĀ	ف	ف	ف	ف	WĀ	ح	ح	ح	ح
QĀF	ق	ق	ق	ق	KHĀ	خ	خ	خ	خ
KĀF	ك	ك	ك	ك	DĀL	د			د
LĀM	ل	ل	ل	ل	DHĀL	ذ			ذ
MĪM	م	م	م	م	RĀ	ر			ر
NŪN	ن	ن	ن	ن	ZĀY	ز			ز
HĀ	ه	ه	ه	ه	SĪN	س	س	س	س
WĀW	و			و	SHĪN	ش	ش	ش	ش
YĀ	ي	ي	ي	ي	ṢĀD	ص	ص	ص	ص

Fig 1. Arabic character in all forms. (EF end form, MF middle form, BF beginning form, and IF isolated form.)

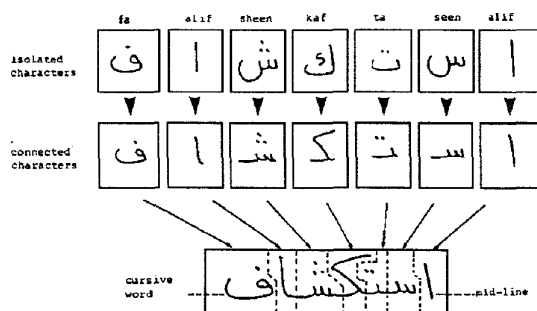


Fig 2. An Arabic word.

Some other characteristics of Arabic script are summarized below.

- Most characters (17 out of 28) have a dot, two dots, or zigzags associated with the character and they are located either above, below, or inside the character.

- Most characters share similar shape with others, e.g. BĀ, TĀ and THĀ; JĪM, HĀ and KHĀ, etc. The position or number of dots in the character makes the only difference.
- Some characters can only appear at the beginning or at the end of a word or sub-word. An Arabic word could have one or more sub-words. This is due to the fact that some characters are not connectable from the left side with the succeeding character.
- There are only three zigzags that represent vowels. Other vowels are represented by diacritics in form of over-scores or under-scores. The use of diacritics is limited to the cases where the word is foreign or where the pronunciation is stressed.
- There are no upper or lower cases in Arabic.



Fig 3. An example of overlapped Arabic words.

2.2 Dissection vs. Recognition-Based Segmentation

The segmentation of an object can be performed by dissection or recognition-based methods. Dissection is meant the decomposition of the image into a sequence of sub-images using general features [6]. It is an intelligence process in that an analysis of the image is carried out. For OCR systems using this technique, they usually plot projection profiles of the image and then use a set of rules to segment the image. The dissection technique is widely used by Latin, Chinese and Japanese OCR systems. It is because characters of these scripts are isolated, hence the character segmentation can be easily achieved by dissection techniques. Although Amin [7] developed a dissection technique for Arabic characters, it seems to be font dependent.

On the other hand, no feature-based dissection algorithm is employed in the recognition-based segmentation technique. It usually uses a mobile window of variable width to provide a sequence of tentative segmentations which are then confirmed (or not) by the character recognition as a result of a coherent segmentation/classification result [6]. This technique is also called "segmentation-free" in other literatures. The major advantage of this technique is that it bypasses the segmentation problem. Therefore it should be suitable to systems which involved serious segmentation problem.

3. THE ARABIC OCR SYSTEM

The implemented Arabic OCR system involves five image processing techniques which are the image acquisition, the preprocessing, the segmentation, the feature extraction and the classification. As the recognition-based technique is employed in the system, the feature extraction and classification are

grouped into one block. Fig 4 gives an overview of the proposed system.

A document is quantized by a flatbed scanner in space and amplitude (i.e. image sampling and gray-level quantization) to acquire a digitized representation. The digital image is then binarized by the Otus method described in [8]. A simple smoothing method is used to minimize the noise in the image due to the shading effect or unevenness of the gray scale [9]. The image is then ready for segmentation. The projection profile method is employed to extract lines from the document. As mentioned earlier in Section 2.1, Arabic words may horizontally overlap with others, therefore a word segmentation method is developed to solve this problem. The algorithm is described in [10].

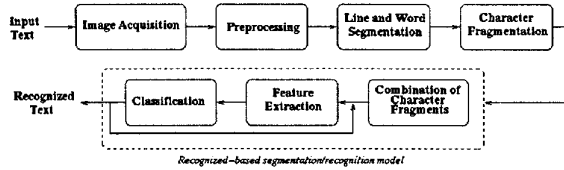


Fig 4. The recognition-based Arabic OCR system.

3.1 Character Fragmentation

The input to the recognition-based OCR system is a sequence of tentative character fragments. It can be done by either the pixel-based or feature-based fragmentation. In order to save the processing time of the system, the feature-based fragmentation is chosen. It involves two steps.

The first step provides coarse fragmentation points. We simplified the dissection technique of Amin [7] by ignoring all the supplementary segmentation rules. In more detail, we plotted the vertical projection profile of a word and calculated the sum of the average value (AV), where

$$AV = (1/N_c) \sum_{i=1}^{N_c} X_i \quad (1)$$

and where N_c is the number of columns and X_i is the number of black pixels of the i th column. Hence each part which shows a sum value less than AV is a tentative fragmentation point [7].

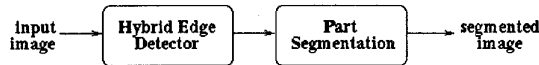


Fig 5. Bennamoun's segmentation technique.

In the second step, we fine tuned fragmentation points by applying the object segmentation method of Bennamoun's vision system [11]. His segmentation method has been practically proved to be a reliably technique for segmenting objects with convex dominant points (CDPs). Fig 5 illustrates this segmentation technique which consists of two stages: the hybrid edge detection and the part segmentation.

Firstly, we used the hybrid edge detector, whose structure is shown in Fig 6, to detect the edges of a word. A hybrid edge detector is used because it can localize good edges and provide good immunity to noise simultaneously. We then extracted the contour of the word and fed it into the part segmentation stage.

We detected CDPs of a word in the part segmentation stage. The algorithm used in the extraction of the CDPs is illustrated in Fig 7. At first, the contour smoothing operation is carried out using a Gaussian kernel with σ_1 so that the problem of discontinuity in the calculation of the derivative of curvature can be avoided. Once a smoothed contour is produced, the curvature is computed using Equation (2).

$$K_s(t) = \frac{\dot{\hat{x}} \ddot{\hat{y}} - \dot{\hat{y}} \ddot{\hat{x}}}{\left(\dot{\hat{x}}^2 + \dot{\hat{y}}^2 \right)^{3/2}} \quad (2)$$

where $\dot{\hat{x}} = \frac{d\hat{x}}{dt}$, $\dot{\hat{y}} = \frac{d\hat{y}}{dt}$, $\ddot{\hat{x}} = \frac{d^2\hat{x}}{dt^2}$, $\ddot{\hat{y}} = \frac{d^2\hat{y}}{dt^2}$, and

\hat{x} and \hat{y} denote the smoothed version of the x and y coordinates of the contour respectively. The uppermost branch of the block diagram shown in Fig 7 extracts all the dominant points on the contour by convolving the curvature with the derivative of the Gaussian function with σ_2 and followed by zero crossing detection. A dominant point is defined as the point for which the derivative of the curvature equals zero. The lowermost branch is responsible to select the convex points for which the smoothed curvature is greater than a certain threshold Th . Both branches are ANDed to produce the CDPs and each CDP is a tentative fragmentation point.

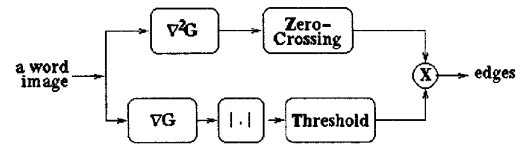


Fig 6. The hybrid edge detector.

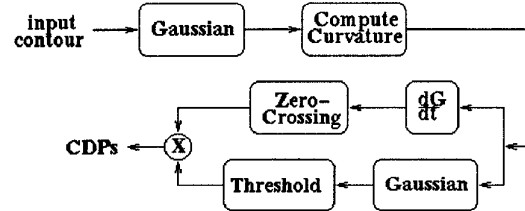


Fig 7. Extraction of the CDPs.

3.2 Feature Extraction

نفس

By using the hybrid edge detector, the contour of a character fragment is extracted. Then, we started from the top right-hand black pixel of the character fragment contour and traced through its whole contour. The tracing process used depends on a 2x2 window. When this window is imposed over a contour, it produces a vector such as those in Table 1. This feature extraction process is similar to the one described in [12]. However, as the input image is different, some modifications to the method have been made. The result of this process is a sequence of Freeman codes. Fig 9 shows the contour of the character ALIF. By applying this feature extraction method, the following sequence of Freeman code is produced:

We then apply the following four formulae to smooth up the code chain.

$$C_i C_j C_i \rightarrow C_i C_i C_j \quad (3)$$

$$C_i C_j C_l \rightarrow C_i C_l C_j \quad (4)$$

$$C_i C_j C_k \rightarrow C_i C_j C_k \quad (5)$$



















$$C_i C_j C_k \rightarrow C_l C_l C_l \quad (6)$$

where

$$C_i, C_j, C_k, C_l \in \{1, 2, 3, 4, 5, 6, 7, 8\}$$

and C_l is the resultant direction of C_i , C_j , and C_k . By applying the above formulae, the above listed sequence becomes:

7,7,1.1.2.2.3.3.3.3.

Mask 2x2	No.	Type	Freeman code
	1		
	2		
	3	←	5
	4	→	1
	5	↑	7
	6	↓	3
	7	↘	2
	8	↙	6
	9	↗	8
	10	↖	4
	11		
	12		
	13		
	14		
	15	↖ (after 3 or 10)	6
	15	↘ (after 4 or 9)	2
	16	↗ (after 5 or 9)	8
	16	↖ (after 6 or 7)	4

The code chain is finally concentrated by dividing the run-length of a code with a threshold T_1 providing that the run-length of that code exceeds a threshold T_2 . The purpose of T_2 is to make the final code chain have a certain degree of robustness to noise. If T_1 is set to 8 and T_2 is set to 3 then the above code chain becomes the following sequence of codes:

3,3,5,7,7,2.

3.3 Classification

The classification process is carried out at the final stage to recognize the character. It assigns an input character to one of many pre-specified classes which are based on the extracted features and their analysis.

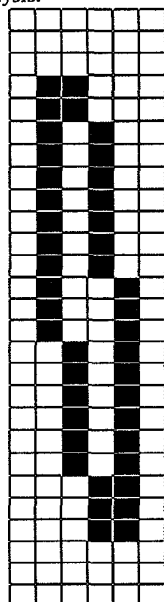


Fig 9. The contour of ALIF.

In this OCR system, each character fragment is numbered from right to left. During the recognition process, the first fragment is fed into the feature extraction process in order to determine the concentrated Freeman code chain. This code chain is then inputted to a structural classifier to find the best match. The structural classifier is a state-diagram of Freeman codes of database samples. In order to minimize the confusion of character fragments with characters and to save the search time, there are four databases. According to the position of a fragment or fragments in a word, we will go to the corresponding database to search for the best match. For example, if there is a combination of first and second fragment, we will go to the database file for beginning characters. If the fragment could not be recognized, a signal is fed back to the character fragment combination process to combine the first and second fragment (refer to Fig 4). Then the above processes are repeated until a character is recognized. If a character is recognized after combining the first n fragments, then this feedback loop will start again to recognize the next character from the $(n+1)$ th fragment onwards.

The above feedback system has a potential problem which is if a character in a word could not be recognized due to some reasons, then the rest of the characters would not be recognized properly. In order to minimize this problem, we repeated the above feedback recognition process again but started from left to right if the word is not wholly recognized in the right-to-left

feedback loop. After that, we combined the result of these two feedback trials to form the recognized word.

4. RESULTS AND DISCUSSIONS

We have fully implemented the recognition-based Arabic OCR system that is described above. The system is written using C/C++ programming language and is run on Pentium 166MHz personal computer. It was applied to Arabic documents which means all four forms, as shown in Fig 1, are mixed together in testing samples. Many tests were taken on printed texts and a recognition accuracy of 90% was achieved. The worst result is shown in Fig 10. It recognizes Arabic characters in around 20 char/sec. In other words, it is a real time system.

The major error of this system happens in the classification stage. Even though we have performed right-to-left and left-to-right feedback recognition, whenever there is a character in a word that could not be recognized, the rest of the characters in the word are not recognized properly. It seems that it could not be solved unless a more complex feedback control strategy is used.

If we compare the recognition accuracy of this system with the other two systems that have been described in [5, 13], it is obvious that the recognition accuracy has increased. This is because of the use of the recognition-based segmentation/classification method. Therefore we believe that recognition-based model is more suitable to the Arabic script or other cursive scripts, like handwritten Latin.

5. CONCLUSION

We have presented a recognition-based approach for the recognition of printed Arabic text in this paper. This system consists of the image acquisition, preprocessing, segmentation, feature extraction and classification. It is similar to usual OCR systems except it has a feedback loop that can control the combination of character fragments to form a character for classification. Because of this feedback loop, the system bypasses the character segmentation process which leads to the 90% recognition accuracy. Its recognition rate is about 20 chars/sec.

As mentioned earlier, the feedback loop that we used has a potential problem. If a character in a word could not be recognized, the rest of the characters are not recognized properly. This affects the accuracy of this system. However, we believe that if a more intelligent feedback loop is developed on controlling the combination of character fragments to form characters, a higher recognition accuracy should definitely be achieved.

6. REFERENCES

- [1] V. K. Govindan and A. P. Shivaprasad, "Character Recognition - Review," *Pattern Recognition*, vol. 23, no. 7, pp. 671-683, 1990.

- [2] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical Review of OCR Research and Development," in *Proceedings of IEEE*, vol. 80, pp. 1029-1058, July 1992.
- [3] I. S. Abuhaiba, S. A. Mahmoud and R. J. Green, "Cluster Number Estimation and Skeleton Refining Algorithm for Arabic Characters," *The Arabian Journal for Science and Engineering*, vol. 16, no. 4B, pp. 519-530, October 1991.
- [4] K. M. Jambi, "Arabic Character Recognition: Many Approaches and One Decade," *The Arabian Journal for Science and Engineering*, vol. 16, no. 4B, pp. 501-509, October 1991.
- [5] A. Cheung, M. Bennamoun and N. W. Bergmann, "Implementation of A Statistical Based Arabic Character Recognition System," (Brisbane, Australia), TENCON'97, pp. 531-534, December 1997.
- [6] R. G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," *IEEE Trans. on PAMI*, vol. 18, no. 7, pp. 690-706, July 1996.
- [7] A. Amin, "Recognition of Arabic Handprinted Mathematical Formulae," *The Arabian Journal for Science and Engineering*, vol. 16, no. 4B, pp. 532-542, October 1991.
- [8] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. on SMC*, vol. 9, no. 1, pp. 62-66, January 1979.
- [9] A. Amin and W. H. Wilson, "Hand-Printed Character Recognition System Using Artificial Neural Networks," pp. 943-945, July 1993.
- [10] A. Cheung, M. Bennamoun and N. W. Bergmann, "A New World Segmentation Algorithm for Arabic Script," (Auckland, New Zealand), DICTA'97, pp. 431-435, December 1997.
- [11] M. Bennamoun and B. Boashash, "A Structural Description based Vision System for Automatic Object Recognition," *IEEE Trans on SMC*, vol. 06, no. 27, pp. 893-906, 1997.
- [12] A. Amin and J. F. Mari, "Machine Recognition and Correction of Printed Arabic Text," *IEEE Trans. on SMC*, vol. 19, no. 5, pp. 1300-1306, October 1989.
- [13] A. Cheung, M. Bennamoun and N. W. Bergmann, "The Arabic Optical Character Recognition Systems: Statistical and Neural Network Approaches," IAIF'97, pp. 293-298, November 1997.

الغرور البشري من نفس الإنسان ، والغرور هو أول مراتب المعصية ، وأول مداخل
الشيطان إلى النفس ، لأنه يجعل النفس تحس بقدراتها وتبعد هذه القدرات ، ولا يغتر
الإنسان إلا ابتعد عن الله ، وحسب أنه يستطيع أن يستغني عنه ، وأنف من الاستغفار
وطلب الرحمة من مولا .

(a)

الغرور البشري من نفس الإنسان ، والغرور هو أول مراتب المعصية ، وأول مداخل
الشيطان إلى النفس ، لأنه يجعل النفس تحس بقدراتها وتبعد هذه القدرات ، ولا يغتر
الإنسان إلا ابتعد عن الله ، وحسب أنه يستطيع أن يستغني عنه ، وأنف من الاستغفار
وطلب الرحمة من مولا .

(b)

Fig 10. (a) The original document. (b) The recognized result.