An Approach to Separate Overlapped Characters of Arabic Script

Amir S. AL-Malah
AL-Mustansiriya University, College of Science, Computer Science Department, Baghdad, Iraq
E-mail: AL-Malahasmy@yahoo. Com
Mobile: 07902244945

Abstract:

This paper presents a new method on separate the overlapped character of handwritten and printed Arabic script. The method depends on vertical scan and region growing method, region growing procedure that starts with a set of seed pixels. The aim is to grow a uniform and connected region from each seed. Applied to cursive Arabic script, where ligatures, overlap and style variation pose challenge to the recognition system.

Keywords: Arabic character segmentation, optical character recognition (OCR), overlapped Arabic character.

1. Introduction:

Optical character recognition (OCR) is a process of automatic computer recognition of characters in optically scanned and degitized pages of text. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical application potentials. It can contribute immensely to the advancement of automation process and can improve the interface between man and machine in many applications. Some practical application potentials of OCR system are: (1) reading aid for the blind, (2) automatic text entry into the computer for disk top publication, library cataloging, ledgering, …etc. (3) automatic reading for sorting of postal mail, bank cheques and other documents (4) document data compression: from document image to ASCII forma. (5) Language processing (6) multi-media system design, …etc.

Depending on versatility, robustness and efficiency, the commercial OCR systems can be divided into four generations:

The first generation systems can be characterized by the constrained letter shapes which the OCRs read.

The next generation is characterized by the recognition capabilities of a set of regular machine printed characters as well as hand printer characters. At the early stages, the scope was restricted to numerals only. The third generation can be characterized by the OCR of poor print quality characters, and hand-printed characters for a large category character set.
The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, table and mathematical symbols, unconstrained hand-written characters, color document, low-quality noisy documents like photocopy and fax,…etc.

At present, more sophisticated optical readers are available for Roman Chinese, Japanese and Arabic text [2]. These readers can process documents which has been typewritten, typeset, or printed by dot-matrix, line and laser printers. They can recognize characters with different fonts and sizes as well as different formats including inter mixed text and graphics. With the introduction of narrow range scanners, measuring 3 to 6 in wide, columnar scanning is now possible. With these scanners an optical reader can recognize multiple columns or sections of page or mailing lists.

Some are equipped with software for spell checking, and for flagging suspicious characters or words [1].

2. ***Arabic character recognition [2, 4, 5, 6]:***

Arabic language provides a rich source of technical challenges for recognition and transliterration algorithms The most obvious characteristics of the Arabic language is that Arabic scripts are inherently cursive; writing isolated characters in “block letters” is an unacceptable and unused writing style.

Arabic is written from right to left, Arabic text (machine printed or hand written) is cursive in general and Arabic letters are normally connected on the base line. This feature of connectivity will be shown to be important in the segmentation process. Some machine printed and hand written texts are not cursive, but most Arabic texts are, and thus it is not surprising that the recognition rate of Arabic characters is lower than that of disconnected characters such as printed English.

The shape of the letter is context sensitive, depending on its location within a word. For example, a letter as ‘ع’ has four different shapes: isolated ‘ع’, beginning ‘عـ’, middle ‘ـعـ’ and end ‘ـع’. Certain character combinations form new overlapped shapes which are often font dependent. Some overlapped involve vertical stacking of characters, see Fig. 1. Since

not all letters connect, word baundary location becomes an interesting problem, as spacing may separate not only words but also certain characters within a word.

In handwritten and decorative styles usually include vertical combinations of short strokes called *ligatures* this feature makes it difficult to determine the boundaries of the characters. Furthermore, characters of the same font have different sizes (i.e. characters may have different widths even though the two characters have the same font and point size). Hence, word segmentation based on a fixed width cannot be applied to Arabic.

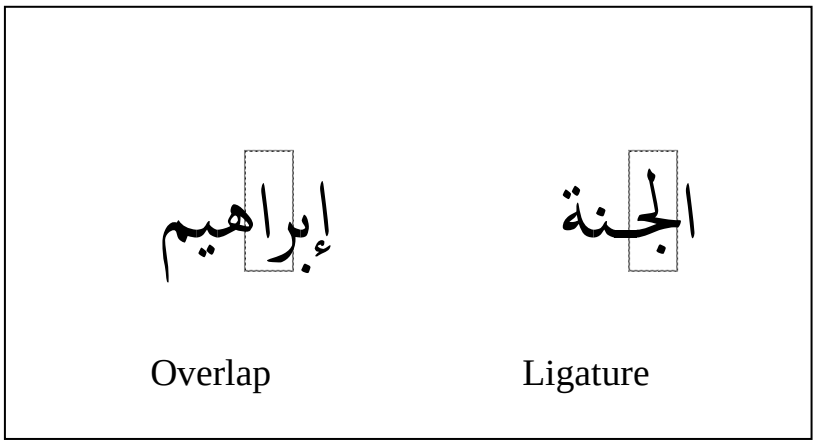

Fig. (1): Sample of Arabic writing in Rokaa contain ligature and overlap

Arabic fonts like any other font have certain feature. These are, the shape feature, which have complex shapes compared to English characters. They may consist of many "strokes" their range from one to four. Another feature of Arabic characters is its connectivity and can have multiple connections for each character depending on its position within a word.

All the above features make recognition of Arabic characters more difficult than others.

3. Properties of Arabic scripts [3]:

There are 28 characters in the Arabic alphabet. Each character has two to four different forms which depend on its position in the word or sub words.

An Arabic word can have one or more sub words, e.g., there are three sub words in the world shown in Fig. 2. Most characters have dot (s) or zigzag (s) associated with the character and this can be above, bellow, or

inside the character. Many characters have a similar shape. The position or number of these secondary strokes makes the only difference, e.g., BA, TA, and THA. We have also noticed that Arabic words may horizontally overlap and characters may stack on others. These induce problems for both the word and the character segmentations. Fig. 1, 3 demonstrate an example of word overlapping, the dotted box is where the overlapping occurs. At this stage, it is not hard to understand that segmentation is a crucial step in the development of an Arabic OCR system.

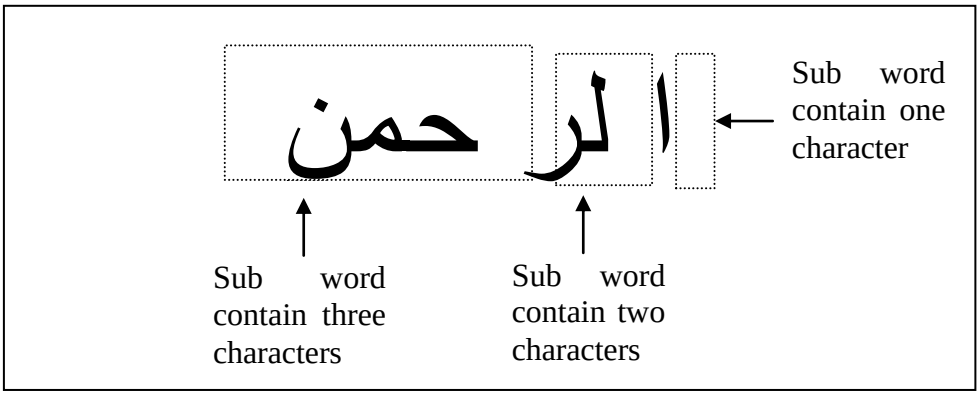

Fig. (2): Word divided into three sub word

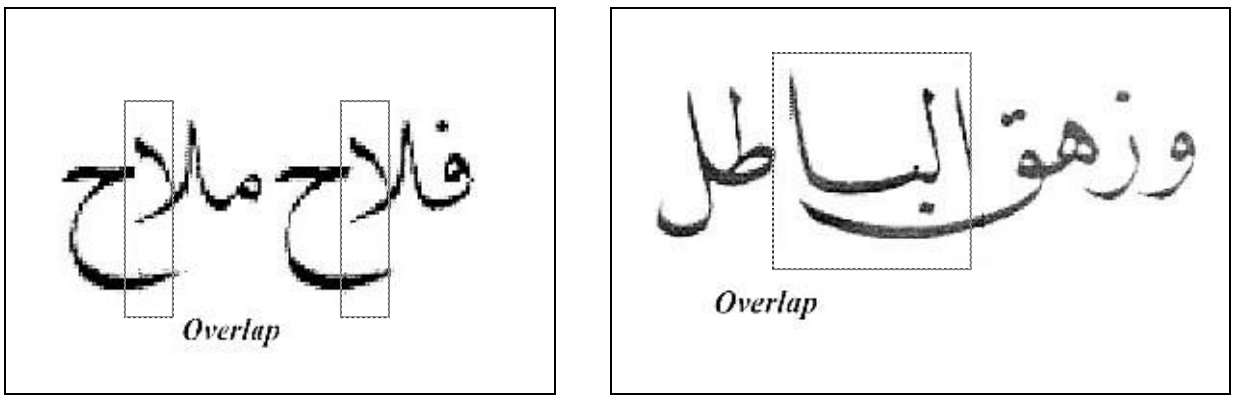

Fig. (3): An example of overlapping Arabic word

4. An overview of the proposed system:

To simplify the segmentation stage we must make attention to the problem of ligature and overlapped characters, we use the method of regain growing to solve the overlapped characters.

Regain growing is bottom-up procedure that starts with a set of seed pixels. The aim is to grow a uniform and connector regions from each seed.

A pixel is added to a region if and only if:

It has not been assigned to another region.
It is neighbor of that region.
The new region created by addition of the new pixel is stile uniform.

Algorithm:

Let ***f*** be an image, and R1, R2, ….Rn a set of regions each consisting of a single seed pixel.

$$\text{point} = \frac{\Sigma \text{pixel Ri}}{\Sigma \text{pixel in thinning RTi}}$$

avpoint = av X point
Repeat
For i = 1…..n
For each pixel ***p*** at the boarder of ***Ri***
For all neighbors of ***p***
Let ***x, y*** be the neighbor's coordinates
Let ***mi*** be the mean gray level of pixels in ***Ri***
If the neighbor is unassigned and $|f(x, y) - mi| <= D$
Add neighbor to ***Ri***, update ***mi***
Until no more pixels are being assigned to regions
If ***Ri*** > ***avpoint*** then put it in new dimensions
If ***Ri*** < ***avpoint*** then put it in previous dimensions

The word image is introduced to the system as a matrix and we threshold it to a blank pixels (foreground) and white pixels (the background).

We scan the image from top – down, right – left, and when we detect the blank pixel we use it as a start point, search for the 8 – neighbors for each point in the region, we stopped when no more pixels satisfy the

criteria for inclusion in that region. We mentioned to the size of region and we compared it to the size of point (point, double point, hamza) if it's size is greater then put it array (new plan), else put it in previous array (plan).

Every point after we register we delete it and make a new scan to detect other region until no more region found.

5. conclusion:

This paper presents a new technique for separating overlapped Arabic characters. The algorithm resulted in 92% for separate overlapped by using region growing method. This method is work on Arabic overlapped sub words and this overlapped sub words does not intersect with each other.

References:

1. U. Pal, B. B. Chaudhuri, Indian script character recognition: a survey.
2. A.Amin, off line Arabic character-recognition: the state of the art, pattern recognition 31 (1998) 517-530.
3. A.Cheung, M.Bennamoun,N.W.Bergmann, An Arabic optical character recognition using recognition-based segmentation, Pattren Recognition 34(2001) 215-233.
4. M.S Khorsheed, Recognising handwritten Arabic manuscripts using a signal hidden Markov model, pattern recognition letters, 24(2003) 2235-2242.
5. Emad J.Mohammed, "Arabic Character Recognition Methodlogy Using Complex Moments", M.Sc. Thesis,comp.Dept.the University of Technology, 1995.
6. A.Amin, Recognition of printed Arabic text based on global features and decision tree learnuing techniques, pattren recognition 33(2000) 1309-1323.