Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

A Skew Resistant Method for Persian Text Segmentation

Sajad Shirali-Shahreza
Computer Eng. Dept.
Sharif University of Tech.
Tehran, Iran
shirali@ce.sharif.edu

M. T. Manzuri-Shalmani
Computer Eng. Dept.
Sharif University of Tech.
Tehran, Iran
manzuri@sharif.edu

M. Hassan Shirali-Shahreza
Computer Eng. Dept.
Yazd University
Yazd, Iran
hshirali@yazduni.ac.ir

Abstract

Using OCR programs is one of the best ways to convert written and printed documents into digital form. The first phase in OCR is segmenting the input image and identifying text and non-text regions. This paper proposes a new method for segmentation of Persian printed texts which is based on the Ink Spread Effect. Considering that the Persian scripts are very different from the English script, most methods proposed for the English script have not rendered good results for the Persian scripts. The method proposed in this paper has been designed considering the special features of the Persian scripts. In addition, one of the most important characteristics of this method is resistance to skew. Moreover, the proposed approach is directly applicable to Arabic scripts.

Keywords*: Page Segmentation, Optical Character Recognition (OCR), Persian Document, Ink Spread Effect.*

1. Introduction

Nowadays, considering the advantages of digital documents, it is necessary to convert written and printed documents to digital form [1,2]. Presently, the best method to convert written documents to digital form is to use OCR (Optical Character Recognition) software. In most OCR methods, the first phase is segmentation of input pages. Different definitions have been proposed for segmentation. However, the concept shared by most of the definitions is to find similar regions in the document and labeling them as text, image, table and so on [2-4].

Much work has been done in recent decades for segmentation and many algorithms have been developed [2,5-8]. Segmentation of pages in special cases, e.g. for English printed pages, is considered as a solved problem [6,7]. Most proposed methods assume that the input is a binary image [2,6,9,10]. Indeed some methods have also been proposed for color images [11], but, due to the complexity of color pictures and the fact that most documents like office letters and books are not colored; methods based on binary or grayscale images are widely used. The method proposed in this paper assumes inputs as grayscale image.

Segmentation methods are generally classified into three groups [14]: Top-down methods [13], bottom-up methods [2,9] and hybrid methods [12]. The method adopted by this paper is a bottom-up one.

From another point of view, segmentation methods can be classified into two groups [4,13]: First, methods designed for a special application and pages with certain characteristics such as table of contents (TOC) [15], TOC and indexes [16] and mailing addresses on postal parcels [17]. These methods use the specific characteristics of those documents like that they have a tabular form, or page numbers are usually short [16]. The second group is general algorithms that do not use certain characteristics of documents. The method proposed in this paper is included under the second group.

Most methods were proposed in the preliminary years and some new methods suppose that the document structures have a rectangular form [1,2,10,18,19] while new methods try to segment complex and irregular pages as well [4,9,12,20]. The proposed method has no hypothesis on the rectangularity of regions.

In terms of sensitivity to skew, most methods are skew-sensitive like [1] and some can tolerance small skew like [4,13,20]. One of the advantages of the proposed method is its skew resistance.

Most of the above mentioned methods are for printed texts. This is natural considering that the main application of OCR methods is for printed texts. Some work has been done indeed for segmenting handwritten texts [7,21-23].

This paper provides a new method for segmenting Persian and as well as Arabic printed texts. It is based

on the Ink Spread Effect. This method is resistant to skew and has no hypothesis about the input page. This method has been designed mainly for Persian and Arabic texts. However, according to a limited number of tests that were performed on English texts, it seems that the method can be used for other languages as well.

The structure of this paper is as follows: in section 2, the special characteristics of the Persian and Arabic scripts are expressed. In section 3, the proposed algorithm will be provided. The results of implementing the algorithm are provided in section 4 and, finally section 5 provides the general conclusion.

2. Special Characteristics of Persian and Arabic Scripts

This section discusses two important characteristics of the Persian and Arabic scripts that distinguish them from the English script. These two characteristics, as mentioned in [10,18,21], highly affect the accuracy of OCR systems.

2.1. Dots

In English alphabet only two letters have dot, "i" and "j", while, even if their dot is removed, it will still be possible to recognize them. But in Persian, 18 out of the 32 and in Arabic 15 out of the 28 letters have dots. In most cases, the letters can only be recognized from each other only according to the position or the number of dots. This makes dots highly important and they should not be deleted under any circumstances. On the other hand, the sizes of the dots are very small and are very close to the sizes of the noises that may appear in images due to scanning or the low quality of paper [24]. For example, in one of the pictures of the test collection that had been scanned with 300 dpi, the dot size of the small fonts (almost size 8) was approximately 2 to 8 pixels while the size of noise was about 5 pixels. This similarity is so strong that even a human user should pay attention to the context in order to be able to recognize the noise in some cases, and this is almost impossible to be done during the segmentation phase. Therefore, no noise removing algorithm was used in this method because the tested methods each resulted in the removal of a considerable percentage of dots.

2.2. Cursive Writing

In Persian and Arabic, letters are connected during writing (words are cursively written). As a result, the connected components in a certain section of the text, such as a line, are not similar and the component sizes are very different from each other. In English, by using the relative uniformity feature of the component size, most non-text components can be removed [9], while this is not possible for Persian texts.

3. The Proposed Algorithm

This section provides the proposed algorithm. In general, this algorithm consists of three phases:
1. Pre-processing; 2. Simulating the Ink Spread Effect; 3. Specifying the segments. In section 3.1, the pre-processing and preparation phase of the image for segmentation has been discussed. Pre-processing action can strongly affect the results [19]. The main idea of the proposed method is the ink spread effect on a paper. This idea and its simulation are explained in section 3.2. In section 3.3, the final phase in which the pieces are specified is explained. In section 3.4, it is shown how the results for very complex or low-quality pages can be improved by running the algorithm for several times.

3.1. Pre-Processing [26]

In this work, it has been assumed that the input image is a grayscale image obtained by scanning a page. The first problem is to convert the grayscale image into a binary image. One of the most common ways is to use a threshold level such as α. Thus, pixels which are greater than α, are considered as white and the rest as black. In some methods, such as [1], default thresholds are used. This is good for uniform pages. In general, however, considering the varying amounts of brightness in different texts, it is not a good method. In our method, the Otsu method, which is provided in [25], has been used. With this method, the threshold is dynamically calculated. As it is mentioned in section 3.4, by applying this method on regions with gray background, which are, for example, created due to scanning color images, the background is entirely removed.

After this phase, connected components are calculated with the method provided in [25]. Then, very large components (which are bigger than a predefined threshold β), are removed as image regions. Considering that the letters are connected, this threshold must be very big because connected components in the titles can be very large, and consequently, if we choose small values for β, then the titles may be removed. In the next phase, the lines and frames are removed. To do this, for each component, minimum and maximum of x and y of its pixels are

calculated, where x and y indicate the coordination of pixel. Then, the smallest rectangle that encompasses the component (with corners (minX, minY) and (maxX, maxY)) is founded. If the proportion of height to width or width to height of the component is more than γ, the component is removed as a line. If the proportion of the black pixels to the area of the rectangle is less than δ, the component is removed as a frame. In addition, components in which the proportion of the black pixels to the area of the rectangle exceeds a certain ζ and also their size exceeds a threshold η (i.e. they were not dots), will be removed as photo regions. Similar techniques are used in other methods such as [20].

3.2. Simulation of the Ink Spread Effect

The main idea of our method was inspired by the ink spread effect on paper. While writing on paper with ink, ink is smeared around words to a certain extent. The progress of ink around words depends on the word size or, in other words, the word volume and the amount of ink used for writing the word. Inspired by this phenomenon, in the Ink Spread Effect simulating method, each connected component extends all around in proportion to its size. If this is done carefully, all the components in a paragraph or a column connect to each other to form a whole component. Also, the separate paragraphs or columns form separate components.

For simulating the ink spread effect, first the spreading radius for each component is calculated. In general, the size of a component has a relation to the square of its font size. Moreover, the distance between lines and words is proportionate to the font size. We found that, if the spreading radius is selected proportionate to the font size, the different words of a paragraph will connect to each other. Font size of a component can be approximately considered to be the square root of the component size. In our method, a linear relation proportionate to the square root of the component size is used for the spreading radius and the coefficients relating to this relation are calculated according to the sample images. While examining the sample images, it was seen that most titles were one line and the distance between titles and the relevant text was equal to the distance between two lines with the title font. Therefore, for big components that are likely titles, the spreading radius is modified in order to prevent connection of the title to the following text. To simulate the ink spread effect, for each border pixels of the component (pixels with a white neighbor), a filled circle with the spreading radius centered on that pixel is drawn.

3.3. Specifying the Segments

In this phase, first the connected components of image resulting from previous part are calculated. Whereas we assume that each text segment has more than one word, the text segment size is bigger than a threshold λ. On the same basis, components smaller than λ are not text segments, so they are deleted. As a result, many components that have been created from noise or parts of photos are deleted and the total number of components is reduced. Each of the remaining components is considered as a text region. The resulting image is used as a mask and, using it, the regions relating to each segment are specified from the output image of phase 3.1.

3.4. Repetition for Improving the Results

On complex pages such as general magazines, the different text regions vary in terms of brightness, and, as a result, the binary-making algorithm does not work well. Therefore, one can first run the entire algorithm so that the regions are specified to a certain extent. Then, in the second iteration, the binary-making algorithm is run separately for each component. This process, removes background of text regions with non-white (gray) backgrounds.

4. Experimental Results

This method was tested on 16 images and the results were carefully examined. This collection consists of pages of a sports journal with a low-quality paper similar to the conditions of newspaper pages [24], a specialized computer book, a general computer magazine with an acceptable print quality and some advertisement brochures. For the specialized book, considering the normal structure, most regions were recognized correctly (Figure1.). Few words that were deleted were cases in which the space between the words of a line had been more than the normal limit in order to make the page more beautiful in typesetting. As in the case of the general computer magazine, because of the clear structures and the good quality of paper, good results were obtained (Figure2). The deleted words were mainly in the title. As to the brochure, the pages were cramped and the fonts were small. In addition, there were numerous pictures and the background was mainly a color background. However, the results are good (Figure3). As to the sports magazine, despite the lower quality, the binary-making operation was carried out very well (especially by repeating the algorithm) and the connection of the

components was due to the close distance of the different columns (Figure4).

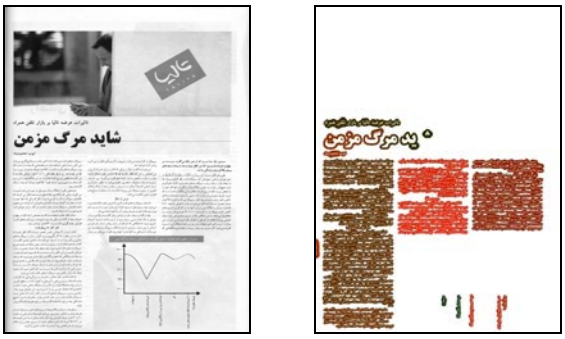
In terms of error, there was only one case of horizontal merge of the text regions and a few cases of vertical mergers. As horizontal merger had an undesirable effect on the quality of the OCR system [5], the quality of this method is high from this point of view. Indeed, in cases in which the texts are very close to the images, the images are considered as a part of text segment. Most non-text segments were relatively small regions and were parts of photos. All texts within tables, frames and with a non-white (gray) background are well segmented. No change was seen in presence of skew. Time needed for processing each image on an ordinary laptop computer is less than one minute.

5. Conclusion

This paper provides a method for segmenting, based on the ink spread effect, which is a new idea. This method is highly resistant to skewing. Areas with a gray background or within frames or tables are specified well and are processed with an acceptable speed. In addition, no training is required and it has been designed for a general state rather than a specific use. Therefore, it can be used for different applications. Another important feature is its full compatibility with the Persian as well as Arabic scripts.

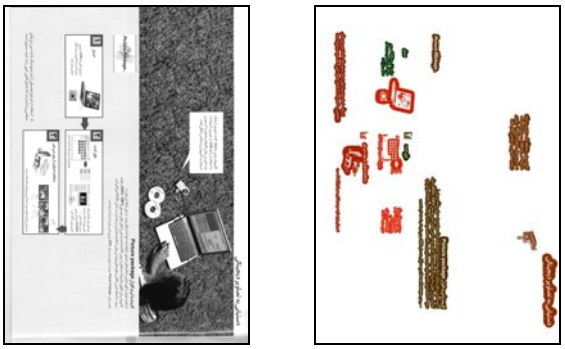
(a) (b)

Figure 1. Sample page from a Book with severe skew (a) original page, (b) segmentation result



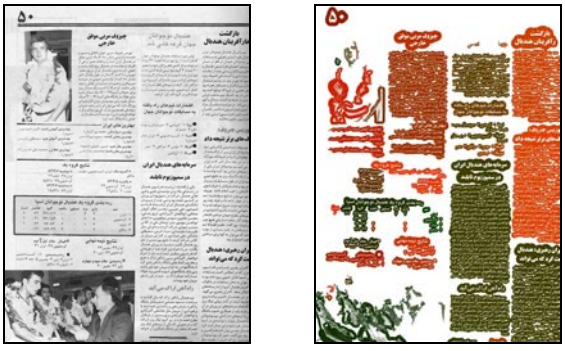
(a) (b)

Figure 2. Sample page from a computer magazine (a) original page, (b) segmentation result



(a) (b)

Figure 3. Sample page of an advertisement brochure (a) original page, (b) segmentation result



(a) (b)

Figure 4. Sample page of a sport magazine (a) original page, (b) segmentation result

6. References

[1] L. Cinque, L. Forino, S. Levialdi, L. Lombardi, S. Tanimoto, "Understanding the Page Logical Structure", *Proceedings of International Conference on Image Analysis and Processing (ICIAP1999)*, Italy, September 1999, pp. 1003-1008.

[2] Q. Wang, Z. Chi, R. Zhao, "Hierarchical Content Classification and Script Determination for Automatic Document Image Processing", *Proceedings of 16th International Conference on Pattern Recognition (ICPR2002)*, Canada, 2002, vol. 3, pp. 77-80.

[3] D.P. Mukherjee, S.T. Acton, "Document Page Segmentation Using Multiscale Clustering", *Proceedings of International Conference on Image Processing (ICIP 1999)*, Japan, October 1999, vol. 1, pp. 234-238.

[4] A. Antonacopoulos, "Page Segmentation Using the Description of The Background", *Computer Vision and Image Understanding*, *Elsevier*, June 1998, vol. 70, Issue 3, pp. 350-369.

[5] S. Agne, A. Dengel, B. Klein, "Evaluating SEE - A Benchmarking System for Document Page Segmentation", *Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Scotland, August 2003, vol. 1, pp. 634-638.

[6] A. Antonacopoulos, B. Gatos, D. Karatzas, "ICDAR 2003 Page Segmentation Competition", *Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Scotland, August 2003, vol.1, pp. 688-692.

[7] Y. Zheng, H. Li, D. Doermann, "Text Identification in Noisy Document Images Using Markov Random Field", *Proceedings of 7th International Conference Document Analysis and Recognition (ICDAR2003)*, Scotland, August 2003, vol. 1, pp. 599-603.

[8] S. Mao, A. Rosenfeld, T. Kanungo, "Document Structure Analysis Algorithms: A Literature Survey", *Proceedings of Document Recognition and Retrieval X, SPIE*, January 2003, vol. 5010, pp. 197-207.

[9] J. Wang, Y. Li, X. Huang, Z. He, "Page Segmentation and Classification Based on Pattern-list Analysis", *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP2004)*, Honk Kong, October 2004, pp.735-738.

[10] M. J. Tabrizi, M. H. Shirali-Shahreza, "A New Page Segmentation Method For Persian Documents", *Proceedings of 15th IASTED International Conference on Applied Informatics*, Austria, 1997 .

[11] H.M. Suen, J.F. Wang, "Text string extraction from images of colour-printed documents", *IEE Proceeding on Vision, Image and Signal Processing, IEE,* August 1996, vol. 143, Issue 4, pp. 210-216.

[12] B. Kruatrachue, P. Suthaphan, "A Fast and Efficient Method for Document Segmentation for OCR", *Proceedins of IEEE Region 10th International Conference on Electrical and Electronic Technology*, Singapore, August 2001, vol. 1, pp. 381-383.

[13] K. Etemad, R. Chellappa, D. Doermann, "Document Page Segmentation By Integrating Distributes Soft Decisions", *Proceedings of 1994 IEEE International Conference on Neural Network*s, USA, June-July 1994, vol.6, pp. 4022-4027.

[14] S. Mao, T. Kanungo, "A Methodology for Empirical Performance Evaluation of Page Segmentation Algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE*, Mar 2001, pp. 242-256.

[15] S. Mandal, S.P. Chowdhury, A.K. Das, B. Chanda, "Automated Detection and Segmentation of Table of Contents Page from Document Images", *Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Scotland, August 2003, vol. 1, pp. 398-402.

[16] S. Mandal, S.P. Chowdhury, A.K. Das, B. Chanda, "Automated Detection and Segmentation of Table of Contents Page and Index Pages from Document Images", *Proceedings of 12th International Conference on Image Analysis and Processing (ICIAP2003)*, Scotland, September 2003, pp. 213-218.

[17] S. Randriamasy, "A set-based benchmarking method for address bloc location on arbitrarily complex grey level images," *Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR1995)*, Canada, August 1995, vol. 2, pp. 619-622.

[18] K. Hadjar, R. Ingold, "Arabic Newspaper Page Segmentation", *Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Scotland, August 2003, pp. 895-899.

[19] K. Hadjar, O. Hitz, R. Ingold, "Newspaper Page Decomposition Using a Split and Merge Approach", *Proceedings of 6th International Conference Document Analysis and Recognition (ICDAR2001)*, USA, September 2001, pp. 1186-1189.

[20] P.E. Mitchell, H. Yan, "Document Page Segmentation and Layout Analysis using Soft Ordering", *Proceedings of 15th International Conference on Pattern Recognition (ICPR2000)*, Spain, September 2000, vol. 1, pp. 458-461.

[21] Shirali-Shahreza, M.H., *Off-line Recognition of Farsi Handwritten Words & Numerals by Neural Networks*, Ph.D. Dissertation, Electrical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, 1996.

[22] S. Nicolas, T. Paquet, L. Heutte, "Text Line Segmentation in Handwritten Document Using a Production System", *Proceedings of 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR2004)*, Japan, October 2004, pp. 245-250.

[23] E. Lecolinet, L. Likforman-Sulem, "Handwriting Analysis : Segmentation and Recognition", *IEE European Workshop on Handwriting Analysis and Recognition: A European Perspective*, Belgium, July 1994, pp. 17/1-17/8.

[24] B. Gatos, S.L. Mantzaris, A. Antonacopoulos, "First International Newspaper Segmentation Contest", *Proceedings of 6th International Conference on Document Analysis and Recognition (ICDAR2001)*, USA, September 2001, pp. 1190-1194.

[25] Shapiro, L.G., and G.C. Stockman, *Computer Vision*, Prentice Hall, 2001.

[26] S Shirali-Shahreza, M.T. Manzuri-Shalmani, M.H. Shirali-Shahreza, "Preparing Persian/Arabic Scanned Images for OCR," *Proceedings of 2nd IEEE International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA'06),* Syria, April 2006, vol. 1, pp. 1332- 1336.