ON OPTICAL CHARACTER RECOGNITION OF ARABIC TEXT

Abdelmalek Zidouri[1], Muhammad Sarfraz[2]

1: Assistant professor, Electrical Engineering Department, KFUPM # 1360, Dhahran 31261.
Email: malek@kfupm.edu.sa

2: Associate professor, Information and Computer Science Department, KFUPM # 1510, Dhahran 31261
Email: sarfraz@kfupm.edu.sa.

ABSTRACT

Although, optical character recognition has made tremendous achievements in the area of desktop publishing, yet a huge amount of work is required to be done. Unlike Roman like languages, there are various languages possessing a large number of fonts and/or having complicated shapes. Arabic language is one of those languages, which is somewhat complicated in its construction. Although a reasonable amount of work has been reported so far for Arabic language but still a good amount of work is needed to be developed. In addition, many other languages also need considerable attention for automatic generation in their recognition. Efficient, robust, and error free methodologies are required to develop systems for such languages so that the recent hardware technologies, to display and print, can be utilized.

This work is devoted to one way of addressing the problem of recognition of the Arabic alphabet. We give a brief survey of the state of the art in Arabic Character Recognition and different methods and approaches to this problem. We show that recognition can be achieved by simple matching to prebuilt prototypes of all the Arabic Character set. This free segmentation approach proved to be efficient for the recognition of one font of the Arabic language. We deal with Arabic as a well-structured language and base our prototype description on a method called "Minimum Covering Run Expression".

We also show that our database of prototypes is easily extendable to allow for multifont recognition of Arabic as a basis for a full Arabic OCR system.

Keywords: *Arabic Character Recognition, MCR Expression, OCR, Strokes, Matching.*

الملخص

رغم الإنجاز الهائل الذي حققته القراءة الآلية في مجال النشر المكتبي؛ إلا أن الباب يبقى مفتوحا على مصراعيه أمام الجهود الجبارة التي ينبغي بذلها في هذا المضمار بالنسبة للغة العربية. خلافا للغات المبنية على الحرف اللاتيني؛ فاللغة العربية تزخر بكم هائل من أنواع الخطوط و أساليب الخط مما يجعل الخط العربي فنا قائما بذاته. إلا أن هذا قد زاد من تعقيد مهمة التعرف الضوئي على الحروف (OCR) التي هي في حد ذاتها عملية صعبة و معقدة بالنسبة للحاسب الآلي.

و تؤثر طبيعة النصوص العربية و غيرها من اللغات تأثيرا مباشرا على عملية التعرف الضوئي على الحروف؛ فخصائص النص العربي مثل تشابك الحروف و تداخلها؛ و بعض الحروف لا يقبل الاتصال إلا من جهة واحدة بالإضافة إلى تباين أشكال الحرف العربي وفقا لموقعه في بداية أو وسط أو نهاية الكلمة كلها تزيد من خصوصية و تفرد البحث في هذا المجال.

نعرض في هذه الورقة طريقة جديدة تعتمد على وصف تركيبي للحروف بناء على طريقة تسمى MCR أو (أقل مسح مغطي). هذه الطريقة تتجنب عملية تقطيع الكلمات إلى حروف و التي هي ليست هدفا في حد ذاتها؛ و يتم التعرف على الحرف و قراءته عندما تتم مواءمته مع نموذج أولي في قاعدة المعطيات التي يتم إعدادها لهذا الغرض.

لقد تمت تجربة الطريقة بنجاح على خط النسخ و نبين هنا سهولة توسيع الطريقة إلى أنواع أخرى من الخطوط العربية.

1. INTRODUCTION

Reading text in documents automatically is the key to get information therein shared and stored in the most optimal way. This will support if not replace the more tedious and expensive task of data input through keyboard. This problem has been largely tackled for Latin script languages, but remains still widely open for Arabic script based languages. The characteristics of the Arabic language do not allow direct implementation of many algorithms used or developed for other languages. The particular characteristics of the Arabic characters and text add to the difficulty of this already challenging problem of character recognition.

Nevertheless, since the early work carried out by [Amin A. et al. 1986, 1992] there have been reports about successful research projects in the field of printed Arabic character recognition. Connectivity of characters being an inherent property to Arabic writing, it is therefore of a primary importance to tackle the problem of segmentation [Iwaki et al., 1985], [Jain et al., 1992], [Chen et al., 2000] in any potentially practical OCR system, especially for Arabic. Several researchers have attempted to solve the problem of segmentation of cursive writing with success for on-line recognition however off-line cursive writing, where the order of the strokes made by the writer is lost, has not been satisfactorily solved.

A good survey for off line Arabic character recognition can be found in [Amin, 1997]. It shows that machine simulation of human reading has been the subject of intensive research for almost three decades. The state of Arabic character recognition research throughout the last two decades is presented in [Al-Badr et al., 1995] and [Amin, 1998]. A large number of research papers and reports have already been published on Latin, Chinese and Japanese characters [Sawaki et al., 1998], [Mori et al., 1992], [Govindan et al., 1990], [Impedovo et al., 1991] and others. However, relatively little work has been conducted on the automatic recognition of Arabic characters. Both machine printed and handwritten Arabic text is connected and cursive, and this complex problem is still an open research field. Windows based software that can interact with a scanner and can recognize a single font of Arabic script

with 85% accuracy at about 16 characters per second, has been presented in [El-Dabi et al., 1990]. Hidden Markov Models and Neural network are also being investigated like in [Amin et al., 1997], [Ben Amara et al., 1998], [Dehghan et al., 2001], [Mahjoub, 1996]. An algorithm developed for recognition of printed Farsi characters with various fonts, irrespective of size, rotation and stork is presented in [Namazi et al., 1996]. The system uses Pseudo-Zernike Moments as input features and the classifier consists of a complex of neural networks (NN) and fuzzy neural networks (FNN). The performance of the authors' system is evaluated on a database consisting of more than 3700 character samples. They claim having achieved a high recognition rate of 99.85%. This would have been an excellent system if Arabic characters were written separated like the Roman alphabet. A high performance Arabic character recognition system is introduced in [Alherbish et al., 1998] and [Bushofa et al., 1997]. The goal stated was to maximize both accuracy and speed. The authors developed a sequential Arabic character recognition system and mapped it into a multiprocessing environment. Experimental results show that the multiprocessing environment is very promising in enhancing a sequential Arabic character recognition system performance according to the authors. This sounds good if the problem of segmentation of text to individual characters is dealt with successfully. In our case, segmentation is achieved automatically whenever a correct final match is made [Zidouri et al., 1994], [Zidouri et al., 1995]. This means that our approach can be thought of as a segmentation free method. Segmentation is just a by-product of recognition. In this aspect, this is similar to the approach of [Cheung et al., 2001], and [Al-Badr et al., 1995] in the sense that it is a recognition-based segmentation method. Our method has been tested with one font and results proved to be encouraging. The multi-font aspect of the problem is under investigation.

2. ARABIC SCRIPT CHARACTERISTICS

Arabic language is one of the most ancient languages and spoken by many people in areas around the globe. The Arabic script and language have resisted any major change for centuries now. Text written or words used more than 1000 years ago are still being used and understood by schoolboys around the Arab world. Nevertheless, with the advent of computer age and information technology, efforts have been directed to adapt the Arabic script for ease of use with the new tools. This without too much loosing on conserving the esthetic and artistic nature of the Arabic alphabet. One such effort has been concerned with automating of the handling of Arabic characters and text. This effort is faced with the usual problems of character recognition in general in addition to problems that are specific to Arabic language only. Arabic presents some specific characteristics that are worth noting for the English reader.

- Arabic is written from right to left.
- It is composed of 28 characters
- The characters change shape depending on their position in a word

- They can be grouped in 100 character shapes
- They present a lot of similarities and composed of many loops and cusps, Tab.1.
- Characters are connected even when typed or printed
- Two kinds of spaces, between words and within a word introduced by characters that have no middle shape **MF**
- The characters are written following a writing line called herein "baseline", which is about 2/3 down the main body of the characters

Table 1 below shows the complete set of the Arabic Alphabet characters in their different shapes: when a character is being isolated or form (**IF**), at the beginning of a word or (**BF**), in the middle of a word or (**MF**), and at the end of a word or (**EF**).

EF	MF	BF	IF	EF	MF	BF	IF
ـا			ا	ـط	ـطـ	طـ	ط
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			د	ـف	ـفـ	فـ	ف
ـذ			ذ	ـق	ـقـ	قـ	ق
ـر			ر	ـس	ـسـ	سـ	س
ـز			ز	ـش	ـشـ	شـ	ش
ـص	ـصـ	ـصـ	ص	ـه	ـهـ	هـ	ه
ـض	ـضـ	ـضـ	ض	ـو			و
ـط	ـطـ	ـطـ	ط	ـي	ـيـ	يـ	ي

Table. 1 Arabic Alphabet in their different shapes:

Isolated (IF),
at the beginning (BF),
in the middle (MF), and
at the end (EF) of a word.

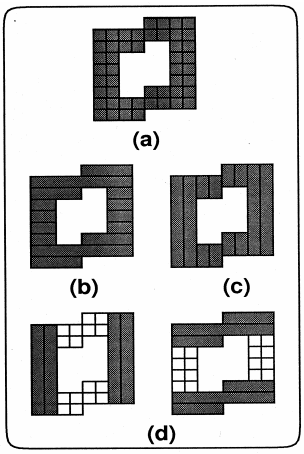

Fig. 1 (a) A binary pattern,
(b) its horizontal run representation,
(c) its vertical run representation, and
(d) its MCR representation showing the stroke decomposition of the pattern.

These characteristics have made Arabic character recognition one of the most difficult problems. It is well known that for Off-line cursive writing, where the order made by the writer is lost, an important problem in the recognition process is to segment the words into separable characters for recognition. Existing algorithms and techniques for Roman characters or Kanji based languages are not readily adaptable to Arabic. For a unified approach to various types of document image processing, a method called Minimum Covering Run (MCR) expression has been proposed by [Douniwa et al., 1992], [Chinveeraphan et al., 1993]

to represent document images. It has been modified by [Chinveeraphan et al., 1994], [Chinveeraphan et al., 1995] to extract strokes from textual patterns more precisely than the original expression, as structural features for recognition. Recently many types of document image processing such as image compression, image understanding, or image database managing have been extensively used. A structural method that tends to deal with these tasks in a unified approach is very much wanted. Also, there is a growing demand for the use of Arabic script in computer and communication technologies due to the wide spread of computers in businesses and home computing. The following sections give a brief description of the modified MCR technique and how it is being used as a preprocessing for Arabic character recognition.

3. MCR DATA DESCRIPTION

The modified MCR stands for modified minimum covering run. Generally, information in document images such as characters or lines is composed of horizontal and vertical strokes. Traditionally patterns are described either by vertical runs or horizontal runs of pixels. In MCR a pattern is described with both types of runs by a minimum number of runs called covering runs. The modified MCR uses some local stroke analysis to account for elongated segments, therefore is faster than the original expression and is better suited for stroke description. This is achieved at the expense of some more runs than the exact minimum, which is calculated with analogy to maximum matching in a corresponding bipartite graph in graph theory. The term "*stroke*" is not as usually defined in the literature as a pen lift or as usually defined in the Japanese or Chinese writing. Our definition of a stroke is somehow theoretical one. The word "stroke" is being used here to mean such "parts" as the four curved segments composing a character zero, or a "circle" shape pattern, i.e. an "O" or a similar shaped pattern would be represented by 2 vertical and 2 horizontal "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. Fig. 1 shows an example of "strokes" for representing a circle shaped binary pattern. However this way of defining a stroke is convenient. For example crossing lines would be decomposed into 4 strokes. In this way we don't have to worry about the order of writing, when describing textual patterns. The MCR expression aims at describing strokes of characters, or lines in document images competently. This increase in accuracy of stroke extraction is, however, achieved at the expense of increase in number of representative runs. It means that a number of covering runs in Modified MCR expression is more than that in MCR expression of the same image. In this way, the Modified MCR expression can be regarded as an intermediate processing which converts a bit pixel representation into (structural) shape representation [Davis, 1986]. That is, patterns such as characters in images are decomposed into horizontal and vertical strokes, and properties of the strokes are encoded. For recognition of characters the strokes as defined, are subdivided into two categories, overlapping (or crossing) parts and non-overlapping (or non-crossing) parts. Only the later are used at the moment to describe the prototypes used for recognition. The

advantage of stroke extraction is in its use in many applications to document image understanding such as segmentation, classification or character recognition. This later application is exploited in this work for printed Arabic characters.

4. SEGMENTATION FREE RECOGNITION

Our system is a recognition-based segmentation method. This means that we overcome the inherent problem of segmentation of words into individual characters for further feature extraction and classification. In our system the recognition is directly achieved by simply matching the candidate pattern to the prototypes that form the database of our system. This database is knowledge-based description of all the 100 classes of the Arabic character shapes. The description is obtained from a training set of documents for which the modified MCR expression is calculated. Successful results have been reported [Zidouri, 2001] for one font Arabic printed text. Fig.2 shows a block diagram of our system. We take a scanned document image. Find its modified MCR expression. We detect the baseline. Then from the results obtained by modified MCR description and our knowledge of the characters, we build a reference database of prototypes for each class of character shapes from a set of test document images. This step is performed only once at the learning stage. Features are extracted from the structural information obtained. We are using 8 topological features. The features used are

- ***{ln, wd}***; geometrical features (length, width)
- ***{tp} = {h,v}***; type of stroke (horizontal or vertical)
- ***{ld, rd}***; direction left and right from the center
- ***{ps} = {lz, bz, mz, uz}***; relative position with respect to baseline

In addition we associate with each stroke a ***region*** label **{rgn}** and a label for the number of ***connected components*** in that stroke **{con}**. Then comes the classification and recognition stage. This is achieved by simple matching of a candidate character on a scanned document to a prototype in the reference database build for this purpose. We match a candidate character **C** to a prototype **P** having the same number of strokes ***k***. All the prototypes are visited in this process, and if for a prototype **P = (S_1, S_2, …, S_k, connection_rule)** there is a candidate character **C = (s_1, s_2, …, s_k, connection_rule)** such that:

$$\forall S_j \in P \;\; \exists s_j \in C$$

where

$j = \{1,2,...,k\}$,

$S_j = (f_{1j}, f_{2j}, ..., f_{mj})$ $m \leq 8$ (f_{mj} is a relationship to, or a value of one of the 8 features used)

$$s_j = (\{\ln_j\},\{wd_j\},\{tp_j\},\{ld_j\},\{rd_j\},\{ps_j\},\{con_j\},\{rgn_j\})$$

if: $\forall f_{pj} \in S_j \quad \exists f'_{cj} \in s_j$

where

f_{pj} is a relationship to, or a value of one of the 8 features in the prototype, and f'_{cj} in the candidate character, and

p = {1, 2, …, m}

c = {1, 2, …, 8}

such that $(f_{pj} \supseteq f'_{cj}) \wedge (Connection\_Rule\_Match)$ then the candidate character shape **C** is matched to the prototype **P**.

We report a recognition rate of more than 97% for a popular Arabic font called Naskh, at a speed of about 10 characters per second.

This work is in progress for expansion to multifont. Fig.3 shows a portion of output showing the error of substituting the letter ف (or F) for the letter ق (or Q) because one of the dots where not properly recognized. This situation can be remedied for in some cases just by adding some more prototypes to the database of reference prototypes. For more robust recognition, it would be interesting to incorporate in MCR the information about runs of white pixels. This will provide the space information between runs of black pixels that we lack in our system. It will

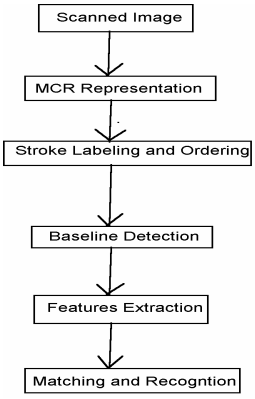

Fig.2 Block diagram of the system

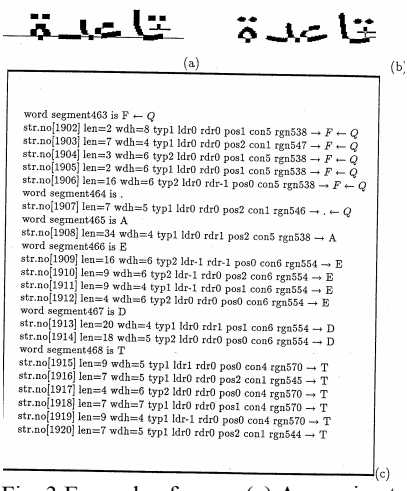

Fig. 3 Example of errors (a) Approximate strokes for visualization showing the detected baseline (b) Non-overlapping strokes extracted by modified MCR, (c) recognition and points to reiection substitution errors

solve for the substitution errors caused by selecting the character with fewer number of dots when the characters' bodies have identical or similar shape and their dots have the same position with respect to the baseline, and they differ only by the number of dots. Fig.3 illustrates an example of such case. The character main body shape common between ڤ "Q-**BF**" and ف "F-**BF**" which differ only by the number of dots makes it very difficult to find rules of connection to discriminate between the two exactly, as the dots are not connected. In this case, knowing the separating space would provide valuable information for discrimination between similar matching.

5. CONCLUSION

This paper gives a brief survey of the state of the art in Arabic Character Recognition and different methods and approaches used towards this goal. We presented our method and showed that in our system we extract features from a structural description of the binary patterns called MCR expression. These features serve to build a set of reference prototypes for the different classes of the character shapes. Recognition is then achieved by simple matching of a candidate character shape to the pre-built prototypes of all the Arabic Character set. We report a recognition rate of more than 97% for a popular font called Naskh at a speed of 10 characters per second. This free segmentation approach proved to be efficient for one font of Arabic printed characters. The multi-font aspect is under investigation.

ACKNOWLEDGEMENTS

This project has been funded by King Fahd University of Petroleum & Minerals under Project # EE/AUTOTEXT/232.
.0

REFERENCES

1. Al-Badr, B., and Haralick, R , 1995, Segmentation-Free word recognition with application to Arabic, *Proc, 3rd Int. Conf. On Document Analysis and Recognition*, Montreal, pp. 355-359.
2. Al-Badr, B., and S. Mahmoud, 1995, Survey and bibliography of Arabic optical text recognition, *Signal Process*. 41, pp. 49-77.
3. Alherbish, J. and R. Ammar, 1998, "High-performance Arabic character recognition," *Journal of Systems and Software* vol. 44 no.1 pp. 53-71
4. Amin A.,1997, "Off line Arabic character recognition – a survey-", *Proceedings of the 4th International Conference on Document Analysis and Recognition ICDAR* Vol. 2
5. Amin A., and G. Masini,1986, "Machine Recognition of Multi-fonts Printed Arabic Text," *Proc. 8th Inter. Conf. on Pattern Recognition*, (Paris), pp. 392—395.

6. Amin A., and H. B. Al-Sadoun,1992, "A New Segmentation Technique of Arabic Text," 11th IAPR, vol. 2, (The Hague), pp.441--445, Aug. 30-Sep. 3 .
7. Amin A., and W. Mansoon,1997, "Recognition of printed Arabic text using Neural networks", *Proc. 4th Int. Conf. On Document Analysis Recognition*, Ulm, Germany, August .
8. Amin A.,1998, "Off-line Arabic character recognition The State of the Art", *Pattern Recognition*, Vol. 31 No. 5, pp. 517-530.
9. Ben Amara, N., Belaïd, A.1998, "Modélisation pseudo bidimensionnelle pour la reconnaissance de chaînes de caractères arabes imprimés", in: Colloque International Francophone sur l'ECrit et le Documernt CIFED'98, Québec, Canada, p. 131-141.
10. Bushofa, B.F.M, Spann, M., 1997, "Segmentation and Recognition of Printed Arabic Characters using Structural Classification. *Image and Vision Computing,* 15, 167-179.
11. Chen, Y. K. and Wang, J. F.,2000, "Segmentation of Single –or Multiple- Touching Handwritten Numeral String Using Background and Foreground Analysis", *IEEE Trans. on PAMI* vol. 22, no.11 , 1304-1317
12. Cheung A., Bennamoun M., Bergmann N. W.,2001, "An Arabic optical character recognition system using recognition-based segmentation", *Pattern Recognition* 34 , 215-233
13. Chinveeraphan, S., Douniwa, K. and Sato, M.,1993, "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.
14. Chinveeraphan, S., Zidouri, A., and Sato, M.,1995, "Modified Minimum Covering Run Expression of Binary Document Images", *IEICE Trans. Inf. &. Syst*., vol. E78-D, no.4, pp.503--507, Apr.
15. Chinveeraphan, S., Zidouri, A., and Sato, M.,1994,"Stroke Representation by Modified MCR Expression as a Structural Feature for Recognition" *in Proc. IWFHR-IV, (Taipei)*, pp. 11--19, Dec. 7-9 .
16. Davis, L. S.,1986, "Two-Dimensional Shape Representation", Handbook of Pattern Recognition and Image Processing (ed. By T. Y. Yong and K. S. Fu) Academic Press, pp. 223-245.
17. Dehghan, M., K. Faez, M. Ahmadi, M. Shridhar,2001, "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM", *Pattern Recognition*, Vol. 34, pp. 1057-1065.
18. El-Dabi, S. S., Ramsis, R. and Kamel, A.1990,"Arabic Character Recognition System: A Statistical Approach for Recognizing Cursive Typewritten Text,", *Pattern Recognition*, Vol. 23 no.5, pp. 485-495.
19. Govindan, V. K. Shivaprasad, A. P.,23-7-1990, "Character recognition – a review, *Pattern Recognition,* 671-683
20. Impedovo, S. Ottaviano, L. and Occhinegro, S.,1991,"Optical Character Recognition – a survey", *Int. J. Pattern Recognition and Artificial Intelligence* 5, Jan., 1-24
21. Iwaki, O., Kida, H. and Arakawa, H., 1985,"A character /graphic segmentation method using neighborhood line density", *Trans. Of the Institute of Electronics and Communication Engineers of Japan*, Part IV, J68D, 4 ,821-828
22. Jain, A. K. and Bhattacharjee, S. K.,1992, "Text segmentation using Gabor filters for automatic document processing", *Machine Vision and Applications* 5, 3 , 169-184

Top

23. Mahjoub, M. A.,1996, “Choix des parameters lies a l’apprentissage dans la reconnaissance en ligne des caracteres arabes par les chaines de Markov caches, in *Forum de la Recherche en Informatique”*, Tunis, July . (French).
24. Mori, S., Suen, C. Y. and Yamamoto, K.,1992, “Historical review of OCR research and development”, *Proceeding of the IEEE* 80, 7 , 1029-1058
25. Namazi, M. and K. Faez,1996, “Recognition of multifont Farsi/Arabic characters using a fussy neural network,” *Proceedings of IEEE region 10 Annual International Conference* vol. 2 pp. 918-922 Nov. 1996
26. Sawaki, M. and N. Hagita,1998, “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.
27. Zidouri, A., Chinveeraphan, S., and Sato, M.,1994, “Arabic Document Image Understanding by MCR expression Method,” *in Proc. ICSS'94 Int. Conf. Signals and Systems*, (Algiers), pp.23--27, Sept.24—26.
28. Zidouri, A., Chinveeraphan, S., and Sato, M.,1995, “Structural Features by MCR Expression Applied to Printed Arabic Character Recognition” *in 8th Int. Conf. on Image Analysis and Processing}*, (San Remo Italy), pp.557--562, Sept. 13-15
29. Zidouri, A., 2001, “A Structural Description of Binary Document Images: Application for Arabic Character Recognition” *Proc. Int’l. Conf. On Image Science, Systems, and Technology*, (Las Vegas), vol. I, pp. 458—464, Jun. 25—28.

Top