Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Text Line Segmentation of Arabic Handwritten Documents using Line Height Method

Mokhtar Abdulrahman Mohammed, M. Ravi Kumar, R. Pradeep
Department of Computer Science
India

Abstract: ***The research on offline handwritten Arabic character recognition has received more and more attention in recent years, because of the increasing need of Arabic document digitization, Segmentation of text lines is one of the important steps in the Optical Character Recognition System. Segmentation is pre-processing step of word and character segmentation. Text Line Segmentation can be viewed simply for handwritten documents which contains distinct spaces between the lines and it is more complex for the documents where text lines are overlapped, touch, curvilinear and variation of space between text lines and skewed documents. In line segmentation, errors are generated due to touching and overlapping character occurrences. Sometimes, interline space and noises make line segmentation a difficult task. In this paper I proposed segment text line handwritten for Arabic language based on the computing the height of lines an image using Line Height Method.***

Keywords: Line Height Method, Optical Character Recognition, Segmentation.

I. INTRODUCTION

Text line segmentation of hand written for Ancient and historical documents, printed or handwritten, strongly differ from the documents (newspapers, scientific journals, magazines, and business letters) because layout formatting requirements were looser. Their physical structure is thus harder to extract. In addition, historical documents are of low quality, due to aging or faint typing. They include various disturbing elements such as holes, spots, writing from the verso appearing on the recto, ornamentation, or seals. Handwritten pages include narrow spaced lines with overlapping and touching components. Characters and words have unusual and varying shapes, depending on the writer, the period and the place concerned. The vocabulary is also large and may include unusual names and words. Full text recognition is in most cases not yet available, except for printed documents for which dedicated OCR can be developed.

Page segmentation into text lines is performed in most tasks mentioned above and overall performance strongly relies on the quality of this process. Pre-processing of document images (gray level, color or black and white) is often necessary before text line extracting to prune superfluous information (non textual elements, textual elements from the verso) or to correctly binarize the image.

We survey the different approaches to segment the clean image into text lines. Taxonomy is proposed, listed as line height, projection profiles, smearing, grouping, Hough-based, repulsive-attractive network and stochastic methods. The majority of these techniques have been developed for the projects on historical documents. Script segmentation of a document images is a very important task for an Optical character recognition (OCR). A lot of research work has been done for script or character segmentation for character recognition of Arabic language script. In any OCR system segmentation phase is a very important step to improve accuracy of printed Arabic language or hand character recognition system in optical character recognition (OCR) heavily depends upon segmentation phase. The terms segmentation means subdivides an images into a particular part (like text or graph separation) its constituent region or object. Basically in the segmentation technique that minces, I tried to extract a specific part (text line based on text line height) of the script document images.

II. RELATED WORK

Some of the schemes that are reported in the recent works for character recognition in Arabic language as follows:

(Xiu, Peng et al. 2006): A new probabilistic segmentation model is proposed. First, a contour-based over-segmentation method is conducted, cutting the word image into graphemes. The graphemes are sorted into 3 queues, which are character main parts, sub-parts (diacritics) above or below main parts respectively. The confidence for each character is calculated by the probabilistic model, taking recognizer output, geometric confidence and logical constraint into consideration. Then, the global optimization is conducted to find optimal cutting path, taking weighted average of character confidences as objective function. Experiments on handwritten Arabic documents with various writing styles show the proposed method is effective [1].

(Abdullah 2007): This research seeks to fulfil the following objectives to overcome the problems of the segmentation such as overlapping, and slanting.

Also To advance the segmentation and recognition of the handwritten Arabic script by providing a wide range of handwriting styles in a database to be used as a test bench. To facilitate the study of the off-line Arabic characters segmentation problem by categorizing and analyzing the segmentation methods. The main contribution of the research is the new algorithm for the off-line handwritten Arabic characters segmentation using Slant-Tolerant Segment Feature (STSF). The research also contributes to the field by introducing the following:

1. An adequate Arabic handwritten database for the experiments.
2. A categorization system for the segmentation methods in which the Segmentation methods are categorized into two approaches: Junction- Seeking Approach (JSA) and Recognize-Segment Approach (RSA)[2].

(Wshah, Shi et al. 2009): have propose a new algorithm for segmentation of off-line handwritten Arabic words. The algorithm segments the connected letters to smaller segments each of which contains no more than three letters. Each letter may be segmented to at most five pieces. In addition to improving the recognition of Arabic words, another potential application of the proposed segmentation method is to build lexicon of small size, consisting of no more than three letter combinations. Generally, it is very hard to generate lexicon for recognition of unconstraint handwritten Arabic documents due to the large number of words of Arabic language [3].

(Kumar, Abd-Almageed et al. 2010): We present a novel graph-based method for extracting handwritten text lines in monochromatic Arabic document images. Our approach consists of two steps-Coarse text line estimation using primary components which define the line and assignment of diacritic components which are more difficult to associate with a given line. We first estimate local orientation at each primary component to build a sparse similarity graph. We then, use a shortest path algorithm to compute similarities between non-neighboring components. From this graph, we obtain coarse text lines using two estimates obtained from Affinity propagation and Breadth-first search. In the second step, we assign secondary components to each text line [4].

(Boussellaa, Zahour et al. 2010): This paper presents a new method for automatic text-line extraction from Arabic historical handwritten documents presenting an overlapping and multi-touching characters problems. Our approach is based on block covering analysis using unsupervised technique. This algorithm performs firstly a statistical block analysis which computes the optimal number of document decomposition into vertical strips. Then, our algorithm achieves a fuzzy base line detection using fuzzy C means algorithm. Finally, blocks are assigned to its corresponding lines. Experiment results show that the proposed method achieves high accuracy about 95% for detecting text lines in Arabic historical handwritten document images written with different scripts [5].

III. CHALLENGES

If the line spacing between lines is very small, it is difficult to segment the lines.If the document having the presence of seeping ink from the other side of the document make image preprocessing particularly difficult and produce binarization errors.

The segmentation problem of Arabic handwriting is considered a challengeable task due to different styles of handwriting and the connectivity of the Arabic letters as shown Fig. 1.

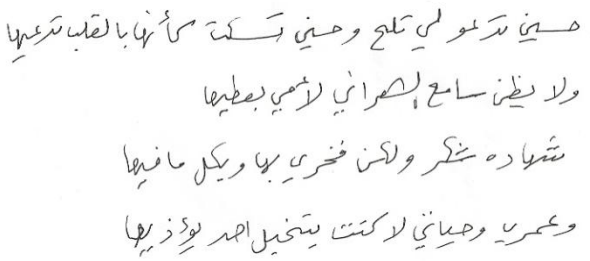

Fig. 1: Connectivity of handwritten Arabic characters.

The challenge also sometime one word in Arabic language written by different users but has the skew, orientation and styles are different as shown Fig. 2.

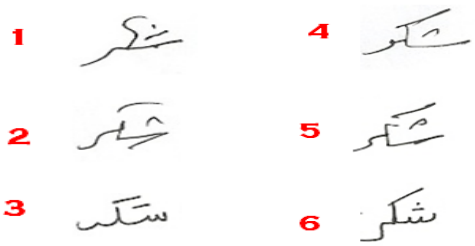

Fig. 2: One word written by different users.

IV. MOTIVATION

Hand writeable data is any day better to face to recognition. It reduces complexities in reading, understanding and helps in simplify the data which were written years before, these data can be scanned with the help of a scanner and is converted into hand Writeable format which is a challenging task in present day and thus in turn the whole process employs less time and labour.

Today there are many OCRs already developed for the prestigious English language and other European languages. There are OCRs also available for some of the Asian languages like Japanese, Chinese etc. so this lead us to propose a system for recognition of regional Arabic language (segmentation handwritten lines).

V. PROPOSED METHOD

The LHM is an algorithm used for segment text line handwritten for Arabic language based by computing the height of lines an image, use count variable for compute lines or rows for text line.

The LHM(Line Height Method) checks and searches about first point in the image and store in variable, after storing it will check for the bottom black point until get space and segment text line by drawing lines as shown Fig. 3.

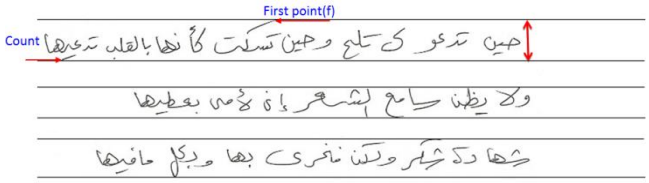

Fig. 3: First point and Count.

By this we can calculate distance between starting black point pixel and the end of black point pixel of text line by use variable count (height) each lines in image, so that we can differentiate the lines in handwritten image as shown Fig. 4.

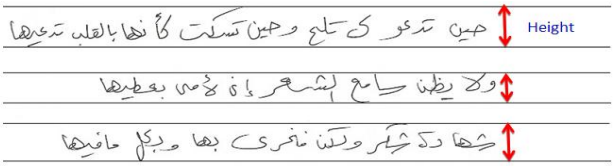

Fig. 4: Height of lines in Arabic language.

VI. INE HEIGHT METHOD (PROPOSED DIAGRAM)

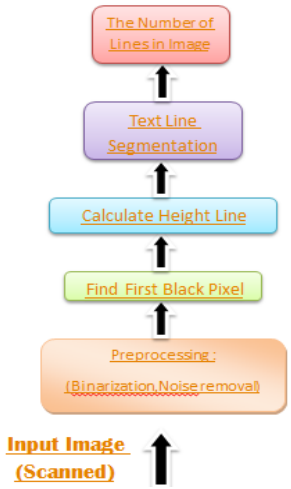

Fig. 5: LHM Diagram.

VII. RESULTS AND DISCUSSION

Results are presented to illustrate the accuracy and efficiency of the proposed method.

The scanned images were then analyzed and then segment their text lines are shown in Table. 1. Moreover, it is a real-time process. The below table illustrates to see the result for the determination of first point (f) and height (count) each lines in image:

Table. 1: First point and Height of lines.

No. of Lines	First point(f)	Height line (count)
1	9	52
2	83	50
3	159	52
4	232	49
5	308	56
6	379	51
7	452	60
8	537	51
9	603	57
10	678	57
11	757	58

The Fig. 6 illustrate the Convergence between the values of heights of lines and also illustrate increased the values of first point of lines in image:

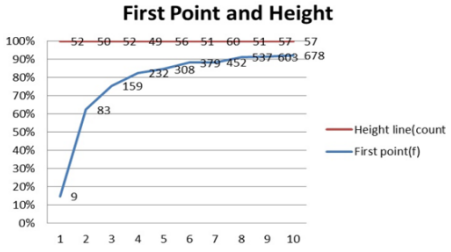

Fig. 6: Convergence values of first point and height (count).

VIII. RESULTS IMAGE

This is the original document. We apply convert to gray level and binary process Fig. 7:

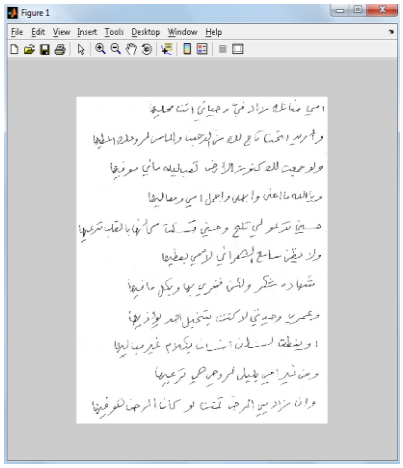

Fig. 7: Input Image.

This is the binarized image in Fig. 8:

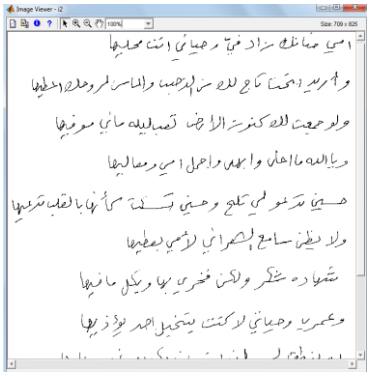

Fig. 8: Binarized Image

This is the segmented line from the binarized Image in Fig. 9:

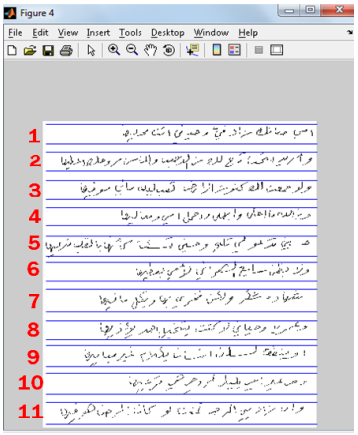

Fig. 9: Segmented Line.

The results obtained from segmentation process, the simple image of handwritten Arabic language was found to be successful. The main advantage of our method is fast. These designed algorithm will strongly support for the segmentation of text line of complex image in future.

IX. CONCLUSION

The intention of my paper is to how I can segment text line of hand written in Arabic language by computing the height of lines in an image, and to define a better algorithm for line segmentation and toward the accurate results.

The LHM(Line Height Method) designed finds the first point or black pixel by f variable of first text line and stop and move to next line, then calculated the height of all lines in same text line that contain the black pixels by count variable until get space, subsequently now segment text line by drawing line under the text line, and repeated same process for all text lines.

REFERENCES

[1] Xiu, P., L. Peng, et al. (2006). Offline handwritten arabic character segmentation with probabilistic model. Document Analysis Systems VII, Springer**:** 402-412.

[2] Abdullah, S. A. (2007). Off-Line Handwritten Arabic Characters Segmentation Using Slant-Tolerant Segment Features (STSF)[PJ6123. S562 2007 f rb], Universiti Sains Malaysia.

[3] Wshah, S., Z. Shi, et al. (2009). Segmentation of Arabic handwriting based on both contour and skeleton segmentation. Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, IEEE.

[4] Kumar, J., W. Abd-Almageed, et al. (2010). Handwritten arabic text line segmentation using affinity propagation. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, ACM.

[5] Boussellaa, W., A. Zahour, et al. (2010). Unsupervised block covering analysis for text-line segmentation of Arabic ancient handwritten document images. Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE.

