International Journal of Computer
(IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Designing an Arabic Handwritten Segmentation System

Mohamed E. M. Musa[a]*, Bodoor A. Bashir[b], Mohamed N. I. Ismail[c]

[a,b]*Sudan University of Science and Technology. Khartoum, Sudan, hafiz@sustech.edu*
[c]*College of Science, King Faisal University, Al-Ahsa, Saudi Arabia*
[a]*Email: hafiz@sustech.edu*
[c]*Email: mismail@kfu.edu.sa*

Abstract

The greatest difficulty facing the recognition of Arabic handwritten words is segmentation, because Arabic handwriting is cursive with complex multi-form styles. Hence, intensive research efforts are needed to reach an effective Arabic handwriting segmentation system. This paper presents a system which uses morphological features of the Arabic characters for segmentation. The proposed system segments non-overlapped (horizontally connected -e.g. "حسن") as well as overlapped (vertically connected - e.g. "نجد") characters. The result is not very good one. However, it arrives at good directives for more research. As the writing was freely without any restrictions, both over-segmentation and under-segmentation problems affect the system.

Keywords: Arabic pattern recognition; word segmentation; morphological features.

1. Introduction

Optical Character Recognition (OCR) is a branch of Pattern Recognition, and it is one of the active research topics in the last four decades of the last century and the first decade of this century. This area is concerned with enabling the computer to recognize the input data in the form of images, which is either handwritten or printed text, then converting this data into electronic form. The goal is to enable the computer to read and thus to get rid of the keyboard as a data entry tool. Designing and implementing Arabic recognition system is one of the major challenges that face Arabic recognition researchers.

---

* Corresponding author.

Therefore, this area is an active research area [1]. The appearance of new results of the research in this area is still going on [2]. One of the main reasons for the difficulty of the Arabic writing recognition is its cursive script.

Among the problems of the Arabic writing recognition systems is the large and different types of Arabic fonts, most of the research presented in this area put some constraints that govern font shape and size and deal with all fonts in the same way [3].

Other problems are the (Over segmentation) and (Under segmentation) problems which appears in some research papers [1, 4, 5]. In the Over segmentation problem a single letter is segmented into more than one segment as in segmenting Sheen (شـ) into Noon (نـ) and Ta (تـ). In the under segmentation problem, two connected letters may be similar in shape to a single letter that is because they are Overlapped letters, Example: The shape of the connected letters Noon and Meem (نمـ) resembles the shape of the Gheen letter (غـ).

The aim of this study is to design an Arabic Handwriting Segmentation System. And the type of writing intended here is a freely writing without any restrictions, where the proposed system deals with words that contain overlapped and non-overlapped letters. The suggested system is based on an algorithm provided by (Muhktar et al) [1].

The organization of the paper is as follows: section II includes an overall summary of the characteristics that characterized the Arabic language, and explain the process of the segmentation. Section III, explains the dataset used in this paper. Section IV, explains the proposed system and its performance. Section V, shows the experiments and results. Section VI contains the conclusion and future works

2. Arabic Word Segmentation

The Arabic script is cursive; this property makes the segmentation of words into characters very difficult. In addition, the Arabic word may consist of one connected part, as in Mohammed "محمد" or a number of connected parts as in the word Hamid "حامد", where it consists of two parts, "حا" and "مد". Some Arabic words may consist of overlapping letters for example; the characters Noon "نـ" and Jeem "جـ" are overlapped in the word "نجد ". One of popular methods for the Arabic word segmentation is the Vertical projection Method [6]. This method is illustrated in Figure (1).

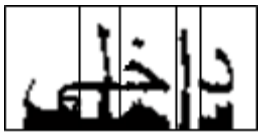


Figure 1: The Vertical projection Method

When applying this method to printed script it gives very high recognition rates [5]. However, its performance on handwritten is poor. This is due to the different forms of writing from one person to another, as well as the fact that some letters are spilited into parts. In addition, this method has big problem with overlapped characters.

One of the previous studies in this topic is the study presented by (El-Gowely et al) [5]. This study developed a special technique for the recognition of the printed Arabic script on multiple fonts using the Vertical projection technique. The reported recognition rate of this work is 94%.

Another promising algorithm Muhktar et al algorithm's called "Off-line handwritten Arabic Character Segmentation Algorithm (ACSA)" [1]. Here is a summary for their algorithm:

- Pre-processing of the image.

- Find the baseline of the word and then separate the word into main parts that include the body of the word and secondary parts that include dots and Hamzas.

-Segmenting the word into Sub-words, so that each part is a set of connected characters.

- Extraction of the Morphological features of the Pixels.

- Analysis of the image into three zones: Upper zone, Median zone and Lower zone depending on the Base line and determining Threshold for the width of the Median zone.

-determine the upper letter The (Ascenders) and the (Descenders) letters. See Figure (2)

- Find points on the (Local Minima (LM)) in the (Outer contour) of the word. A group of laws are applied on these points to determine whether these points separate between the characters properly or not. If all the laws are achieved on a specific point then the point is accepted as a separator point between two characters, otherwise it is rejected.

This algorithm achieved a ratio of 86% for correct segmentation of the words into characters and a ratio of 9% for under segmentation. Under segmentation means that two overlapped characters are segmented as one character, as shown in Figure (3). A ratio 5% is went into Over segmentation. This appears in the words that contain either of the letters Seen "سـ" or Sheen "شـ", in such case the letter is segmented into more than one part. These experiments were applied on a small unpublished data set consists of 100 handwritten words. This data set is collected by the authors to test the algorithm.

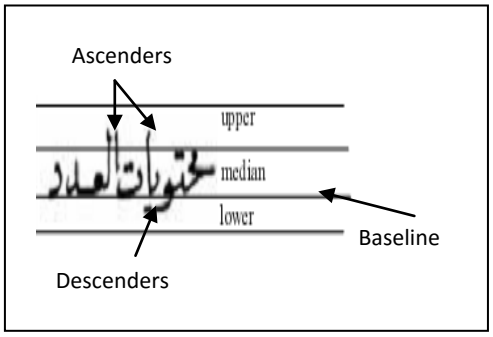


Figure 2: The three areas of the image and the baseline and the top characters (Ascenders) and subscripts (Descenders)

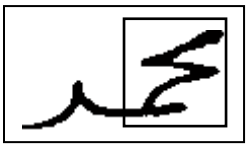


Figure 3: Example of overlapping characters

3. SUST-ARG names dataset

The Dataset used in this paper, SUST-ARG, is a dataset designed and collected by the Pattern Recognition Research Group in the College of Computer Science and Information Technology, Sudan University of Science and Technology. It is a local application form for graduates. The form contains the necessary data needed to obtain the graduation certificate, and the data include the quadruple name of the applicant in Arabic and English.

4. The proposed word Segmentation

The proposed words Segmentation system consists of two main phases, namely:

1. preprocessing phase. 2. Segmentation phase.

The Segmentation Phase consists of several sub-stages, as explained on the following sections.

4.1. Segmentation of the word into sub-word

Each word is segmented into parts such that each connected part of the word represents specific component. The process of segmenting the word into parts goes through the following phases:

a. Consider each connected part of the word as a component, even dots and Hamzas are taken as separate component.
b. Include Dots and Hamzas with their relevant components. Figures (4) and (5) show the steps of segmenting the word into parts.

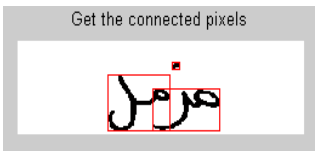


Figure 4: Finding connected part

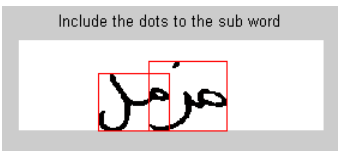


Figure 5: Include Dots with their relevant components

4.2. Sub-words' features extraction

The thinning process of the image is performed before the feature extraction to make thickness of the lines of the image one pixel. Feature extraction process for each part of the word is performed in two phases:
a. Feature extraction for each pixel.

b. Feature extraction for each letter-candidate

The pixel features are:

- ♣ Start point.
- ♣ End point.
- ♣ Branch point.
- ♣ Loop point.
- ♣ Connection point

An algorithm was designed to find the Start point of writing and steps of the algorithm is as follows:

1. Find the baseline of the word.
2. Find the first point with a value of 0 while searching from right to left in the image and from top to the baseline of the word.
3. Test the point resulting from step 2, if it falls within a loop shape then the values of its coordinates is etained and the type of starting point is considered to be a Loop point. Otherwise go to step
4. Starting from the point resulting from step 2, we should pass through the writing curve. When we move to any point, the feature is checked, if the feature is end point stop.

The process of Feature extraction for letter-candidate is done by identifying the letter as Ascender or Descended letter, and this based on the baseline. The baseline is calculated using the Horizontal projection method of the image. See Figure (6)

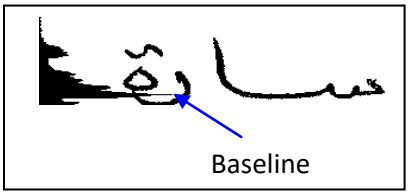


Figure 6: Calculating the baseline using Horizontal projection

4.3. Segmenting sub-word into characters

For the algorithm used in the segmentation process we have taken advantage of the algorithm offered by Mukhtar et al [1] which was mentioned before, with the some amendments. One of the most important amendments that were added to the algorithm is an algorithm for the segmentation of (Overlapped characters) which is explained in the next section.

4.4. Overlapped characters segmentation algorithm

From the dataset used in this study, it was noted that most of the Overlapped characters were: "جـ ، حـ ، خـ ، مـ" and the letters "جـ ، حـ ، خـ" have the same form in writing when they overlap with other characters, see Figure (7). It was also noted that the letter "مـ" has the same form in writing when it overlaps with other characters, as Figure (8) shows.

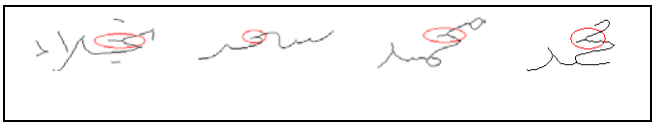


Figure 7: Examples of which the letters "جـ ، حـ"appeared overlapping with other letters.

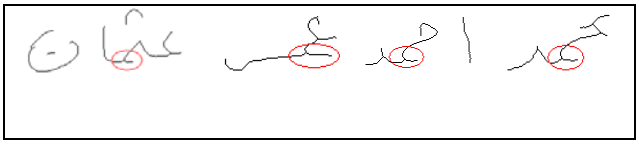


Figure 8: Some words in which the letter "مـ" appeared overlapping with other letters.

This extension is based on matching the form of letter when overlapping, as Figures (7) and (8) illustrate with other parts of word. In the case of matching, The method is determining the accepted points for segmentation which depends on the form of the letter when overlapping so that:
• If the form of the character when overlapping is similar to the form of one of the letters "جـ ، حـ ، خـ", which was illustrated in Figure (7), only one point is determined to be accepted as illustrated in Figure (9).
• If the form of the character when overlapping is similar to the form of the letter "مـ", which was illustrated in Figure (8), two points were determined to be acceptable as illustrated in Figure (10).

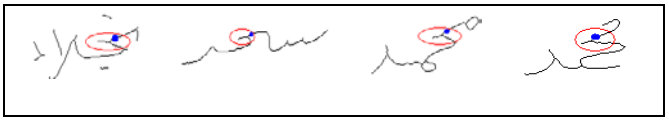


Figure 9: Determining the Acceptable point of the words that show one of the letters "جـ ، حـ ، خـ"overlapping

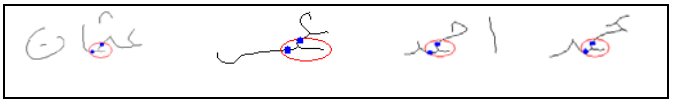


Figure 10: Determining the Acceptable points of the words that show the letter "مـ" overlapping

5. Experiments and results

Experiments were conducted on 287 words chosen from the used data set. Overall results are shown in Table (1) which shows the number of words that have been tested, the total number of Sub-words and the sum of their letters, as well as the results of the performance of the proposed system, as illustrated in the number of Sub-words and the number of segmented characters.

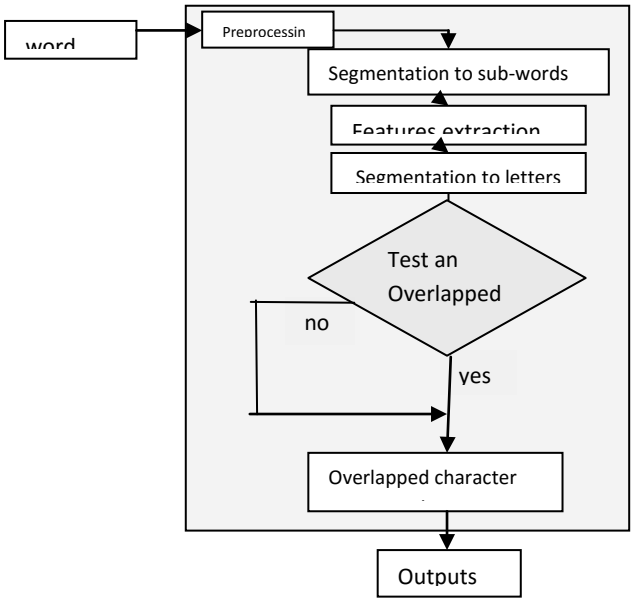


Figure 11: The general structure of the Words Segmentation Proposed System.

Table 1: Results of the performance of the proposed algorithm

	Number of words	number of (Sub-words)	number of letters
Ground truth	287	450	1159
Performance of the proposed algorithm	287	450	784
Success rate	-	100%	67.64%

Tables (2) and (3) illustrate the performance of the proposed system for the words that contain overlapping and non-overlapping letters, respectively. Table (4) shows some of the samples that were used testing the system. These samples include words that contain overlapping and non-overlapping characters.

The system was tested on words containing the overlapped characters which most of them are the words "أحمد", "محمد" and "عمر" . If we exclude the word "محمد" the success rate of the proposed algorithm becomes 83.41%. The words "أحمد" and "عمر" contain three connected letters and the second letter meem "ـمـ" is the overlapped character. Therefore, if this letter is successfully segmented, the word would successfully be segmented into

characters. With regard to the word "محمد" , the proposed system has been tested by 79 words including 316 characters and only 159 characters where successfully segmented by the system and thus the success rate of proposed algorithm in this case was 50.31% and this figure would mean that this algorithm fails to the segmentation of 157 characters. The reason for this result is due to the prevalence of the word "محمد" and therefore the multiplicity of its writing ways.

Table 2: Results of the performance of the proposed system for the words that contain Overlapped characters

	number of words	number of (Sub-words)	number of letters
Original numbers	137	190	548
Performance of the proposed algorithm	-	190	353
Success rate	-	100%	64.41%

Table 3: Results of the performance of the proposed system for the words that contain Non-overlapped characters

	number of words	number of (Sub-words)	number of letters
Original numbers	150	260	611
Performance of the proposed algorithm	-	260	431
Success rate	-	100%	70.54%

Table 4: Results of the performance of the proposed algorithm on the words that contain Overlapped characters and it is limited in the words "أحمد","محمد" and "عمر"

	number of samples	number of characters		Success rate
		Original	Segmented	
أحمد	5	15	13	86.66%
عمر	46	184	153	83.15%
محمد	79	316	159	50.31%

Table (4) the performance results of the proposed algorithm on the words "أحمد", "محمد" and "عمر ". Figure (12)

shows some samples of the word "محمد" that the algorithm Under segment them, on the other hand, Figure (13) shows Over segmentation ones.

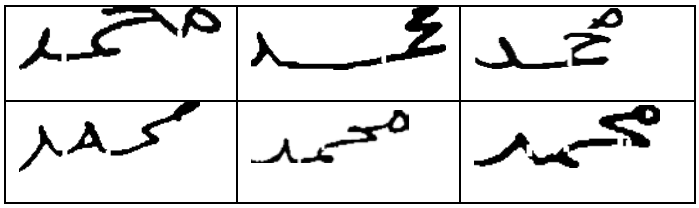


Figure 12: Some samples of the word "محمد" that the proposed algorithm failed to segments

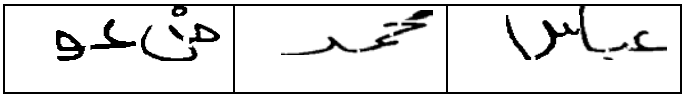
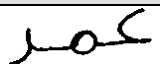
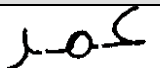
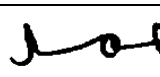
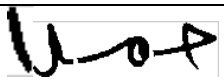
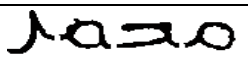
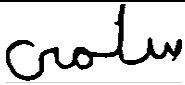
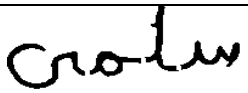
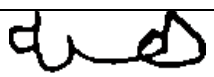
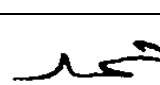
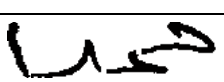
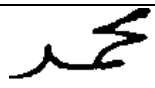
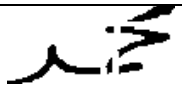
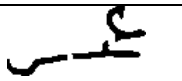


Figure 13: Some words which have been wrongly segmented or over-segmented

Table 5: Some of the words that have been tested on the proposed algorithm and these words contain the Overlapped and Non-overlapped characters

			Segmented Letters					
#	**Inputs**	**Outputs**	1	2	3	4	5	6
1								
2								
3								
4								
5								
6								

7								
8								
9								

6. Conclusion

In this paper, a new system for the segmentation of the Arabic Handwritten words was presented. The idea of the proposed algorithm is based on the algorithm proposed by Mukhtar et al [1] which are based on outer contour of the image and the Morphological features. The most important amendments to Mukhtar et al algorithm is an extension for dealing with overlapped letters. The proposed system has a segmentation rate of 67%, according to the used dataset. The writing method of data set is free without any restriction on the type and size of the font.

It was noted that when testing the system on the word "محمد" written on an Overlapped basis the system gives the segmentation rate of 50%

Part of the suggested future work of this study is to conduct post-processing phase, including: gathering parts of the characters resulting from over segmentation. Since there are some small parts that result from the segmentation process which are parts of the same character as shown in Table (5) in the word No. 7, where word "محمد" segmented into 6 characters as the third part is part of the letter "حـ "and the Fifth part is part of the Second meem "مـ". The idea here is to enter many combined small parts to a probabilistic classifier, which will rate this combined parts and give us the chance to find the correct one which is expected to have high probabilistic rate.

References

[1] Sari, Souici and Sellami, "Off-Line Handwritten Arabic Character Segmentation Algorithm: ACSA ", Proceeding of the eighth International workshop on frontiers in handwriting recognition, (2002).

[2] Abuhaiba, "A Discrete Arabic script for Better Automatic Document Understanding ", the Arabian Journal for Science and Engineering, (April 2003).

[3] Touj, Amara and Amiri, "Two Approaches for Arabic Scrip recognition-based Segmentation Using the Hough Transform ", Document Analysis and Recognition, Volume 2, Page (s): 654 - 658, (2007).

[4] Ayman Mohammad Bahaa Eldeen Sadeq, "Intelligent Neural System for Character Recognition", A

Thesis Submitted in Partial Fulfillment of the Requirements of the Degree of Master of Science in Electrical Engineering (Computer & Systems), (1999).

[5] Bushofa and Spann, "Segmentation and recognition of Arabic Characters by Structural Classification ", Image and Vision Computing, (1997).

[6] Fakir, Hassani and Sodeyama, "On the Recognition of