# ADAPTIVE DISSECTION BASED SUBWORD SEGMENTATION OF PRINTED ARABIC TEXT

A. Zidouri
*Department of Electrical Engineering*
*King Fahd University of Petroleum and Minerals*
*Dhahran 31261, Saudi Arabia*
Email: malek@kfupm.edu.sa

M. Sarfraz, S. A. Shahab and S. M. Jafri
*Department of Information and Computer Science*
*King Fahd University of Petroleum and Minerals*
*Dhahran 31261, Saudi Arabia.*
Email: {sarfraz,sadnan,shomaail}@ccse.kfupm.edu.sa

## ABSTRACT

Numerous segmentation and recognition techniques have been proposed in literature for Arabic OCR system. Correct and efficient segmentation of Arabic text into characters is considered to be a fundamental problem. While OCR systems for other languages do not need segmentation for printed text for successful recognition, it is essential to design robust and powerful segmentation algorithms or employ segmentation free recognition schemes for printed Arabic text. Even more, in recognition of handwritten characters, segmentation is considered to be indispensable. Most of current segmentation technique suffers from over segmentation and under segmentation in addition to not being adaptive in nature. In this paper, we have proposed a new sub-word segmentation scheme, which is independent of font size and font type.

**Keywords:** *Arabic Character Recognition, Word Segmentation.*

## 1. INTRODUCTION

Printed Arabic text is like handwritten Latin text, such that connection of characters is an inherent property for Arabic script whether it is typed, printed or handwritten. Most of errors and deficiencies of Arabic recognition systems comes from the segmentation stages.

Various segmentation algorithms have been proposed in the literature. Given the vast number of papers published on OCR, it is impossible to include all the segmentation methods in this survey. Instead we have made a representation selection to illustrate the different principles that can be used.

## 2. RELATEDWORK

Amin in his paper titled Offline Character Recognition [5,6] says that recognition of Arabic characters follows two approaches. *An* Analytical approach wherein the word or sub-word is segmented into characters, Recognized and then combined again to form the Word and a Global approach where in the recognition is performed on the whole representation of words. Whereas the former approach involves segmentation, the latter approach avoids the segmentation stage.

Zidouri et al in [13] proposed the ORAN System (Offline Recognition of Arabic Characters and Numerals) based on Modified Minimum Covering Run Expression (MCR) [8]. Modified MCR has the ability to represent the binary patterns into horizontal and vertical parts referred to as strokes. Their method is independent of Segmentation of Text to characters. The main difficulty associated in Cursive text recognition is the segmentation of words to characters. In their approach they overcome the problem of segmentation by pseudo-Segmentation of characters to small strokes. The obtained strokes are further divided into Overlapping parts and non-Overlapping parts. The Non Overlapping parts are labeled and ordered and a set of features are selected. The geometrical and topological features of the strokes are used for detecting the baseline and for feature extraction. Next matching of candidate character shapes to reference prototypes is performed. The training procedure is carried out to improve the recognition rate of the system. Recognition is achieved by simple left to right analysis and matching to reference prototypes. However the limitation is that, the system is designed for single font and characters

1550-6037/05 *$20.00* © *2005* IEEE

to be recognized are of the same font as the training set.

Almuallim and Yamaguchi [2] proposed a structural recognition technique for Arabic Handwritten words. Their system consists of 4 Phases: Preprocessing the text, Segmentation of words into strokes, classification of strokes using their geometrical and topological properties and combining the strokes into the strings of characters that represents the recognized word. Since it is difficult to segment a cursive word to letters, words are segmented into separate strokes and classified as complementary characters, strokes with a loop and strokes without a loop . These strokes are further classified using their geometrical and Topological properties. Finally the relative positions of the classified strokes are examined and the strokes are combined in several steps into the string of characters that represents the recognized word. System failures in most cases are due to incorrect segmentation of words.

Al-Yousefi and Udpa [4] introduced a Statistical Approach for the Recognition of Arabic Characters. The character is segmented to Primary and secondary parts. Secondary parts are then isolated and identified separately and moments of horizontal and vertical projections of the primary parts are computed. Finally a Statistical classification approach is used to classify the characters.

Segmentation can also be achieved using closed Contour. SARAT System [12] used outer contours to segment Arabic words to characters. The word is divided to a series of curves by determining their start and endpoints of words. Whenever outer contour changes sign from positive to negative a character is segmented.

Amin and Al-sodoun [3,7] adopted a multifont technique for segmenting Arabic text. The Technique is divided to **4** major steps. First the image is scanned and is transformed to a binary image. Next the image is preprocessed to remove any noise associated with the image. In the third stage the skeleton of the image is traced from left to right using a **3** x 3 window and a binary tree is constructed. Freeman code is constructed to describe skeleton code. Finally the binary tree is segmented to subtree such that each subtree describes a character in the binary image. This technique overcomes the problem of under segmentation and over segmentation of the traditional segmentation method that is based on baseline segmentation technique. It also overcomes the difficulty of placing the segmentation points within the characters.

Hamami and Berkani in [11] employed a simple segmentation method. Their method is based on the observation of projections of pixel rows and columns, and uses these projections to guide the segmentation process. They start by segmenting the image horizontally into separate lines. Then, each line is segmented vertically to separate parts\words. The previous two steps are mainly achieved through observing the rows or columns that have no black pixels. Afterwards, the baseline is detected and it is scanned in order to estimate the start and end points of separate characters. Like in their method, under-segmentation, which is a common problem shown in **Figure 1,** is treated by considering the entire under-segmented set **of** characters as a single character. On the other hand, the other common problem of over-segmentation is solved by taking care of the situations in which it may occur and resolve them accordingly.

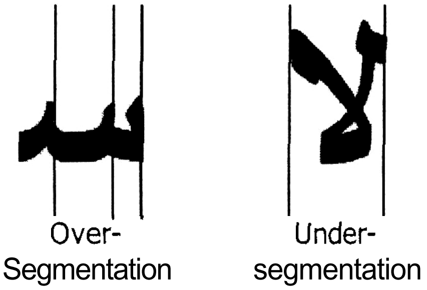


**Figure 1** Common Problems in Arabic Character Segmentation

However this technique fails for Overlapping Characters. As shown in the **Figure 2** when the two characters overlap the algorithm fails.

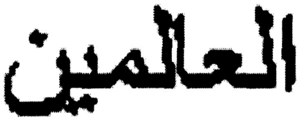


**Figure 2** Overlapping **of** two Characters

Finally Badr and Haralick [1] describe the design and implementation of a system that recognizes machine printed Arabic words without prior segmentation. The technique is based on describing symbols in terms of shape primitives. At recognition time, the primitives are detected on a word image using mathematical morphology operations. The system then matches the detected primitives with symbol models. This leads to spatial arrangements of the matched symbol models. The system conducts a search in the space of spatial arrangements of models and outputs the arrangement with the highest probability as the recognition of the word.

Other segmentation techniques in the literature include the segmentation of text to words and then to characters based on zoning technique [9], Construction of a binary tree by tracing the skeleton of the given image and segmentation of the binary tree to subtree such that each subtree represents a character image [3], stroke segmentation [10] where Arabic words are segmented to primary strokes and secondary strokes.

## 3. PROPOSED APPROACH FOR SUB-WORD SEGMENTATION

In our project for Arabic OCR system, after applying preprocessing techniques, horizontal and vertical segmentations are employed to segment a page into separate lines and lines into sub-words respectively. Arabic characters are connected on the baseline. In order to dissect sub-word into characters, we exploit the fact that, junction point between connected characters lies at baseline as shown in fig. 2.2. But in certain cases like case of an isolated “ ”, it will suffer from "over segmentation", in other cases, some words may have overlapping characters such as “ "and thus suffer from "under segmentation" as in Fig. 2.1. Even some segmentation techniques, might work for one font but fail to segment words if font type or size is changed. We have developed a general technique, which is independent of font size and font type.

Consider the following notation:

$\Theta$ = Width of single dot in the document.
$L_s$ = Width of smallest character
$L_s'$ = Width of two smallest character if appear together
$L_m$ = Maximum Width of character in isolated form
$B[x, y]$ = Location of Baseline
I = Image of subword
I'= Image of subword without dots
E= Empty image of size I.

To improve segmentation efficiency we opted to remove stress marks like dots from characters. Their original position is remembered and reintroduced only in the recognition phase. In order to remove dots from subword, we have employed connected component approach with **8** neighbors.

**Steps in character segmentation**

1. Skeletonize I'
2. Scan from right to left in row-wise fashion, to find **a** band of horizontal pixels having length >= $L_s$
3. Take vertical projection on the scanned band found in step2. If no pixel is encountered, draw a vertical guide band on **E.**
4. Use special mark for the guide bands which are drawn due to the scanned band (found in step2) below the baseline $B[x, y]$.
5. Repeat the procedure, for all the rows.

After performing above mentioned steps, an image E with several guide bands is obtained. In order to select, correct guide band for sub-word dissection, we extract several features from each guide band:

| Feature | Description |
|---|---|
| F1 | Width of the guide band |
| F2 | Distance from 1st predecessor from right, zero in case of 1st guide band |
| F3 | Distance from 2nd predecessor from right, zero in case of 1st and 2nd guide band. |
| F4 | 1 if guide band drawn due to scanned band is above baseline<br>0 if guide band drawn due to scanned band is below baseline |
| F5 | Midpoint of guide band |

The judicious selection of guide band is driven through several rules. The feature sets {F1...F5) of each guide band are tested for each rule. If it satisfies rules then it is selected otherwise it is rejected.

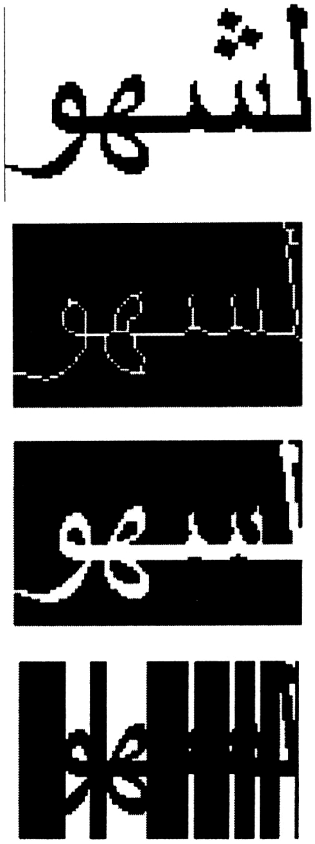
**Rule#1:** Choose guide band having highest relative width (F 1) and F4 = 1
**Rule#2:** Choose guide band if F2 > L, and F4 = 1
**Rule#3:** Choose guide band if F2 <= Ls and F3 > $L_s$' and guide band is not the last one.
**Rule#4:** Choose guide band if F1 >= $L_m$ and F4 = 1

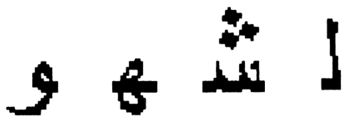
For the 1st guide band in the sets, even if it fails to qualify **Rule#1-4** and the guide band next to it satisfies **Rule#2** then it should be selected.

If all the guide bands fail to satisfy any rule, then we apply less constrained rule base i.e. removing F4 condition except **Rule#4**.

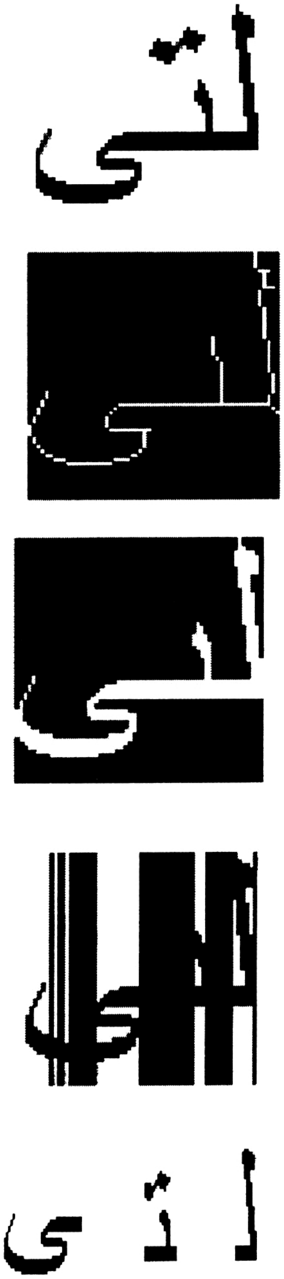
Example:

**1. Sub-word with dots and character "ش"**





**2. Sub-word with dots**



## 4. DISCUSSION

This segmentation scheme has been implemented in a MATLAB environment. Results are quite promising. Few points are still under investigation:

1. The problem of character overlapping for some Arabic fonts causes some under segmentation of some characters like the special character " " composed of "ل' and "ا"

2. The problem of ligature for *Arabic traditional font* generates also under segmented characters.

These problems can be solved by considering some group of characters that always appear together in a special form, as a separate class. Some other miss-segmentation will be dealt with in the recognition stage. The segmentation is not an aim by itself; some characters can be classified in a first run during the segmentation process by simple matching. Arabic words are sometimes composed of groups of connected and non connected portions that we refer to as sub-words. Sub-words can **be** composed of one or more characters. So segmentation of sub-words is applied only to those sub-words that are composed of more than one character..

## 5. CONCLUSION

In this paper an efficient segmentation of printed Arabic text into characters is considered. **As** segmentation and recognition are closely dependent of each other, and segmentation is not an aim by itself, our approach defers the ligature problem to the recognition phase. In this paper, we have proposed a new sub-word segmentation scheme, which is independent of font size and font type.

**Acknowledgements**

***This project is funded by King Fahd University of Petroleum & Minerals under Project # EE/AUTO-TEXT/232.***

## 6. REFERENCES

[1] B. Al – Badr and Robert M. Haralick, "Segmentation free Word Recognition with application to Arabic", Intelligent Systems Laboratory, FT –10, Univ. of Washington, U.S.A.

[2] H. Al-muallim and S. Yamaguchi, "A method for recognition of Arabic Cursive Handwriting", IEEE Trans. Pattern Anal. Mach. Intelligence, PAMI – 9, pp.715-722, (1987).

[3] H.B. Al – Sadoun and A. Amin, "A new Structural technique for recognizing printed Arabic Text", Int. J. Pattern Recognition Artificial Intelligence., vol.9, pp.101- 125, (1995).

[4] H. Al –Yousefi and S.S. Udpa, " Recognition of Arabic Characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.853-857, (1992).

[5] A. Amin, "Off line Arabic Character Recognition – A Survey", IEEE, pp. 596-599, (1997).

[6] A. Amin, "Off-Line Arabic Character Recognition system: State of the Art", Pattern Recognition, Vol. 31, No. **5**, pp 517-530, (1998).

[7] A. Amin and H. Al- Sadoun, "A segmentation Technique of Arabic Text", Proc. 11th International Conference on Pattern Recognition., The Netherlands, pp 441-445, (1992).

[8] S. Chinveerphan and A. B.C. Zidouri and M Sato, "Modified MCR Expression using Binary Document Images", IEICE Trans. Information & System., Vol. E78 –D, No. 4, pp. 503-507, (April 1995).

[9] M. Fakir, M.M Hassani and C. Sodeyama , "Recognition of Arabic Characters using Karhunen – Loeve Transform and Dynamic Programming", IEEE International Conference on System Man and Cybernetics 12-13-14, Japan, pp. 868 –873 ,(October 1999).

[10] H. Goraine, M. Usher and S. Al-Emami, " Offline- Arabic Character Recognition", University of Reading, Dept. of Cybernetics, Reading, Berkshire, UK, pp. 71-73,(1992).

[11] L. Hamami and D. Berkani, "Recognition System for Printed Multi-Font and Multi-Size Arabic Characters", Arabian Journal for Science and Engineering, Vol 27, Number 1B, 57-72 (April 2002)

[12] V. Margner, SARAT – A system for the recognition of Arabic Printed Text, Proc. 11th Int. Conf. on Pattern Recognition, pp. 561-564, (1992).

[13] A, Zidouri, S. Chinveerphan, and M.Sato "Recognition of Machine Printed Arabic Characters and Numerials Based on MCR." IEICE Trans. Information & System., Vol. E78 –D, No. **12 ,pp.** 1649-1655, (December 1995).