International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:5, 2014

Persian/Arabic Document Segmentation Based On Pyramidal Image Structure

Seyyed Yasser Hashemi, Khalil Monfaredi

***Abstract*—**Automatic transformation of paper documents into electronic documents requires document segmentation at the first stage. However, some parameters restrictions such as variations in character font sizes, different text line spacing, and also not uniform document layout structures altogether have made it difficult to design a general-purpose document layout analysis algorithm for many years. Thus in most previously reported methods it is inevitable to include these parameters. This problem becomes excessively acute and severe, especially in Persian/Arabic documents. Since the Persian/Arabic scripts differ considerably from the English scripts, most of the proposed methods for the English scripts do not render good results for the Persian scripts. In this paper, we present a novel parameter-free method for segmenting the Persian/Arabic document images which also works well for English scripts. This method segments the document image into maximal homogeneous regions and identifies them as texts and non-texts based on a pyramidal image structure. In other words the proposed method is capable of document segmentation without considering the character font sizes, text line spacing, and document layout structures. This algorithm is examined for 150 Arabic/Persian and English documents and document segmentation process are done successfully for 96 percent of documents.

***Keywords*—**Persian/Arabic document, document segmentation, Pyramidal Image Structure, skew detection and correction.

I. Introduction

IN order to segment a document which is an important step in Optical Character Recognition (OCR) systems, the document image is divided into homogeneous zones, each consisting of only one physical layout structure, such as text, graphics, and pictures. Therefore, the performance of OCR systems depends heavily on the implemented document segmentation algorithm. Several document segmentation algorithms have been proposed during the last three decades [1]-[10].

The various approaches toward document segmentation are typically categorized as “bottom-up”, “top-down”, and “textural analysis” methods. The “bottom-up” methods [1]-[3] start from pixels or the connected components, determine the words, merge the words into text lines, and finally merge the text lines into paragraphs. The main disadvantage of these approaches is that the identification, analysis, and grouping of connected components are, in general, time-consuming processes, especially when there are many components in the image. The “top-down” approaches [4]-[7] look for global information e.g. black and white stripes on the page and use them to split the page into columns, the columns into blocks, the blocks into text lines, and finally the text lines into words. Low time complexity of these methods in comparison to the prior methods, i.e. “bottom-up” approaches and their natural top-down view from coarse to fine resolution as is preferable for human beings’ eyes are of the most important advantages of this method. On the other hand, in “top-down” techniques it is unfortunately difficult to segment the complex document layouts which include some nonrectangular images and various character font sizes. Some other recently proposed document segmentation methods [8]-[10], consider the homogeneous regions of the document image such as text, image or graphic as a textured region. Thus, document segmentation is implemented based on the textured regions found in gray scale images. Very high time complexity is the main problem associated with these texture-based approaches, since many masks are used for extracting local features and also different tuning filters are used to capture a desired local spatial frequency and the orientation characteristics of a textured region. Since the Persian documents have some special characters which does not exist in the English documents, so the aforementioned methods cannot be directly used for Persian document segmentation.

The special characters of Persian documents are as follows:

1) The Persian scripts are cursive and each connected component includes more than one character. On the other side their arrangement and size may also vary tremendously.
2) There are 32 basic characters in the Persian alphabet. These characters may change their shapes according to their positions (beginning, middle, end or isolated) in the word. Since each character can take four different shapes, thus we have 114 different shapes considering all of Persian alphabets.
3) Special stress marks called dots are the other characteristics of the Persian scripts. Most of the Persian characters have one, two or three dots. These dots may be situated at the top, inside or bottom of the characters.

From the script identification point of view, it is concluded from the above mentioned expressions about the special characters of Persian documents that these scripts’ word sizes are non-uniform. The word size may vary according to the number of cursive characters and dots in the word.

In this paper, we propose a novel method for Persian document segmentation using pyramidal image structure. This paper is organized as follows: In Section II, the proposed algorithm is described in detail. Experimental results of the

Seyyed Yasser Hashemi and Khalil Monfaredi are with the Department of Computer Engineering, Miyandoab Branch, Islamic Azad University, Miyandoab, Iran (e-mail: hashemi.uni@gmail.com, khalilmonfaredi@gmail.com).

proposed algorithm are presented in Section III. Finally, Section IV discusses the paper results.

II. Proposed Method

Many document segmentation algorithms are presented for English Documents which most of them do not provide good results for Persian/Arabic documents due to their differences mentioned above. To make these methods suitable for Persian scripts, some of their parameters must be specialized. We have proposed a parameter free segmentation method for Persian/Arabic documents which gets rid of these restrictions. The proposed method is interestingly capable to segment the Persian documents composed of different font sizes, different lines spaces and also different structure layouts. In the proposed method, the low resolution version of the document is firstly processed and then the document's high resolution version is analyzed in detail. This manifests the pyramidal nature of the proposed method. The pyramidal tree structure with vertical and horizontal analysis is used in order to segment the document. First, we extract the edge of the document using the horizontal projection profile. If the edges are placed in regular intervals, the area is segmented as a text area. Otherwise, we will find position(s) to split the area into smaller regions, recursively. The proposed algorithm steps are listed as:

1) Document skew detection and correction (Section II, *A*).
2) Pyramidal quad tree structure Construction and multi scale image generation $\left(I_i\ (i=1,\ldots,N)\right)$ from the input document image (I0) (Section II, *B*).
3) Considering the original document image as initial member in collection of segmented regions (R)
4) Vertical analysis and then splitting R regions of (Ii) image of document in horizontal direct (Section II, *C*).
5) Horizontal analysis and then splitting R regions of Ii image of document in vertical direct (section II, *D*).
6) Repeating Steps 4-5 until the region becomes a single homogeneous region.

A. Document Skew Detection and Correction (SDC)

In this step, the skew angle of the document (θ) must be estimated. The proposed method uses a document SDC based on Centre of Gravity (COG). To determine the skew angle, first step is the Baseline Identification (BI). The angle between the baseline and direct horizontal lines determine the skew angle. Therefore, the most important step in this process is to identify the baseline. Baseline of the document is a line that passes through the COG along the horizontal axes. In this algorithm, we detect skew angle by finding Actual Region of Document (ARD) using connected component analysis, identifying its COG and identify the baseline of document. The angle between the baseline and horizontal lines specifies the skew angle. The algorithm steps are as:

1) Document segmentation (CC identification)
2) Identification of the ARD. For this purpose, four CCs that have the more distance from the C0, C1, C2 and C3 corners (shown in Fig. 3) will be selected.
3) Finding the COG. COG is calculated using (1):

$$\begin{aligned} COG_x &= \frac{1}{6A}\sum (x_i+x_{i+1})(x_i y_{i+1}-x_{i+1}y_i) \\ COG_Y &= \frac{1}{6A}\sum_{i=0}^{N-1} (y_i+y_{i+1})(x_i y_{i+1}-x_{i+1}y_i) \\ A &= \frac{1}{2}\sum (x_i y_{i+1}-x_{i+1}y_i) \end{aligned} \tag{1}$$

'A' is the area of polygon.

4) Baseline identification. A line (baseline) from COG to the center of the line that connects the two upper and lower left corner of the ARD (midpoint).
5) Calculation of the amount of document skew angle which is the angle between baseline and the horizontal line that passes through the midpoint.
6) Rotation of the document (see Fig. 5 which is the rotated version of Fig. 1)

B. Pyramidal Image Structure

The pyramidal image structure is a simple and robust technique to provide several resolutions of an image [11]. An image pyramid is a collection of decreased resolution images which are arranged in the shape of pyramid in a way that the base of the pyramid contains a high-resolution while the apex contains a low resolution approximation of the image (Fig. 2).

Fig. 3 represents a simple system for constructing image pyramids. The Ii+1 output is used to approximate upper level low resolution image from the original image to be included in the pyramid. To do this, the OR logic is first applied to the adjacent odd and even columns of the image Ii, and then it is applied to the adjacent odd and even rows of the resulted image Ii3.Therefore, the number of pixels in Ii+1 is on quarter of the number of pixels in Ii. This process is repeated N times (number of levels of pyramidal image structure) that is given by (2):

$$\begin{aligned} N &= \left\lfloor \log\frac{l}{100} \right\rfloor \\ l &= \min\left\{I_0.Width, I_0.Hight\right\} \end{aligned} \tag{2}$$

Fig. 4 depicts the final multi scaled images constructed using the proposed system of Fig. 3.

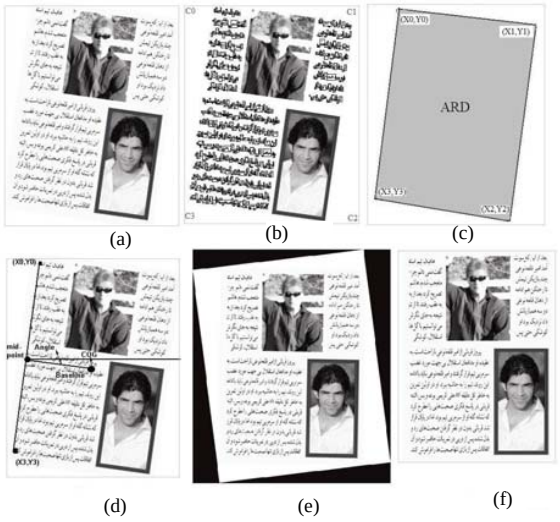


Fig. 1 (a) Skewed document image, (b) Document segmented to connected component, (c) The ARD in skewed document, (d) calculate the amount of document skew angle, (e) deskewed document, (f) final result

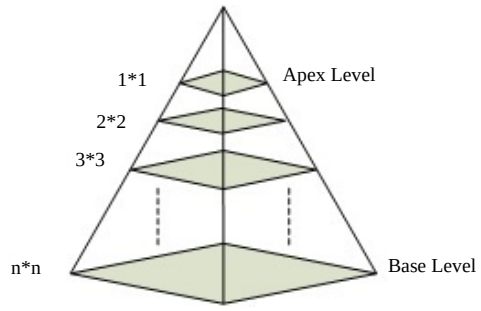


Fig. 2 Pyramidal Image Structure

C. Vertical Analysis

If the text lines of the document are aligned vertically, vertical analysis will segment the document horizontally starting with the lowest resolution image. If the document image of level 'L+1' has no horizontal grooves, the vertical analysis will be performed on the document image of level 'L'. The vertical projection profile as the first step of the vertical analysis is given by (3):

$$P_V(n)=\frac{1}{H}\sum_{y=1}^{H} I_L(n,y) \qquad 1\le n\le W \tag{3}$$

where IL(x,y) is the intensity value of the W×H image at the Lth level. In the second step PV(n) is transformed to the binary signals as described in (4):

$$t_V(n)=\begin{cases}1.0 & P_V(n)>0.05\\ 0.0 & otherwise\end{cases} \tag{4}$$

Analyzing the binary signals, tv(n), is a good decision factor to segment the document horizontally. If all tv(n) signals have the constant value of 1, it reveals that the region has no grooves and thus the segmentation cannot be performed. If tv(n) have at least one signal with 0 value, then document segmentation in the horizontal direction can be performed.

In this process, the descending edge, zero value, and ascending edge of the signal determines the segmentation point (Fig. 5). Fig. 6 shows the result of vertical analysis on one region of the document which has three segmentation points.

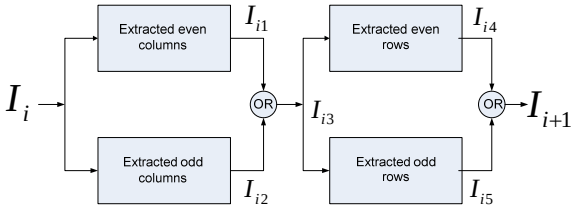


Fig. 3 A simple system to construct each stage of the image pyramid

Fig. 4 Multi-scale images: (a) level 0 (480x670), (b) level 1 (240x335), (c) level 2 (120x167), and (d) level 3 (60x84)

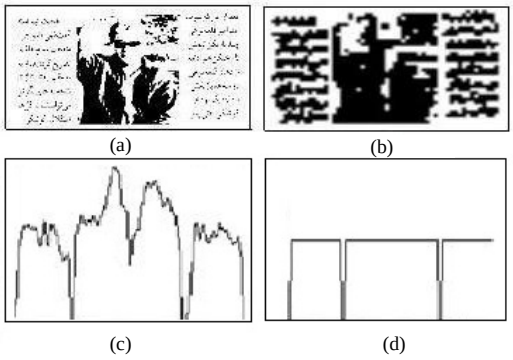


Fig. 5 Vertical analysis (a) original document, (b) image in level 3, (c) the vertical projection profile of image in level 3, (d) the binary signal of the vertical projection profile.

Fig. 6 The document segmented with vertical analysis

D.Horizontal Analysis

The regions obtained from vertical analysis, are further segmented vertically using horizontal analysis. In order to speed up the process, the horizontal analysis is performed recursively. At the first step, the horizontal projection profile is calculated by (5) for each region:

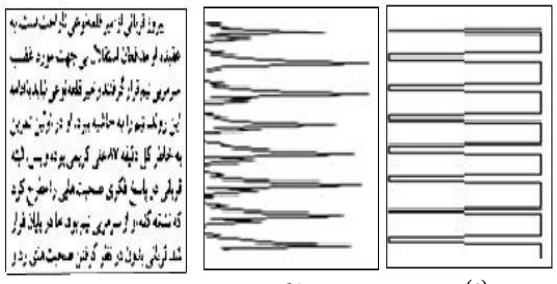
$$P_H(n)=\frac{1}{W}\sum_{x=1}^{W} I_L(x,n) \qquad 1\le n\le H \tag{5}$$

where IL(x,y) is the intensity value in the W×H image of the Lth level and PH(n) are normalized between zero to one. Fig. 7 shows PH(n) for one region of the document.

At the second step, the normalized projection profile is transformed to binary signals as described in (6):

$$tP_H(x)=\begin{cases}1.0 & P_H(x)>0.05\\ 0.0 & otherwise\end{cases} \tag{6}$$

Fig. 7 shows the binary signal.



(a) (b) (c)

Fig. 7 (a) One region of original document, (b) the normalized projection profile, (c) the binary signal of the normalized projection profile

At the third step, the difference of signals calculated using (8):

$$dP_H(n)=diff(tP_H(n)) \tag{8}$$

At the fourth step, ascending and descending edges are calculated using (9) and (10)

$$UHE(n)=\begin{cases}1 & dP_H(n)>0\\ 0 & otherwise\end{cases} \tag{9}$$

$$DHE(n)=\begin{cases}1 & dP_H(n)<0\\ 0 & otherwise\end{cases} \tag{10}$$

where UHE(n) and DHE(n) determine the ascending and descending edges of the signal, respectively.

At the fifth step, the distance between black and white areas of the signal is calculated using (11) and (12):

$$PW(n)=DHE(n)-UHE(n) \tag{11}$$

$$PB(n)=UHE(n+1)-DHE(n) \tag{12}$$

PW(n) and PB(n) are the distance between white and black tPH(n) signals, respectively. Sum of PW(n) and PB(n)

D(p) which is the decision value for document segmentation is derived from Pi(n) by (13), (14), (15) and (16) as:

$$m=\frac{\sum_{n=1}^{N} P_I(n)}{N} \tag{13}$$

$$V=\frac{\sum_{n=1}^{N} (P_I(n)-m)^2}{N} \tag{14}$$

$$P=2-\frac{2}{1+e^{-V}} \tag{15}$$

$$D(P)=\begin{cases}1 & p>TH\\ 0 & otherwise\end{cases} \tag{16}$$

Using the decision value of D(p), we can estimate whether a region is homogeneous or not.

Where ‘N’ is the number of grooves, we set the threshold value TH=0.5. It is achieved for ‘V’ approximately equal to 1.099 in (14). This value for ‘V’ is independent of the character font sizes, text lines spacing, and the document layout structures, and is equally applied to each region to decide whether it is a homogeneous region or not.

There are three types of horizontal analysis using D(p), tPH(n) and Pi(n):

1) *For constant signal tPH(n) equal to 1:* In this type, if tPH(n) relates to upper level (L=1, 2,..., N) of the region of the document image, the horizontal analysis is further repeated for lower levels, but if tPH(n) relates to the original document image, the region is a non-text region (graphics, pictures,…).
2) *For symmetric signal tPH(n):* If the decision value D(p) is 1, the variance of the Pi(n) values are low and the region is considered a text region and hence it requires no further splitting.
3) *For non-symmetric signal tPH(n):* If the decision value D(p) is 0, the variance of Pi(n) values are high and the region is not a homogeneous region; and hence further splitting is inevitable.

Segmentation process is repeated until all regions in all levels have a constant signal tPH(n) equal to 1 or symmetric tPH(n) signal.

E. Determination of Splitting Position

Determination of splitting position is necessary when a region is not homogeneous. This occurs in two cases considering the horizontal direction, in one case at least one white area is larger than the other white areas and in the other case at least one black area is larger than the other black areas. In these cases, we use the following method to find a suitable position for splitting the region.

Let 'W' denotes the set of the white areas of a region (wi) and 'B' denotes the set of the black areas of the region (bi), sorting W (or B) in a ascending order in terms of wi (or bi) magnitude if wi>wmed and wi>7wmax, split wi and if bi>bmed and bi>7bmax, split wi-1. Where, wmed is the median element of W, bmed is the median element of B, wmax is the last element of W, and bmax is the last element of B.

The processes of Sections III, *B* and *C* are repeated for each region until no further splitting is required. Fig. 8 shows the result of proposed document segmentation method.

Fig. 8 Document segmented with proposed method

TABLE I
PERFORMANCE EVALUATION

	Correct	Fail	Precision (%)
Region location	870	**31**	**96.5**
Text	623	**18**	**97.4**
Non-Text	247	**13**	**95**

III. EXPERIMENTAL RESULTS

The Proposed method has been implemented on duo core 2.0 GHZ in 2014. We have considered different skewed and non-skewed documents from different sources like journals, textbooks, newspapers and also handwritten documents. For experimentation purpose 150 documents are considered which 50 of them are handwriting documents. The obtained results are reported in Table I.

IV. CONCLUSION

We have introduced a document segmentation method that segments the Persian/Arabic document into homogeneous regions. The proposed efficient, simple and fast method works based on pyramidal image structure. The skew of the document is corrected first and then a pyramidal image structure is constructed for multi-scale analysis and finally the document image is segmented by vertical and horizontal analysis. The proposed method was examined on different skewed and non-skewed documents achieved from different sources. Experiments show more accurate and high speed results in comparison to the previously reported methods. This method focuses on the Persian/Arabic document segmentation, which also exhibits good results for other scripts such as English scripts. This work can be also extended for special works such as license plate recognition, postal service, and noisy documents. This method was implemented on 150 different documents (90 Persian/Arabic and 40 English and 20 hybrids of English and Persian/Arabic) and the rate of accuracy is 96%.

REFERENCES

[1] F. Legourgiois, Z. Bublinski, and H. Emptoz, "A Fast and Efficient Method for Extracting Text Paragraphs and Graphics from Unconstrained Documents", *Proc. 11th Int'l Conf. Pattern Recognition* , 1992, pp. 272-276.

[2] D. Drivas and A. Amin, "Document segmentation and Classification Utilizing Bottom-Up Approach", *Proc. Third Int'l Conf. Document Analysis and Recognition*, 1995, pp. 610-614.

[3] A. Simon, J. Pret, and A. Johnson, "A Fast Algorithm for Bottom-Up Document Layout Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, vol. 19, pp. 273-276.

[4] J. Ha, R. Haralick, and I. Phillips, "Recursive X-Y Cut Using Bounding Boxes of Connected Components", *Proc. Third Int'l Conf. Document Analysis and Recognition*, 1995, pp. 952-955.

[5] J. Ha, R. Haralick, and I. Phillips, "Document Page Decomposition by the Bounding-Box Projection Technique", *Proc. Third Int'l Conf. Document Analysis and Recognition*, 1995, pp. 1119-1122.

[6] Yi Xiaoa, Hong Yana' "Text region extraction in a document image based on the Delaunay tessellation", *Pattern Recognition*, 2003, pp. 799-809.

[7] Jie Xi, Jianming Hu, Lide Wu, "Document segmentation of Chinese newspapers", *Pattern Recognition*, 2002, pp. 2695-2704.

[8] A. Jain and Y. Zhong, "Document segmentation Using Texture Analysis", *Pattern Recognition*, 1996, vol. 29, pp. 743-770.

[9] A. Jain and S. Bhattacharjee, "Text Segmentation Using GaborFilters for Automatic Document Processing"' *Machine Vision and Applications*, 1992, vol. 5, pp. 169-184.

[10] M.Acharyya, M.K.Kundu, "Document Image Segmentation Using Wavelet Scale-Space Features", *IEEE Transaction on circuits and systems for video technology*, DEC 2002, vol. 12, no. 12.

[11] R.C.Gonzalez, R.E.Woods, "Digital Image Processing", S*econd Edition*, 2002, by Prentice-Hall

Seyyed Yasser Hashemi was born in Miyandoab, Azarbayjane Gharbi, Iran, in 1985. He received the B.Sc. and M.Sc. degrees from Islamic Azad University of South Tehran Branch, in Computer Engineering field. He is with Computer Department of Islamic Azad University, Miyandoab Branch since 2008. He is the author or coauthor of more than ten national and international papers and also collaborated in several research projects. His current research interests include voice and image processing, pattern recognition, spam detecting, optical character recognition, cloud computing and parallel genetic algorithms