Structural Features by MCR Expression for Printed Arabic Character Recognition

A.B.C. ZIDOURI, S. CHINVEERAPHAN, and M. SATO

Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori-ku, Yokohama, JAPAN 226

Abstract. This paper discribes how stroke features in document images are extracted and used for the recognition of printed Arabic characters. It is of importance to provide a good base representation that facilitate analysis and processing of document images. The strokes are extracted by a method called Minimum Covering Runs (MCR)[1]. This method of representing binary images by a minimum number of horizontal and vertical runs is used as a preprocessing step. The strokes are labeled and ordered, a feature space for the 100 shapes of the 28 Arabic characters is build. The system is under developement but the recognition rate obtained at this stage, 95.5% is encouraging.

1 Introduction

Text is a major part of information in document images, and automatically recognising the characters is not an easy problem. This is true especially for languages like Arabic which is written cursively (connected) even when printed. It is of importance to provide a good base representation that facilitate the processing and analysis of such huge amounts of information. Recognising well formed and neat characters in many languages have been solved for, and research is being focussed on more challenging problems of poor quality print omnifont or unconstraint handwriting. However, cursive writing like Arabic, where the segmentation problem of text into distinguishable characters is the source of many errors of several OCR, is still not yet fully explored. Structural features of basic patterns in document images such as characters or tables are mainly horizontal and vertical stroke components.

Our approach to Arabic cursive writing, using stroke components, is a novel one. It considers Arabic like any other stroke-like languages. This has been made possible by the structural features that MCR expression offers to represent text and tabular components in binary document images[2].

Several researchers have attempted to solve the problem 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[3]. A great amount of further effort is required in this domain.

Our method is similar to that of El-Dabi *et al.*,[4] in the sense it deals with the problem of separating the characters after they are recognised. This is because the segmentation process is not an aim by itself. However they proposed a system

Name	ALIF	BAA	TAA	THAA	JEEM	HAA	KHAA
Form_I	ا	ب	ت	ث	ج	ح	خ
Form_B		بـ	تـ	ثـ	جـ	حـ	خـ
Form_M		ـبـ	ـتـ	ـثـ	ـجـ	ـحـ	ـخـ
Form_E	ـا	ـب	ـت	ـث	ـج	ـح	ـخ
Name	DAAL	DHAL	RAA	ZAY	SAAD	DHAD	TTAA
Form_I	د	ذ	ر	ز	ص	ض	ط
Form_B					صـ	ضـ	طـ
Form_M					ـصـ	ـضـ	ـطـ
Form_E	ـد	ـذ	ـر	ـز	ـص	ـض	ـط
Name	ZHAA	KAAF	LAAM	MEEM	NOON	AYN	GHAYN
Form_I	ظ	ك	ل	م	ن	ع	غ
Form_B	ظـ	كـ	لـ	مـ	نـ	عـ	غـ
Form_M	ـظـ	ـكـ	ـلـ	ـمـ	ـنـ	ـعـ	ـغـ
Form_E	ـظ	ـك	ـل	ـم	ـن	ـع	ـغ
Name	FAA	QAAF	SEEN	SHEEN	HAH	WAW	YAA
Form_I	ف	ق	س	ش	ه	و	ي
Form_B	فـ	قـ	سـ	شـ	هـ		يـ
Form_M	ـفـ	ـقـ	ـسـ	ـشـ	ـهـ		ـيـ
Form_E	ـف	ـق	ـس	ـش	ـه	ـو	ـي

Fig. 1. The 100 shapes of the 28 Arabic character set in their different forms: Beginning, Middle, End or Isolated Form.

for typed Arabic text, which involves a statistical approach using accumulative invariant moments as identifier. Invariant moments are very sensitive to slight changes in a character shape. They reported a 94% recognition rate and a speed of 10.6 characters/minute. In our case the characters are segmented once their composing parts are successfully recognised. Although our system is also font dependant, nevertheless we achieved a better recognition rate, at a speed of recognition faster of at least one order of magnitude. Our test set is composed of 10 documents of 500 characters each.

2 Review of MCR Expression

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. Using this fact, 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

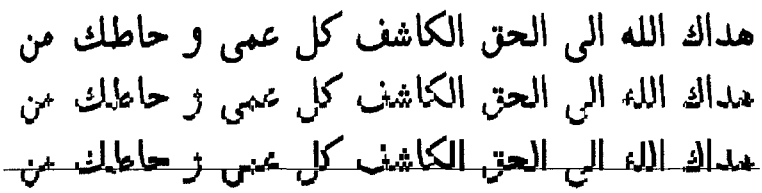

Fig. 2. A part of Text with detected baseline, its non-overlapping strokes extracted by modified MCR expression, and approximate strokes for visualisation.

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. This in turn is the same as finding the maximum matching, which has been solved for by graph theoritical algorithms. The modified version of MCR that we are using, which extracts strokes more accurately, is based on 2 essential procedures, a local stroke analysis used to find elongated segments and stroke like patterns in a binary image and a maximum matching constructing algorithm in the corresponding graph to find the remaining parts. The technique can partition characters and lines in images into horizontal and vertical segments. The use of horizontal and vertical terms in our case is very flexible as all the binary patterns are classified either as horizontal or vertical strokes according to a stroke decision criteria. Horizontal strokes are represented by sets of adjacent horizontal covering runs while vertical strokes are represented by vertical covering runs.

Then a description of characters in forms of strokes (parts) connections is constructed. An example is shown in fig.2 of a part of an Arabic text with detected baseline, the non-overlapping parts extracted by modified MCR, and approximate strokes for visualisation.

3 Characteristics of Arabic Writing

The characteristics of the Arabic language do not allow direct implementation of many algorithms used for other languages. Segmentation of words into characters to be recognised is still the most difficult and source of errors of OCR systems that are developed for cursive writing.

Arabic is written cursively whether printed or handwritten, from right to left, with an alphabet set of 28 characters, which take up to four shapes each, depending on their position. Fig.2 shows the set of the Arabic characters in their different forms within a word: beginning form, middle form, end form or isolated form. Seven characters do not allow for connection from the left side and have only two shapes each.

One of the main characteristics of Arabic writing is that it is written following a baseline. This is where almost all characters are connected to each other for most of the fonts and writing styles. This fact of character connection by some

short horizontal lines or strokes is used to detect what we call baseline. 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. Using this information and the fact that most Arabic characters themselves have their horizontal part written on this baseline, we detect the line containing the largest number of horizontal strokes and label it as baseline fig.3.

4 Features Extraction and Character Reference Building

Modified MCR gives a good structural information to provide the dynamic information from a static image, needed for character recognition. After the MCR expression is obtained for a binary image the non overlapping parts of a pattern or a character are labelled and ordered in a top down, left to right priority, to follow the Arabic way of writing to allow for future eventual connection to a speech synthesiser machine after the recognition is accomplished. The writing line or baseline is detected and a feature vector is then associated with each character, it takes into account the following parameters:

1. Number of parts constituting the character.
2. Size of the parts (length).
3. Position with respect to the baseline.
4. Sequence or order information.
5. Type of strokes (horizontal or vertical).
6. Direction or angle information.
7. Width of parts.

It must be noted that not all these parameters are needed at the same time to identify a specific character. These are used only in some ambiguous cases of multi-response. Fig.4 shows an Arabic word and the different parts parameters.

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 others are a lower zone below, and a middle and upper zones above the baseline. Two levels have been chosen above the baseline for the reason that the main information of the Arabic text is contained in the upper part of the main body of each character[5]. Also most stress marks like dots are above the main body of the character (four times as many as those which have dots below).

5 Segmentation And Recognition

In the framework of structural approaches to OCR most methods are based on representation, feature extraction, description and classification. For cursive writing it is necessary to overcome the complicated problem of letter separation.

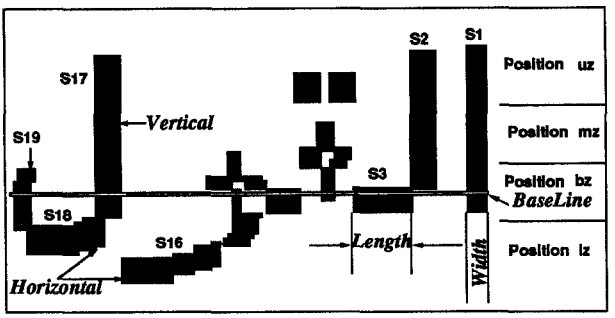

Fig. 3. An Arabic word and its decomposition into parts showing the features used

In our novel approach with MCR representation we are able to segment words or subwords into characters at the same time the recognition is achieved. The document image is first scanned at 216 by 216 resolution and input to the system in the CCITT format size. Then the MCR expression of image is found and the features extracted. The recognition is performed by matching to the reference prototypes build for this purpose. A reference model is built for the 100 shapes of the Arabic characters in their different positions, in the beginning of, within, or at the end of a word or subword, or when it is in its isolated form. Some additional prototypes were necessary for the special ligatures and special characters or stress marks like LAAM-ALIF (ﻻ), TAA-MARBOOTA (ۃ) or HAMZA (ء). The prototypes are made flexible to account for variations in the character shape due to noise from printing processes or quantisation noise of the scanning device. These reference characters were build from observation of the patterns obtained by MCR expression for Arabic printed text documents.

The feature vector built from the character pseudo-segmentation into parts provide the dynamic information from the static image, needed for characters recognition. The recognition is performed from left to right to be in a natural way to allow for an eventual link to a speech synthesiser machine. The recognition rate for the model document is more than 99%. The recognition drops however to just above 95% for the test documents at this early stage, but most of the errors are due to characters being rejected not misrecognised. These being mainly connected characters, as most isolated characters are recognised successfuly. We use a threshold as a unit length that is the size of a dot. From experimental results we found this threshold to be $th = 1/6$ the length of character ALIF. We call this a *baselength*. This is used for zoning of lines of text into 4 horizontal zones, as mentioned previously, and for the parameter size of the strokes. The text is segmented into words or subwords and this in turn separated into characters. Arabic characters take a special shape at the end of a word or a subword. This is used to segment text into words or subwords. Unfortunately there is no way to tell a word from a subword which could be composed of as few as one character unless high level recognition is sought as the use of a dictionary or a spell checker.

The segmentation of words into characters is automatically done after a character is recognised. There is still room to train the system and allow for the prototypes to cover more variations in the shapes of the characters. Most of the errors are now due to characters which are very similar and differ only by one dot like FAA and QAAF or TAA and THAA, and when character MEEM or AYN are within a word or subword. These are represented in some cases by only 2 strokes which makes the discrimination very difficult. The limitation of the system now is that it assumes the same font is used for all the text to be recognised and to be the same as the modelling font. Although a low level recognition scheme, we achieved a reasonably fast recognition speed of more than 10 characters/sec, for a test set of more than 5000 characters. There will be always a trade off between accuracy and effectiveness of a system and its cost and complexity. Up to now we have trained our system only to one type of font. We believe that there is plenty of room for improvement and work to be done.

6 Conclusion

In this paper we have shown that MCR provides a good representation of stroke components in text. As an application we used it to printed Arabic cursive writing to overcome the difficult problem of segmentation inherent to all cursive writings. We labelled the extracted strokes and rendered them meaningful as structural features for recognition and segmentation. The system is in progress and at this stage we achieved a minimum recognition rate of 95.5% for a test set of 10 documents of about 500 character each of the same font as the training set.

References

1. Chinveeraphan, S., Douniwa, K., and Sato, M., *"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.
2. Chinveeraphan, S., Zidouri, A., and Sato, M., *"Stroke Representation by Modified MCR Expression as a Structural Feature for Recognition,"* IWFHR-IV, (Taipei), pp. 11–19, Dec. 7-9, 1994.
3. Senior A. W., and Fallside F., *"Using Constrained Snakes for Feature Spotting in Off-line Cursive Script,"* Proceedings of the Second Int. Conf. on Document Analysis and Recognition (Tsukuba Japan), pp.305–310, Oct. 20-22, 1993.
4. El-Dabi, S. S., Ramsis, R., and Kamel, A., *"Arabic Character Recognition System: A Statistical Approach for Recognising Cursive Typewritten Text,"* Pattern Recognition, vol. 23, no.5, pp. 485–495, 1990.
5. Margner, V., *"SARAT - A System for the Recognition of Arabic Printed Text,"* 11th IAPR, vol. 2, (The Hague), pp. 561–564, Aug. 30-Sep. 3 1992.