Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 115 – 121

Procedia Computer Science

The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Segmentation of Arabic handwritten text to lines

MOKHTARI Younes[a], Yousfi Abdellah[b]

[a]*Equipe TSE, ENSIAS, Université Mohammed V, Rabat- Maroc*
[b]*Equipe ERADIASS, FSJES, Université Mohammed V, Rabat- Maroc*

Abstract

Automatic recognition of writing is among the most important axes in the NLP (Natural language processing). Several entities of different areas demonstrated the need in recognition of handwritten Arabic characters; particularly banks check processing, post office for the automation of mail sorting, the insurance for the treatment of forms and many other industries. One of the most important operations in a handwriting recognition system is segmentation. Segmentation of handwritten text is a necessary step in the development of a system of automatic writing recognition. Its goal is to try to extract all areas of the lines of the text, and this operation is made difficult in the case of handwriting, by the presence of irregular gaps or overlap between lines and fluctuations of the guidance of scripture to the horizontal. In this paper, we have developed three approaches of handwritten Arabic text segmentation, then we compared between these three approaches.

Keywords—Image processing; Handwritten; Segmentation; histogram; Windows slippery.

1. Introduction

The handwriting recognition system is a technology that allows us to convert different types of documents (such as scanned paper documents, digital photos, etc.) into modifiable and exploitable formats.
In literature, there are two types of handwriting recognition:

- Static recognition also called offline recognition treats handwritten texts, in this context it is impossible to know how the different patterns were drawn.
- Dynamic recognition also called online recognition; it's kind of monitoring the writing, the lying and the raising of the electronic pen on a surface or a tab. 5

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
doi:10.1016/j.procs.2015.12.056

Handwriting recognition is generally done in five steps: pre-treatment, segmentation, learning, recognition and correcting recognition errors. In this paper, we interest of segmentation steps.

In literature, several approaches have been proposed in the field of handwriting segmentation. For example:

- BENSLIMANE and ZAROUNI used the partial differential equations (PDE) in order to detect the contours in the old documents. This method has been tested on more 100 manuscripts of different structures, mono, multi- oriented and multiscripts of different qualities .this approach can detect 2092 lines, with a detection rate of 83.2%. [3]
- ZAHOUR, TACONET and RAMDANE trimmed the Arabic manuscript text into 8 columns and horizontally projected each column. This method conducted over a dozen texts shows early encouraging results.
- BENNASRI and ZAHOUR proposed a method to extract the lines of Arabic manuscript text without any constraint for the writer. After detecting the starting points of all lines by a partial projection, he proceeds into monitoring the partial outline of each line, this method gives a rate of 98% 1. The drawback of these approaches is that the execution time is very long.

2. Segmentation

Segmentation is one of the most important steps in any handwriting recognition system; there exist two types of segmentations:

- Global segmentation: It allows us to isolate words of the documents and reference them by their position in the text. This type requires wide lexis so that the recognition could function properly. [5]
- Local segmentation, it exceeds the limits of the global approach, its goal is to make a segmentation into characters or graphemes, and this type gives better recognition rate, but the rate of implementation remains larger. [7]

In literature, there exist several segmentation techniques:

- segmentation by the contour, it rests only on filters and some thresholding techniques for the detection of objects and some homogenous regions, the weak point of this technique is the execution time that is very large.
- Segmentation from the histograms , it remains the most used approach in the recognition systems due to its high speed of retrieving results , but it is sensitive to the problem of overlapping between adjacent lines
- Sliding windows: wiping is the principle of this technique; it is based on the division of the image and the choice of the window size in order to get a good result.[4]

The problem of Arabic handwriting segmentation is still difficult, and this is due to the problem difficulty of overlapping between the lines, derived from the inclination of the writing, and erroneous positions of points and diacritical marks above and below the characters.[8]

In this article, we propose three approaches to carry out the segmentation of Arabic manuscripts to lines, in order to solve these types of problems and to reduce the segmentation execution time. These approaches are:

- Method based on the horizontal projection and the classification of the lines according to their heights.
- Method based on cutting by a sliding column of a fixed size.
- Hybrid method based on the two previous ones.

1.1 Segmentation using line classification (first approach)

Our first approach is to make a segmentation of a handwritten document in Arabic, by combining between the horizontal projection (1) and classification of line heights. According to this method, the document acquired by the scanner contains three types of lines:

- Diacritical lines are lines with small heights; this approach eliminates these lines because it considers them a noise.
- The normal lines are lines with average heights; this approach calculates the average heights (ħ) of all lines of document.
- The overlapping lines are lines with great heights, they are the result of the overlapping of ascending and descending characters but also the pasted characters of the adjacent lines (see figure 1).

$$f(y) = \sum_{x=0}^{W} I(x, y) \tag{1}$$

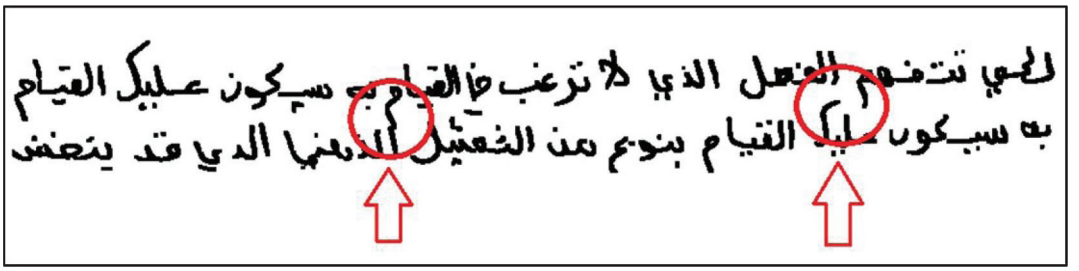

Fig. 1. Image representative of an overlap between the lines.

This approach determines the accurate number of lines in the overlapped block by dividing the block height to the exact height of a row (ħ) and then divides the block on the number of lines found.

The major drawback of this approach reside in case when lines do not have a fixed height, its mean lines contain simple characters with a small height. In this case, the system classify characters with diacritical lines and disturbance, thus lines contains characters with big size, will be classified with the overlapping lines. From the results obtained from this approach, we notice that there are another class that belongs to the normal class and influence on the segmentation process.

1.2 Segmentation by sliding window (2nd approach)

We have developed a method for segmentation by using fixed-size windows , developed by Mr. Abderrazak Zahour, Bruno Taconet et Said Ramdane, Their segmentation method is based on the document subdivision into 8 column, Then complete an horizontal projection on each column, all of that to provide blocks. [2]

Our method is based on the subdivision of the document into columns by using a sliding window of a fixed size.

This column has a size of one third, and moves by a step of one sixth of the width of the image, thus this approach splits the image into three different columns, and at each column performs a horizontal projection in order to detect the lines of the document, then it computes the average height to remove any lines having a height lower to one third of the average, finally it chooses the column with the largest number of detected lines, and after that it divides the acquired document by the number of lines found (Figure 2).

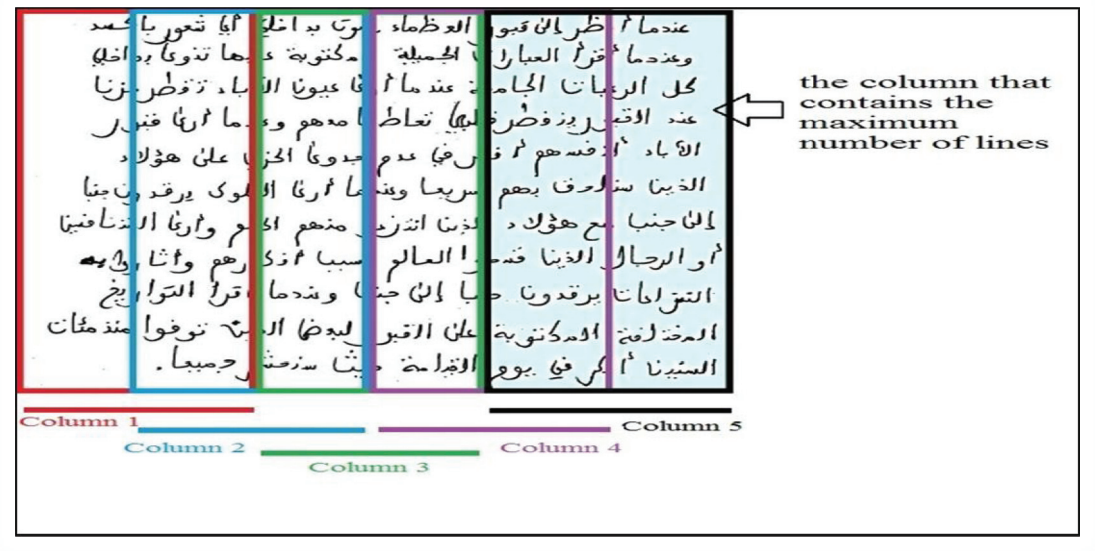

Fig. 2. Division of the sliding window.

1.3 Segmentation by hybrid method (3rd approach)

In this approach, we had the idea of combining between Approach 1 and Approach 2, we uses the first approach to classify lines to detect the overlapping zones, and the sliding window can only be applied only on the overlap zone. This will reduce the execution time of the third approach.

We See in this figure 3 that the handwritten document was segmented by green and red rectangles, the green rectangles are normal lines segmented by only an horizontal projection, but for red rectangles are the overlapping lines, we have treated it by the horizontal projection and the sliding window, the role of this window is the detection of hidden lines in red rectangles.

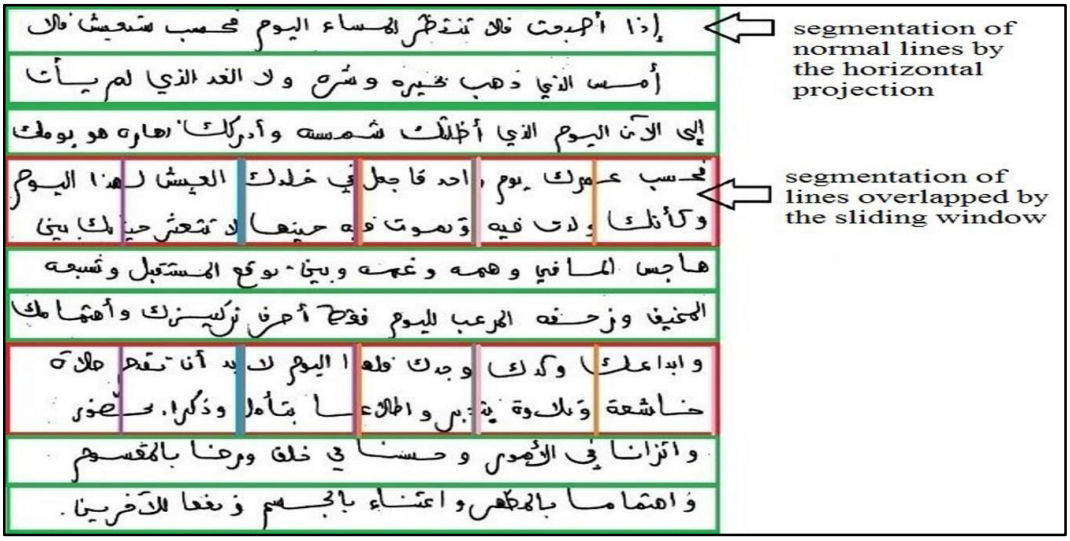

Fig. 3. Segmentation by the hybrid method

3. Test and results

To evaluate our approaches, we worked on a test corpus that consists of 90 written documents (average size 1.20 Mo 1000 lines) by different scripters on which different cases of overlapping lines are inserted. The obtained results for the three methods are presented on the table below:

Table 1: results of the three methods

Methods	Execution time (S)	Results (%)
Segmentation using line classification	8	76
Segmentation by sliding window	20	94
Segmentation by hybrid method	12	90

The first method has a minimum execution rate (8S) which increases the value of this method to the level of the execution time, but the results are encouraging (76%), this method gives good results when the handwriting is clear, and the vertical projection detects the three types of lines already cited.

The second method has higher execution time than the first (20S), among the strengths of this method is the detection rate (94%) and its ability to detect all lines of a non-inclined document, which contains many overlaps. From the results of the first two methods, we have thought of a third method that combines together the strengths of the last ones, and which has given us an average execution time (12S), in addition to a result of 90%.

These results show that our proposed third method can detect most of the baselines, with a higher execution time speed as compared with the first two ones.

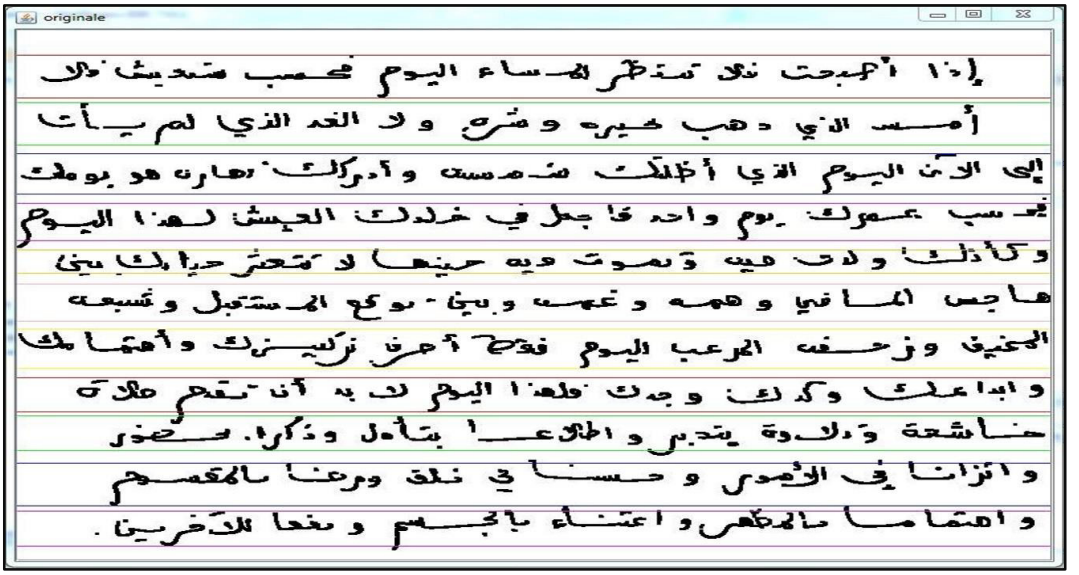

Figure 4: Segmentation of a handwritten text.

4. Conclusion

In this paper, we have reviewed the work done in segmentation of handwritten documents in Arabic, based on the horizontal projection and classification lines into three classes [4], as well as the division of the document with a sliding column of a fixed size. Based on these approaches, we have performed the two first segmentation methods and according to these obtained results, we have developed a third method of segmentation, this method is more efficient in terms of execution time and detection rate as compared to the first two approaches, which shows the importance of this method.

The major problem of these three methods is the inclination of the documents that influences on the results of segmentation, in our next work we will remedy to this problem in order to maximize the number of lines detected.

References

1. BENNASRI A, ZAHOUR. A, TACONET. B. Havre: computer laboratory. *The lines of Arabic manuscript text*.
2. ZAHOUR. A, BRUNO. T, RAMDANE. S, Team GED: University of Havre. *Contribution to the segmentation of ancient manuscripts.*
3. BENSLIMANE. R, ZAROUNI. I, Fes: Laboratory of Transmission and Information Processing USMBA, top school of technology fès. *Image segmentation by region based active contours*.
4. EL GAJOUI. K, ATAA ALLAH. F, OUMSIS. M, Rabat: Laboratory of Computing Research and Telecommunications Faculty of Science. *The OCR multilingual documents.*
5. EL QACIMY. B, Ait Kerroum. M, Hammouch. A, RABAT : GTI-LGE, ENSIAS, University Mohamed V Souissi, Rabat. *Automatic recognition of characters Arabic manuscripts in off-line mode.*
6. BOUKERMA. H, MESLATI. D, ANNABA: Faculty of Engineering Department of Computer Science Doctoral School East - Pole ANNABA. *Combining classifiers for recognition of handwritten Arabic script.*
7. BENOUARETH. A, ENNAJI. A, SELLAMI. M, ANNABA: Laboratory Research Informatics Institute Computer-University Badji Mokhtar – Annaba, Algeria. *Recognition of Words Arab Manuscripts by combination of a Global Approach and Analytical Approach.*

8. AL-HAJJ. R, MOKBELL. C, LIBAN: Balamand University, Faculty of Engineering P.O.Box 100 Tripoli. *Recognition of Arabic cursive combining classifiers in MMCs facing windows.*
9. MOKHTARI. Y, YOUSFI. A. MOROCCO: Workshop on Codes, Cryptography and Communication Systems, AL JADIDA, *Segmentation of Arabic handwritten text to lines*.