ORAN SYSTEM: A BASIS FOR AN ARABIC OCR

Abdelmalek Zidouri

Electrical Engineering Department, King Fahd University of Petroleum and Minerals,
P.O. Box 1360, Dhahran 31261, Saudi Arabia

malek@kfupm.edu.sa

Abstract:

In this paper we present a system for document understanding and for recognition of printed Arabic text. Arabic characters must be segmented before recognition. We overcome the problem of segmentation by our proposed ORAN system (Offline Recognition of Arabic characters and Numerals). ORAN is based on a method called Modified MCR. Using a stroke index, we can parse compound document images into three categories: text, picture and graphical patterns. The eventual Arabic text block is considered for further analysis and recognition. Its Modified MCR expression is obtained. The recognition is achieved by simple matching of candidate characters to reference prototypes. The prototypes are designed according to some topological features of the strokes obtained with respect to baseline detection and a zoning scheme. The recognition rate obtained for a popular Arabic font is higher than 97%.

KEYWORDS: Document Understanding, Arabic Character Recognition, MCR Expression, Strokes, Connected Components.

1. Introduction

Character recognition is one of the oldest fields of research. It is the Art of automating both the process of reading and keyboard input of text in documents. A major part of information in documents is in the form of alphanumeric text. However characters' automatic recognition is not an easy problem, especially for languages using the Arabic script. Arabic is written cursively (connected) even when machine printed or typed. Despite the rapidly increasing electronic sources and libraries, paper documents will live with us for a long time to come. Automatic handling of such documents needs robust and reliable document understanding techniques. For Chinese, Japanese and Roman characters based languages, the problem of recognizing well formed and neat characters have been largely solved for. There are several commercial OCR's available in the market today, example Omnipage, Recognita, IRISpen, Kanjiscan, YondeKoko to name just a few. Research is now being focused on more challenging problems of poor quality print, broken and touching characters or unconstraint handwriting [2, 3, 4, 5] and others. Comprehensive surveys can be found in [6, 7, 8]. However, cursive writing like Arabic, where the segmentation problem of text into distinguishable characters is the source of many errors for several OCR, is still not yet fully explored. It is only in the last two decades where some significant work has been devoted to Arabic script based languages [9, 10, 11, 12, 13, 14]. A state of the art of Arabic character recognition research is presented in [15] and [16].

Different approaches have been tried. Some with certain success, but until now there exist no competition in the Arabic OCR field. The characteristics of Arabic do not allow direct implementation of many algorithms used for other languages like English or Chinese. This is mainly because Arabic characters are always connected even when typewritten and they have four different shapes according to their position in a word. Most of the difficulties lie in the separation of words or sub-words into individual characters for subsequent classification and recognition. It is therefore of a primary importance to tackle the problem of segmentation successfully in any potentially practical Arabic OCR system [17]. As segmentation is not an aim by itself, in our system we overcome this problem of segmentation of connected characters inherent to Arabic writing. The system is based on a method called Modified MCR (Minimum Covering Run) [1, 18]. This expression method was developed to represent binary document images by a minimum number of both types, horizontal and vertical runs. This is in analogy with *bipartite* graph in graph theory. From correspondence between binary image and bipartite graph, where runs correspond to partite sets and edges of the graph correspond to pixels in the image, finding the MCR expression amounts to constructing a minimum covering in the corresponding bipartite graph. **Fig.1** shows a pattern, its Modified MCR expression and corresponding bipartite graph. The characters are represented by a number of horizontal and vertical parts called strokes. Specific features are then associated with these strokes. Thus separating words into characters is done once the characters composing parts are successfully recognized. This is similar to the approach of [9] and [19] in the sense it deals with the problem of separating the characters after their recognition. These strokes are labeled and a database of prototypes for each character shape is build from the features of these strokes. Recognition is achieved by simple matching to the reference prototypes as it is explained in the following sections. **Fig.2** shows a block diagram of the complete system.

2. Modified MCR Expression for Arabic Script

Arabic (Farsi or Urdu) is written from right to left cursively most of its characters are connected even when typed or machine printed. Arabic characters take different shapes (two shapes for six of them and four shapes for the remaining 22 characters) depending on their position in a word; see **Table.1**.

Arabic characters are written following a baseline. Most of the information is contained above the baseline. Arabic is widely used now in the information technology sector around more than 20 countries. At present the choice is very limited for available commercial OCR systems of Arabic language [20]. The relative cost of an Arabic OCR to a Japanese OCR for example is more than five to one. It is even more than that when compared to its English counterpart. Not to speak about the compatibility and user friendly issues that can be raised.

E	M	B	I	E	M	B	I
ـا	ـا	ا	ا	ـظ	ـظـ	ظـ	ظ
ـب	ـبـ	بـ	ب	ـك	ـكـ	كـ	ك
ـت	ـتـ	تـ	ت	ـل	ـلـ	لـ	ل
ـث	ـثـ	ثـ	ث	ـم	ـمـ	مـ	م
ـج	ـجـ	جـ	ج	ـن	ـنـ	نـ	ن
ـح	ـحـ	حـ	ح	ـع	ـعـ	عـ	ع
ـخ	ـخـ	خـ	خ	ـغ	ـغـ	غـ	غ
ـد	ـد	د	د	ـف	ـفـ	فـ	ف
ـذ	ـذ	ذ	ذ	ـق	ـقـ	قـ	ق
ـر	ـر	ر	ر	ـس	ـسـ	سـ	س
ـز	ـز	ز	ز	ـش	ـشـ	شـ	ش
ـص	ـصـ	صـ	ص	ـه	ـهـ	هـ	ه
ـض	ـضـ	ضـ	ض	ـو	ـو	و	و
ـط	ـطـ	طـ	ط	ـي	ـيـ	يـ	ي

Table.1 Arabic Alphabet in their different shapes: Isolated (I), at the beginning (B), in the middle (M) and at the end (E) of a word.

Our approach to Arabic cursive writing, based on Modified MCR Expression is a new one. We consider Arabic like any other stroke-like languages. This has been made possible by the structural features that the Modified MCR expression offers to represent text in binary document images.

The MCR technique is useful to several tasks in document image analysis and understanding. This method was developed in the Precision and Intelligence Laboratory of Tokyo Institute of Technology. The structural information obtained is used in ORAN system, as a preprocessing for character recognition. After the modified MCR expression is obtained for a document image, the strokes are divided into overlapping and non-overlapping parts. The non-overlapping parts of a pattern are labeled and ordered with respect to their absolute position in the document image. This is in a top down, left to right priority, to follow the Arabic way of writing. The labeling process of the parts obtained provides dynamic information from a static image, needed for character recognition. These parts are described according to 8 topological features. These are obtained after a zoning scheme, where the line of text is detected and divided into four zones: an upper zone, a middle zone, a baseline zone and a lower zone.

This has been decided because most of the information from shape of the body of Arabic Characters lies in the upper portion of the characters. Also most stress marks (dots of characters and character **"ء"** called "Hamza") and diacritics occur effectively above the writing line that we call baseline.

3. System Overview

ORAN can be a basis for a complete Arabic OCR system. We can perform image block classification in a compound document image into three categories. The classified Arabic text block in a document will be further processed for character recognition. This is based on the Modified Minimum Covering Run Expression. The block classification scheme is based on a feature called stroke index for document image analysis [21]. This representation is a structural description for patterns in document images by minimum selected horizontal and vertical runs called covering runs. This has proved to be very useful in extraction of stroke components in text.

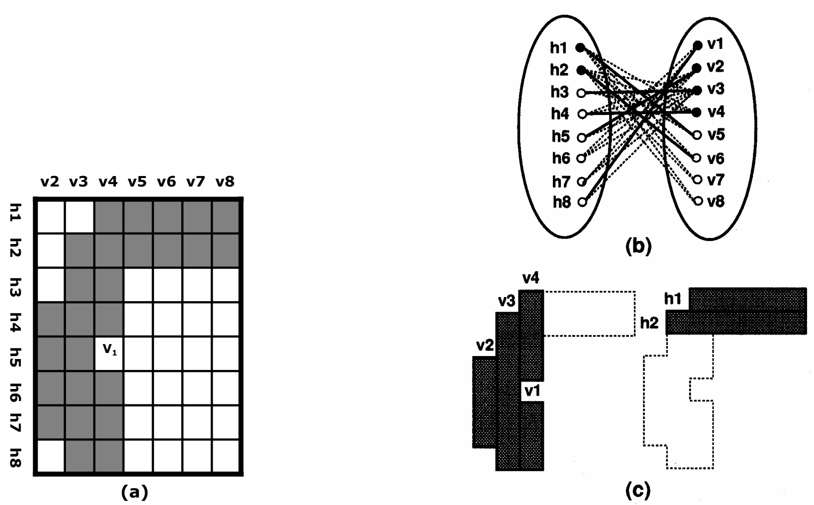

Fig.1. a) Simple Pattern b) Corresponding Bipartite Graph c) Covering Runs by Modified MCR

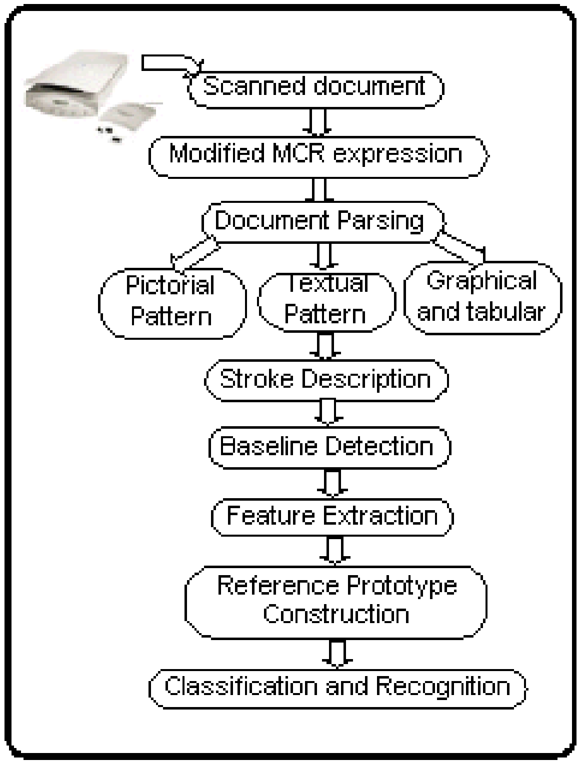

Fig.2. Block diagram of ORAN system

3.1. Document Parsing

One of the current trends in document image processing such as image compression, database management, image analysis and understanding is to consider these tasks through a unified approach. The MCR expression method was developed in this scope. Using information from the MCR expression for classifying various types of regions according to the possibility of strokes extraction, we introduced a feature called stroke index for document image analysis. As document images may present a variety of patterns such as graph, text, picture or dithered image, it is necessary to classify these different patterns into categories for image analysis and understanding. This stroke index is a function of the number of horizontal and vertical runs in the original image and of number of covering runs by the MCR expression. Mapping of typical image patterns on the index feature space is shown in **Fig.3**. The capability of the strokes to be used as a basic part in representation of characters and lines makes it important to know the possibility of strokes extraction from document images. The computed values of the MCR stroke index for these five typical patterns map into four distinct regions as indicated by the arrows on the figure. The picture or dithered image in pattern I and II, mapped near the origin, indicating that there is no possibility of stroke extraction from these patterns. The type III pattern which has a mixture of vertical and horizontal lines mapped in region III which has high $\mathbf{S_H}$ value (horizontal stroke component) and high $\mathbf{S_V}$ value (vertical stroke component). The other two patterns mapped in regions IV and V corresponding to high $\mathbf{S_V}$ and high $\mathbf{S_H}$ respectively.

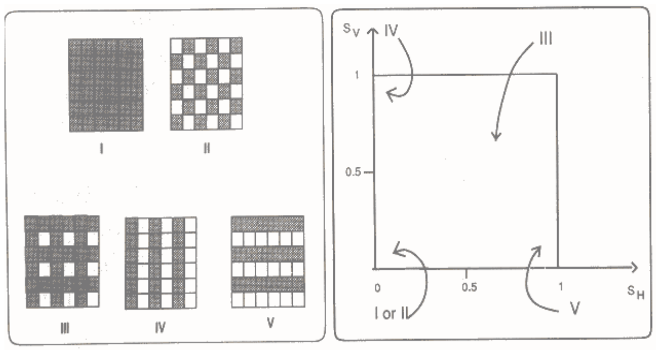

Fig.3 Typical image patterns and their mapping on the Stroke Index Feature-Space.

One of the applications of this stroke index is the classification of different patterns on a document image. It can also be used to detect the orientation of a document whether it is in a portrait or landscape mode. It also gives an insight on the possibility of stroke extraction from document images. We have experimented on several document images containing graphs, pictures and different blocks of text and results were satisfactory. The compound documents have been divided into different blocks then the stroke index values for each block were computed. Mapping values of the stroke index for different blocks of the documents on the $\mathbf{S_H}$-$\mathbf{S_V}$ Feature Space classifies the various types of patterns into distinct regions as shown in **Fig.4.**

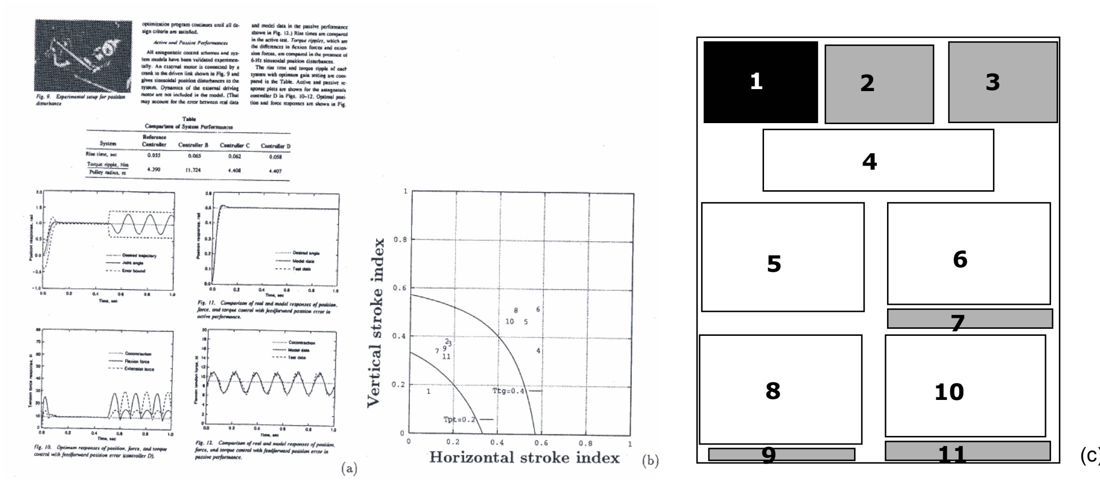

Fig.4. Classification of patterns on a compound document image to 3 categories: Text, Picture and Graph.

3.2. Stroke representation by Modified MCR

The MCR expression method was developed to express binary document images by a minimum number of both types horizontal and vertical runs. It represents binary images with no redundancy and without any loss of information. Some of the horizontal and vertical runs called covering runs are suitably selected to represent the image with a minimum number of runs. In binary images, no runs from the same direction cross each other and every black pixel can be considered as a crossing point of one horizontal run and one vertical run. Therefore an analogy between binary image and bipartite graph has been possible. It has been shown that horizontal and vertical runs of binary image can be thought of as *partite* sets of a *bipartite graph*. From this correspondence between the binary image and the bipartite graph, where runs correspond to partite sets and edges of the graph correspond to pixels in the image, finding the MCR expression amounts to constructing a minimum covering in the corresponding bipartite graph [22]. This in turn is the same as finding the maximum matching, which has been solved for by graph theoretical algorithms [23, 24]. The word “stroke” is being used here to mean such *parts* as the four curved segments composing a character *zero*, or a circular shaped pattern, i.e. an “O” or a similar shaped pattern would be represented by 2 vertical and 2 horizontal “strokes”.

For example crossing lines would be decomposed into 4 strokes. A character “C” or a similar curved pattern at the end of many Arabic characters will be represented by 1 vertical and 2 horizontal strokes and so on. In this way we don't have to worry about the order of writing, when describing textual patterns. Modified MCR expression provides efficient structural information of stroke components in text. We use it to extract the features needed for classification and recognition. For later use we subdivide these strokes into two categories: crossing and non crossing parts. An example of an Arabic word, its description and the non-crossing parts used for recognition, is shown in **Fig.5**.

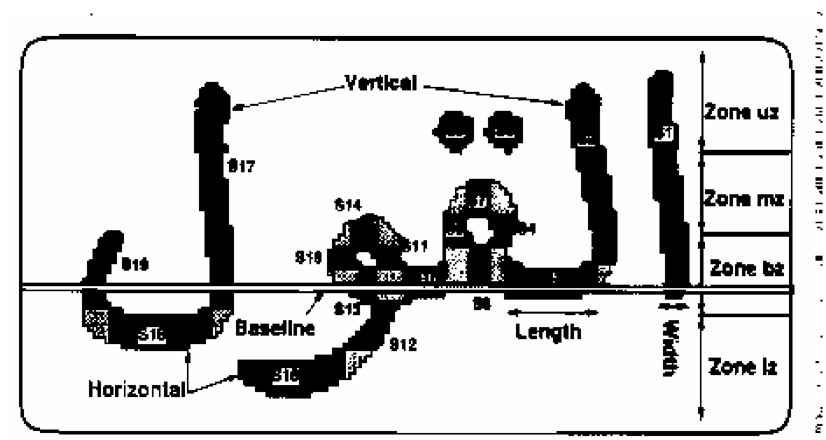

Fig.5. An Arabic word and its representation by the non crossing parts, showing the features used.

3.3. Baseline Detection

Arabic writing is in some way like handwritten Latin. Characters are connected even when printed or typewritten. This fact of character connection by some short horizontal lines or strokes is used to detect what we call baseline. We detect the baseline by searching for the line containing the largest number of horizontal non-overlapping parts in the text document. This baseline is used to associate with every part, the relative position with respect to the writing line as a feature that we use in the recognition phase of the system. In fact, on Arabic typewriters, the key for such horizontal line to connect characters is one of the most used keys. This is to make clear writing and easy comfortable reading for the eye. This line actually does not contain any information apart from the connectivity whether it is short or extended in length like in some titles or before the last character of many words. **Fig.6** shows such an example.

جـــــامعـــــــــة المــــــــلك فهـــد للبــــــترول والمعـــــــــادن

Fig.6. Example of extended connecting line for comfortable reading of Arabic

Using this information and the fact that most Arabic characters themselves have their horizontal part written on this baseline, we look for the line containing the largest number of horizontal strokes and label it as baseline, see **Fig.7**. The detection is by projection of the center position of all horizontal parts on the vertical axis. The baseline is used as a reference to divide the text line into four horizontal zones where the zone zero is that which contain the baseline. The three other zones are the middle zone, the upper zone and the lower zone. Two zones have been chosen above the baseline zone (the middle and upper zones) because the main information of Arabic text is contained in the upper part of the main body of each character. Also most stress marks like dots are above the main body of the character. There are four times as many characters with dots above their main body shape as those which have dots below their main body shape. The baseline proved to be a good reference for the Arabic characters feature selection.

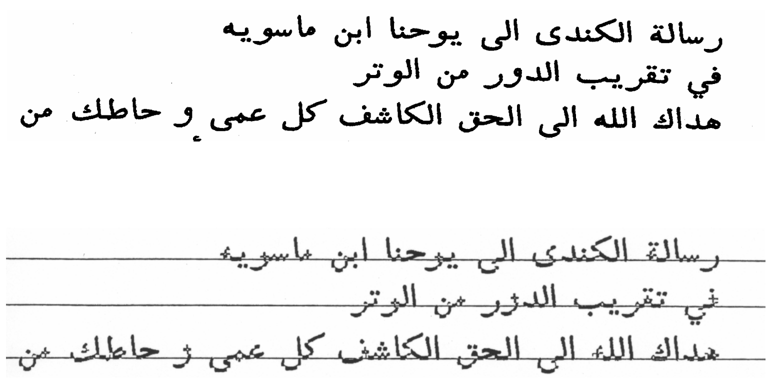

Fig.7. Example of Arabic Text, Corresponding Modified MCR description showing the detected baseline

3.4. Features Selection

After modified MCR expression is obtained for a document image, the strokes are divided into crossing and non-crossing parts. The non crossing parts of a pattern are labeled and ordered with respect to their absolute position in the document image. This is done in a top down, left to right priority, to follow the Arabic way of writing. The labeling process of the parts obtained provides the dynamic information from a static image, needed for character recognition. The following information is obtained for every part:

- ***{ln, wd}***; geometrical features (length, width)
- ***Number of strokes*** in a Connected component
- ***Region label***
- ***{tp} = {h,v}***; type of stroke (horizontal or vertical)
- ***{ld, rd}***; direction left and right from the center 7 directions each.
- ***{ps} = {lz, bz, mz, uz}***; relative position with respect to baseline

These features are based on the detected baseline. For these topological features to be invariant for scale a unit of the size of a *dot* of an Arabic letter like (ب) has been taken as a "*Baselength*". Although strokes are divided into only two types namely horizontal or vertical, we still keep trace of the original curvature and slant shape by the direction feature or angle information. These features are used to build the reference prototypes.

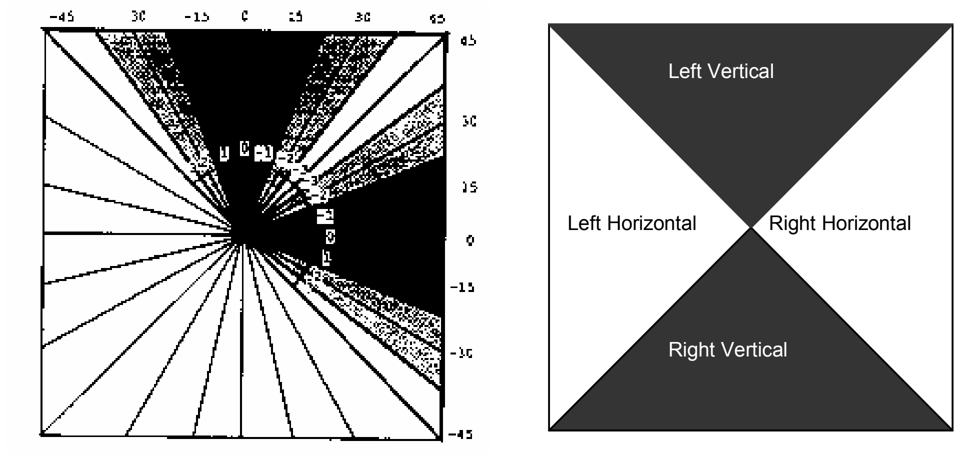

Fig. 8 The features right direction ***rd*** and left direction ***ld***

3.5. Matching and Recognition

The recognition is based on the matching of a candidate character with a reference prototype.

From observation of the modified MCR output we designed prototypes for every character shape of the Arabic Alpha Numeric set. The design is made manually. The prototypes are designed based on the description obtained from modified MCR. This is done according to some model documents used for training and tuning the system. This is done only once at the design stage. Arabic characters take up to four shapes. In total we provide a database of prototypes of 100 character shapes. Some characters are simple like the letter **ا**, therefore the corresponding prototype in the databank is also simple. The number and complexity of prototypes increase with the characters that have multiple loops or cusps and dots like هـ ,ش or ك. We have an average of three prototypes per character shape. A class of character shape composed of ***k*** strokes is described as follows:

$$Chr\_{_C} = (S\_{_{c1}}, S\_{_{c2}}, \ldots, S\_{_{ck}};\ \textit{Connection\_Rule})$$

$$\text{where } S\_{_{cj}} = (\{ln_j\}, \{wd_j\}, \{tp_j\}, \{ld_j\}, \{rd_j\}, \{ps_j\}, \{con_j\}, \{rgn_j\})$$

$$j \text{ from } 1 \text{ to } k.$$

For example, the second character in the word of **Fig.5** would be described as follows:

$$Chr\_{}_{LAAM\_B} = (S_{1LAAM\_B}, S_{2LAAM\_B}, Connection\_Rule)$$

where

$$S_{1LAAM\_B} = [\{ln > l5\}, \{normal\}, \{v\}, \{*\}, \{*\}, \{uz\}, \{*\}, \{*\}]$$

$$S_{2LAAM\_B} = [\{ln > l1\}, \{normal\}, \{h\}, \{*\}, \{*\}, \{bz\}, \{*\}, \{*\}]$$

and

$$Connection\_Rule \equiv \{rgn(S_{1LAAM\_B}) = rgn(S_{2LAAM\_B})\}$$

* is a don’t care value.

A complete description and recognition result of the previous word is given in **Fig.9**.

The system has been trained with several model documents of printed Naskh font, with a laser printer. The Modified MCR expression output carefully observed and prototypes for each character shape are designed. The test data set is composed of different documents of characters of the same and different size as that of the training set, but of the same font. To make the reference prototypes cover a wide range of possible variations, without on the other hand, cause misrecognition by substitution, we used the model document as tests in the beginning to tune the system and correct for any errors until we got about 100% recognition for the model documents. Some prototypes have been readjusted accordingly. In some cases new ones have been designed when necessary.

The recognition rate obtained for the test data set of many documents composed of more than 6000 characters was more than 97%.

```
Word Segment 1 is Chr-Shape ا
str.no[1] len=39 wdth=4 typv ldr0 rdr0 pos2 con1 rgn1 → ا
Word Segment 2 is Chr-Shape ل
str.no[2] len=31 wdth=5 typv ldr0 rdr0 pos2 con13 rgn2 → ل
str.no[3] len=16 wdth=5 typh ldr0 rdr0 pos0 con13 rgn2 → ل
Word Segment 3 is Chr-Shape ق
str.no[4] len=5 wdth=4 typv ldr0 rdr-1 pos1 con13 rgn2 → ق
str.no[5] len=7 wdth=5 typv ldr0 rdr0 pos2 con1 rgn3 → ق
str.no[6] len=3 wdth=7 typh ldr0 rdr0 pos0 con13 rgn2 → ق
str.no[7] len=4 wdth=6 typh ldr0 rdr0 pos1 con13 rgn2 → ق
str.no[8] len=5 wdth=4 typv ldr0 rdr0 pos1 con13 rgn2 → ق
str.no[9] len=7 wdth=5 typv ldr0 rdr0 pos2 con1 rgn4 → ق
str.no[10] len=7 wdth=5 typh ldr0 rdr0 pos0 con13 rgn2 → ق
Word Segment 4 is Chr-Shape و
str.no[11] len=3 wdth=5 typv ldr1 rdr1 pos0 con13 rgn2 → و
str.no[12] len=8 wdth=5 typv ldr-1 rdr-2 pos-1 con13 rgn2 → و
str.no[13] len=2 wdth=6 typh ldr0 rdr0 pos0 con13 rgn2 → و
str.no[14] len=3 wdth=6 typh ldr0 rdr0 pos1 con13 rgn2 → و
str.no[15] len=3 wdth=5 typv ldr0 rdr0 pos0 con13 rgn2 → و
str.no[16] len=10 wdth=5 typh ldr0 rdr-1 pos1 con13 rgn2 → و
Word Segment 5 is Chr-Shape ل
str.no[2] len=39 wdth=5 typv ldr0 rdr0 pos2 con3 rgn5 → ل
str.no[2] len=18 wdth=6 typh ldr0 rdr-1 pos-1 con3 rgn5 → ل
str.no[2] len=13 wdth=4 typv ldr-1 rdr0 pos1 con3 rgn5 → ل
```

Fig.9. Sample of output program for recognition of the word in Fig.5

4. Discussion: Merits and Limitations

Our system in its recognition part is like the approach in [9], in the sense that it does not address the problem of recognition and segmentation separately. As the segmentation is not an aim by itself, we can perform recognition directly whenever a correct match is achieved. There are other reports about successful projects in this field for both printed and handwritten Arabic [10,14,26]. The high recognition rates for some of these results are only for already supposed segmented characters [14,26]. In [10] the authors used a simple segmentation method. Their method is based on the projection profiles on both rows and columns. They use these projections to guide the segmentation process. This approach fails for overlapping characters. The segmentation of words into characters is the most crucial stage in an Arabic character recognition system.

The block classification part is for any document image. The classification failed when some blocks are mixed like too much text with drawings. Actually this was difficult to classify even manually whether it is of graphic or text type. The recognition is limited for printed Arabic text only. We build the reference prototypes for a popular Arabic font called Naskh. The recognition rate drops considerably if we use the system as it is for other fonts and styles or for handwriting. However we argue that our database is still very small and we can expand it easily. We are using about three prototypes per character shape in average. This leaves room for adding more prototypes to remedy for the rejection errors. Adding confidence weight to the recognition stage will improve the recognition rate. The enhancement and extension of the system for the multi-font aspect is in progress as a funded project by King Fahd University of Petroleum & Minerals. The speed of recognition is about 50 characters per second. This is reasonable at this stage and we can afford adding some preprocessing and more improved feature selection. Our method is segmentation free. The segmentation is the source of most errors for other systems.

5. Conclusion

We presented in this paper a system called ORAN for document understanding and recognition. It classifies patterns in compound document images into three categories: text, picture and graphics. It performs the recognition of printed Arabic text. It is a knowledge based system, using a structural description of the text documents and performs recognition with simple matching of candidate characters to reference prototypes that are built for this purpose. ORAN is a segmentation free method. It is a basis for a complete OCR system. The extension and enhancement to multifont recognition is under progress. We report a recognition rate of more than 97% for a test set of more than 6000 characters at a speed of about 50 characters per second. The system trained over a set of documents to take into account the variations due to noise from printing processes or quantization noise of the scanning device. A recognition rate as high as 99.7% for the model documents is obtained.

Acknowledgement

The author would like to thank Prof. M. Sato of Tokyo Institute of Technology and Dr S. Chinveeraphan of Thaicom, Thailand and KFUPM for their valuable contribution to this work.

6. References

[1] Chinveeraphan, S., Zidouri, A., and Sato, M., *"Modified Minimum Covering Run Expression of Binary Document Images"*, IEICE Trans. Inf. &. Syst., vol. E78-D, no.4, pp.503--507, Apr. 1995.

[2] Sawaki, M. And N. Hagita, "Text-line extraction and character recognition of document headlines with graphical designs using complementary similarity measure," IEEE Transactions on PAMI vol. 20 no 10 pp. 1103-1109 Oct. 1998.

[3] Srihari, S. N. S.-H. Cha, and S. Lee, "The Discriminatory Power of Handwriting," Invited Paper, *Proceedings SPIE, Document Recognition and Retrieval IX*, San Jose, CA, pp. 129-142, January 2002.

[4] I.S. Oh, C.Y. Suen "A Class-modular feed forward neural network for handwriting recognition" Pattern Recognition 35, pp 229-244 , (2002).

[5] Cheng-Lin Liu, Hiroshi Sako, Hiromichi Fujisawa, "Performance evaluation of pattern classifiers for handwritten character recognition" IJDAR 4, 3, 191-204, (2002).

[6] Plamondon, R., and S. N. Srihari, On-line and Off-line Handwriting Recognition: A Comprehensive Survey, *IEEE Transactions on Pattern Analysis and Recognition*, 22(1), 63-84, January (2000).

[7] Y.Y. Tang, S.W. Lee, and C.Y. Suen, "*Automatic document processing - a survey*", Pattern Recognition, vol. 29, no. 12, 1931-1952, 1996.

[8] Mori, S., Suen, C. Y. And Yamamoto, K. Historical review of OCR research and development, Proceeding of the IEEE 80, 1029-1058, 7 (1992).

[9] Cheung A., Bennamoun M., Bergmann N. W., "An Arabic optical character recognition system using recognition-based segmentation", *Pattern Recognition* 34, 215-233, 2001.

[10] L Hamami And D. Berkani "Recognition System for Printed Multi-Font and Multi-Size Arabic Characters" *Arabian Journal for Science and Engineering*, Vol. 27, No. 1B, April 2002.

[11] Al-Ohali, Y., M. Cheriet, and C. Suen, *"Databases for recognition of handwritten Arabic cheques"*, pp. 111-121. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam, September 11-13 2000.

[12] M. Fakir and M. M. Hassani, "Automatic Arabic Characters recognition by moment invariants" Colloque international de telecommunications, Fes, Morocco, pp 100-103, 1997.

[13] A. Zidouri, "ORAN: Offline Recognition of Arabic characters and Numerals", International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Susa, Tunisia, Mar 17-21, (2003).

[14] Abuhaiba, I. S. I., Mahmoud, S. A. And Green, R. J., "Recognition of Handwritten Cursive Arabic Characters," IEEE Trans. on PAMI, vol. 16, no.6, pp. 664-672, Jun. 1994.

[15] Al-Badr, B., And S. Mahmoud, Survey and bibliography of Arabic optical text recognition, *Signal Process*. 41, pp. 49-77, 1995.

[16] Amin A., "Off-line Arabic character recognition The State of the Art", *Pattern Recognition*, Vol. 31 No. 5, pp. 517-530, 1998

[17] D. Motawa, A. Amin and R. Sabourin *"Segmentation of Arabic Cursive Script."*, pp. 625-628 IEEE 1997

[18] S. Chinveeraphan, A. B. C. Zidouri, And M. Sato, "Modified Minimum Covering Run Expression of Binary Document Images," IEICE Trans. Inf. &. Syst., vol.E78-D, no.4, pp.503--507, April, 1995.

[19] B.Al-Badr, R.M.Haralick "*Segmentation-Free Word Recognition with Application to Arabic"* International Journal on Document Analysis and Recognition, Vol. 1 No. 3 pp.147-166, 1998.

[20] T. Kanungo, G. A. Marton, and O. Bulbul. "OmniPage vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products". *SPIE Conference on Document Recognition and Retrieval VI*, pages 109-121, January, 1999

[21] A. Zidouri, "Powerful Tool for Arabic Document Understanding", The First GCC Industrial Electrical & Electronics Conference. Manama, Bahrain, May 13-14, 2003.

[22] Chinveeraphan, K. Douniwa, and M. Sato, *"Minimum Covering Run Expression of Document Images Based on Matching of Bipartite Graph"* IEICE Trans. Inf. & Syst., vol. E76-D, no.4, pp.462--469, Apr. 1993.

[23] Hopcroft, J.E., and Karp, R.M., *"An* $n^{5/2}$ *algorithm for maximum matching in bipartite graphs."* SIAM, J. Comput, vol. 2, no.4, pp.225--231, 1973.

[24] Chinveeraphan, S., Zidouri, A., and Sato, M., "*Fast Algorithms for Minimum Covering Run Expression*", IEICE Trans. Inf. & Syst., vol. E77-D, no.3, pp.317--325, Mar. 1994

[25] A. Zidouri, M. Sarfraz, S. N. Nawaz and M. J. Ahmed "PC based Offline Arabic Character Recognition System" ISSPA, Seventh Int'l Symposium on Signal Processing and its Applications. PARIS July 4-7 2003.

[26] M. M. Fahmy, S. Al-Ali "Automatic Recognition Of Handwritten Arabic Characters Using Their Geometrical Features" Studies in informatics and Control, Vol. 10 No. 2, June 2001.