PERSIAN/ARABIC DOCUMENT SEGMENTATION BASED ON HYBRID APPROACH

Seyyed Yasser hashemi [1], Parisa Sheykhi Hesarlo[2]

[1, 2] Department of Computer Engineering, miyandoab Branch, Islamic Azad University, miyandoab, Iran

ABSTRACT

Document segmentation is an essential requirement for automatic transformation of paper documents into electronic documents. However, some restrictions such as variations in character font sizes, different text line spacing, and also non-uniform document layout structures altogether makes designing a general-purpose document layout analysis algorithm much more sophisticated. Thus in most previously reported methods these parameters are inevitably included. This mentioned issue becomes much more acute and excessively 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/Arabic scripts. In this paper, we present a novel parameter-free method based on hybrid method for segmenting the Persian/Arabic document images which also works well for English scripts. The proposed method is capable of document segmentation without considering the character font sizes, text line spacing, document skew, and document layout structures. This algorithm is examined for 150 Persian/Arabic and English documents and document segmentation process are done successfully for 97.3 percent of them at worst condition.

KEYWORDS

Persian/Arabic document, document segmentation, Pyramidal Image Structure.

1. 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 (F. Legourgiois et al., 1992; D. Drivas et al., 1995; A. Simon et al., 1997) 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 (J. Ha et al., 1995; J. Ha et al., 1995; Yi Xiaoa et al., 2003; Jie Xi et al., 2002, Rafi Cohen et al 2013) 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 (A. Jain and Y. Zhong et al., 1996; A. Jain and S. Bhattacharjee et al., 1992; M. Acharyya and M. K.Kundu et al., 2002), 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. Junxi Sun et al., 2008; propose a texture-based Bayesian document segmentation method. In this method Bayesian method is used to fuse texture likelihood and prior contextual knowledge to achieve document segmentation. The texture likelihood is based on a complex wavelet domain hidden Markov tree (HMT) model and the prior contextual is based on a hybrid tree model. 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:

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.

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.

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/Arabic document segmentation using pyramidal image structure. This paper is organized as follows: In Section 2, the proposed algorithm is described in detail. Experimental results of the proposed algorithm are presented in Section 3. Finally the paper is concluded in section 4.

2. MATERIALS AND METHODS

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

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 aforementioned differences. Thus, in order to make these methods suitable for Persian scripts, some of those method's parameters must be specialized. We have proposed a parameter free segmentation method for Persian/Arabic documents which eliminates these restrictions and provides excellent results. The proposed method is interestingly capable of segmenting 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 original image will be analyzed and a pyramidal tree structure (images at different resolutions) is created. At next phase, the bounding boxes are extracted from images

at lowest possible resolution and candidate regions are selected removing the bounding boxes' overlap areas. Then the regions are classified with horizontal and vertical analysis and are further investigated with textual and statistical analysis of the high resolution sample to recognize different regions of the document such as text, image and drawing/table. The Fig. 1 shows the proposed approach followed in this paper.

2.1. 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:

- Connected Component identification(CC)
- 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. 2(c)) will be selected.
- Finding the COG. COG is calculated using (1).

$$COG_x = \frac{1}{6A}\sum (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$COG_Y = \frac{1}{6A}\sum_{i=0}^{N-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$A = \frac{1}{2}\sum (x_i y_{i+1} - x_{i+1} y_i) \qquad (1)$$

'A' is the area of polygon.

- 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).
- Calculation of the amount of document skew angle which is the angle between baseline and the horizontal line that passes through the midpoint.

Rotation of the document (see Fig. 2)

2.2. Pyramidal image structure

The pyramidal image structure is a simple and robust technique to provide several resolutions of an image. 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. The number of pixels in IL+1 is on quarter of the number of pixels in IL. This process is repeated N times (number of levels of pyramidal image structure) that is given by (2).

$$N = \left\lceil \log \frac{l}{100} \right\rceil, \; l = \min\left\{ I_0.Width, I_0.Hight \right\} \qquad (2)$$

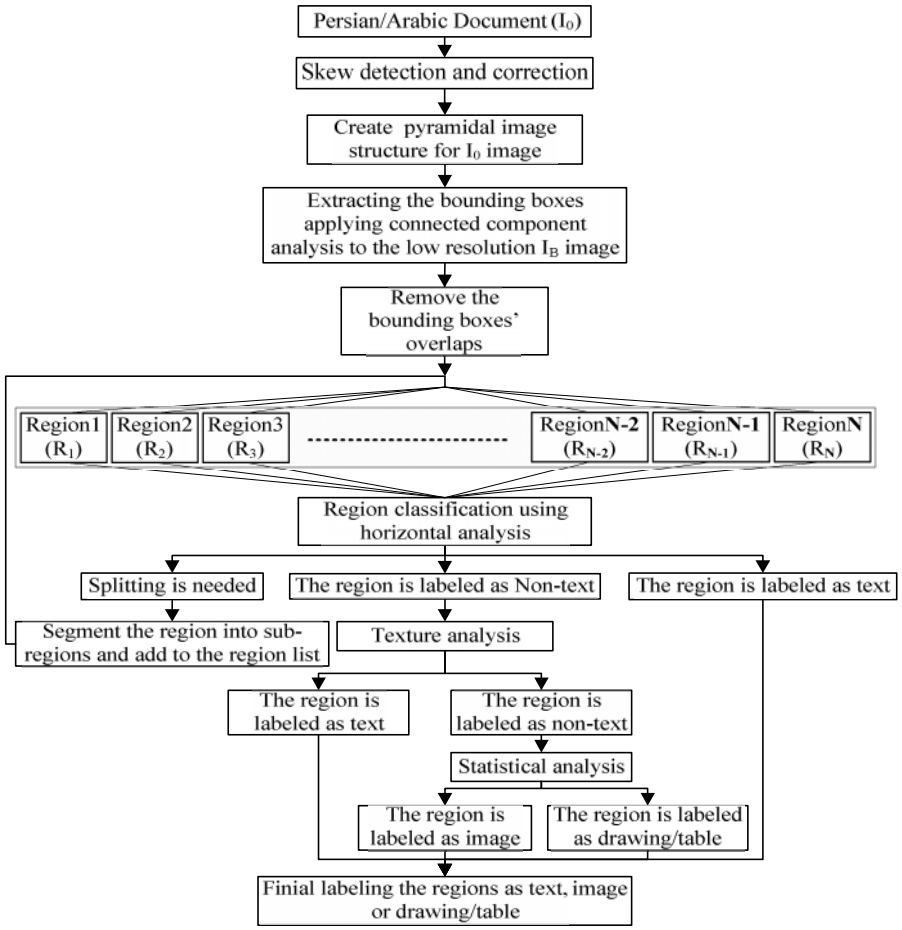

Figure 1. Overall diagram of the proposed method

The intensity of every pixel of the image for 'L+1' level is calculated using pixel intensity of level 'L' with the equation (3) as follows:

$$I_{i,j}^{L+1} = \frac{\left(\sum_{m=0}^{1} \sum_{n=0}^{1} \left(I_{2i+m,2j+n}^{L} \right) \right)}{4} \qquad (3)$$

Where $I_{i,j}^{L+1}$ is the intensity of (i,j) in level 'L+1'. The pyramidal image structure for I_0 image is shown in Fig. 3.

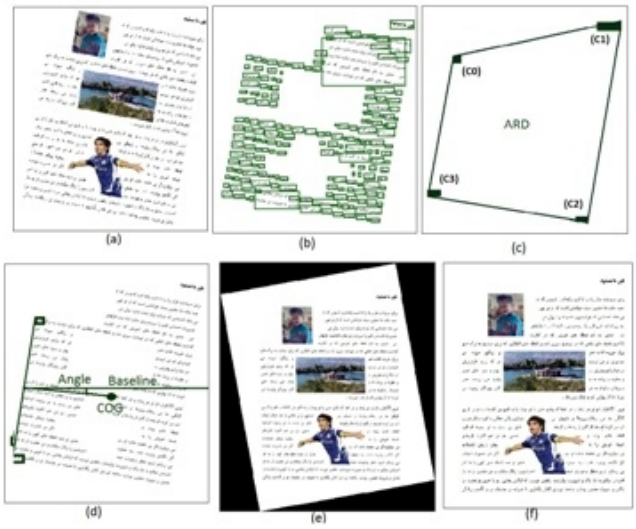

Figure 2. (a) Skewed document, (b) Document segmented to connected component, (c) The ARD in skewed document, (d) calculate the amount of document skew angle, (e) non-skewed document, (f) final result

2.3. Bounding box extraction

This algorithm uses a bottom-up approach to extract connected components (CCs) from the document image. The first step in CCs extraction is to locate rectangular regions called Rects (M. Viswanathan, 2002), A Rects may be thought of as a rectangular region of loosely connected black pixels. More specifically, a Rects has at least one black pixel in every 9-pixel square. A Rects is defined in this way so black regions may be found without scanning every pixel. The second step is to merge adjacent Rects to form CCs .In terms of implementation, Rects are stored in a linked list. This type of structure is ideal because the Rects need to be quickly traversed and random access is not required. The process of locating Rects involves scanning 9-pixel squares of each paragraph in raster order. The top left corner of a Rects is defined by the first 9-pixel square found containing a black pixel. The right edge of the Rects is located by searching right until a white 9-pixel square is found. Similarly, the bottom edge is located by scanning down until a white 9-pixel square is found. CCs are defined as collections of adjacent Rects (Mitchel P. E and Yana H., 2004). CCs are also stored in a linked list, which enables them to be quickly and easily traversed for classification. The process used to construct CCs involves traversing the (initially empty) list of CCs for each Rects in the Rects list. Each Rects is subsequently added to single CCs or used to define new CCs. If a Rects is found to be adjacent to more than one CC, the CCs are merged into a single CCs. Note that since the CCs are defined directly from the Rects, there is no need to refer to the actual image. After extracting CCs, they merge with each other in vertical and horizontal directions to create the bounding boxes.

In multi-resolution analysis, the lowest levels of the pyramid can be used for overall analysis of the image. In this step, the bounding boxes of low resolution image are used to extract image regions. The bounding boxes extraction in lowest resolution of the I0 image is shown in Fig. 4.

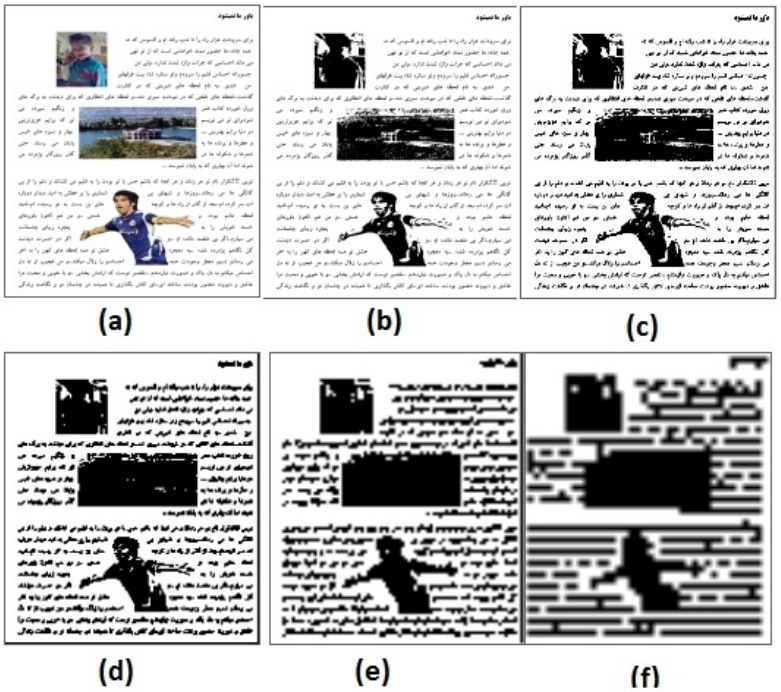

Figure. 3. The Pyramidal image structure for I0 image, (a) original image, (b) binary image, (c) level 0, (d) level 1, (e) level 2, and (f) level 3.

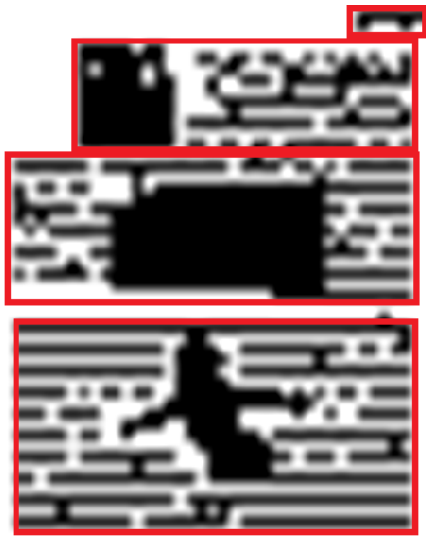

Figure. 4. The bounding boxes extraction in lowest resolution of the I_0 image.

2.4. Remove overlaps connected components

In the low-resolution document segmentation, the image of document I_{0_0} is divided into a set of regions that called $R_1, R_2, ..., R_n$. Each region R_i is defined by coordinates of bounding box. Some of regions, R_i, are overlapped to other. For next stage of high-resolution page

segmentation, we should remove overlapped components from each region R_i. For each region, we remove overlapping components as follow:

First, we obtained connected components of region R_i in low-resolution image P_L. If R_i have only one connected component, then R_i do not have any overlapping components. If in region R_i number of connected components is more than one, we leave the maximum of connected component and remove other connected components at the high-resolution.

2.5. Horizontal analysis

The Abstract section begins with the word, "Abstract" in 13 pt. Times New Roman, bold italics, "Small Caps" font with a 6pt. spacing following. The abstract must not exceed 150 words in length in 10 pt. Times New Roman italics. The text must be fully justified, with a 12 pt. paragraph spacing following the last line.

Persian text region can be easily distinguished from other regions by 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 (4) for each region.

$$P_H(n) = \frac{1}{W}\sum_{x=1}^{W} I_B^L(x,n) \qquad 1 \le n \le H \tag{4}$$

Where $I_B^L(x,n)$ is the intensity value in the W×H image of the L_{th} level and $P_H(n)$ are normalized between zeros to one. Fig. 5 shows $P_H(n)$ for one region of the document.

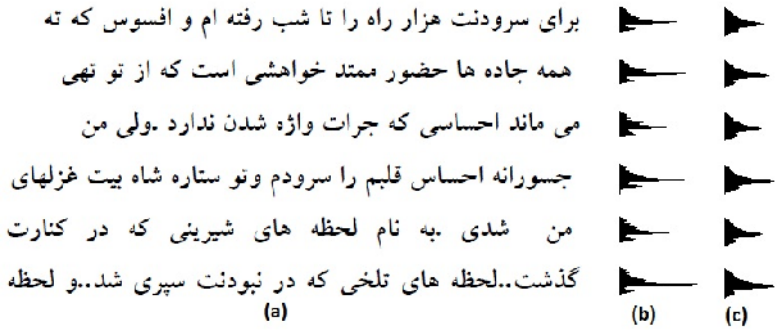

Figure. 5. (a) one region of the original document, (b) the horizontal projection profile, (c) the normalized horizontal projection profile.

At the second step, the normalized projection profile is transformed to binary signals as described in (5).

$$tP_H(x) = \begin{cases} 1.0 & P_H(x) > 0.05 \\ 0.0 & otherwise \end{cases} \tag{5}$$

At the third step, the difference of signals calculated using the equation (6).

$$dP_H(n) = diff(tP_H(n)) \tag{6}$$

At the fourth step, ascending and descending edges are calculated using (7) and (8)

$$UHE(n) = \begin{cases} 1 & dP_H(n)>0 \\ 0 & otherwise \end{cases} \quad (7)$$

$$DHE(n) = \begin{cases} 1 & dP_H(n)<0 \\ 0 & otherwise \end{cases} \quad (8)$$

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 (9) and (10).

$$PW(n) = DHE(n) - UHE(n) \quad (9)$$

$$PB(n) = UHE(n+1) - DHE(n) \quad (10)$$

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 (11), (12), (13) and (14) as:

$$m = \frac{\sum_{n=1}^{N} P_I(n)}{N} \quad (11)$$

$$V = \frac{\sum_{n=1}^{N} (P_I(n)-m)^2}{N} \quad (12)$$

$$P = 2 - \frac{2}{1+e^{-V}} \quad (13)$$

$$D(P) = \begin{cases} 1 & p>TH \\ 0 & otherwise \end{cases} \quad (14)$$

Using the decision value of D(p) we can estimate whether a region is homogeneous or not.
Where 'N' is the number of grooves and we set the threshold value TH=0.5. It is achieved for 'V' approximately equal to 1.099 in (12). 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):

- 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,…) and in next step with textural and statistical analysis will be investigated in detail.

- For symmetric signal tPH(n): If the decision value D(p) is 1, the variance of the $P_i(n)$ values are low and the region is considered a text region and hence it requires no further splitting.

- For non-symmetric signal tPH(n): If the decision value D(p) is 0, the variance of $P_i(n)$ values are high and the region is not a homogeneous region; and hence further splitting is inevitable. Thus this region is sub classified using projection profile information.

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

2.6. Determination of Splitting Position

The Keywords section begins with the word, "Keywords" in 13 pt. Times New Roman, bold italics, "Small Caps" font with a 6pt. spacing following. There may be up to five keywords (or short phrases) separated by commas and six spaces, in 10 pt. Times New Roman italics. An 18 pt. line spacing follows.

When a region is determined to require further splitting because it is not a homogeneous region, there are two cases in the horizontal (Vertical) direction, as shown in Fig. 6. The case in Fig. 6a needs to be split more because one white area is larger than the other white areas and the case in Fig. 6b needs to be split more because one black area is larger than the other black areas. In these two cases, we find a suitable position for splitting, using the method given below, and split it into two regions. The processes described in Sections 2.5 and 2.6 are repeated for each region until no further splitting is required.

Let W denote the set of the white areas of a region, wi.
Let B denote the set of the black areas of a region, bi
Sort the set W(or B) in the increasing order of the magnitude of w_i(or b_i)
w_{med} the median element of W
b_{med} the median element of B
w_{max} the last element of W
b_{max} the last element of B
if $w_i > w_{med}$ and w_i . w_{max}, split w_i
if bi > b_{med} and b_i . b_{max}, split $w_{i-1.}$

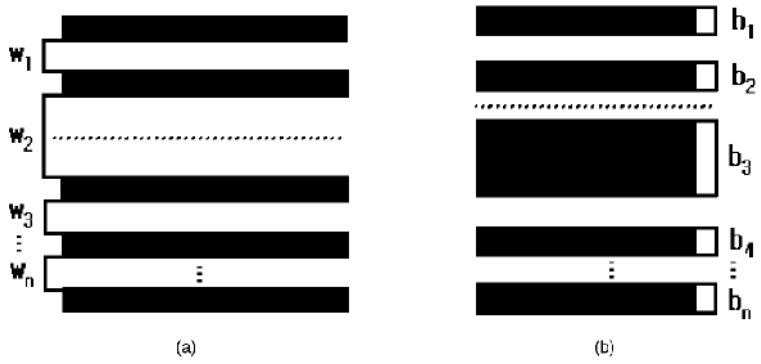

Fig. 6: Two cases requiring further horizontal splitting :(a) A distinct white area exists. (b) A distinct black area exists.

2.6. Non-homogeneous regions analysis

.

In this step, the non-homogeneous regions are analyzed and segmented to homogeneous sub regions. For this purpose, each region is segmented to initial sub region in horizontal direction, based on median of characters high in document (Implications of previous steps). Then by repeating the processes described in Sections 2.5 and 2.6, we segmented a non-homogeneous region into maximal homogeneous sub regions that labeled as text or non-text sub regions. Finally, the sub regions with same label are merged and the final result of this step is obtained.

2.7. Image and drawing/table recognition

The non-text regions which was labeled before, are labeled as image or drawing/table in this section using statistical analysis. In order to do this, the number of blackpixels of each region is applied to the equation (15).

$$\begin{cases} \text{the region is image if } \dfrac{A(R_i)}{W(R_i)\times H(R_i)} \geq 0.4 \\ \text{the region is drawing/table if } \dfrac{A(R_i)}{W(R_i)\times H(R_i)} < 0.4 \end{cases} \tag{15}$$

3. EXPERIMENTAL RESULTS

The Proposed method has been implemented on duo core 2.66 GHZ in 2013. We have considered different 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. The comparative results of the proposed method taking other artworks into account are considered in Fig. 8 and Fig. 9 and the final result for a sample image is shown in Fig. 7.

Fig. 7: The final result for a sample image

4. CONCLUSION

We have introduced a document segmentation method that segments the Persian/Arabic document into homogeneous regions. The proposed efficient and comprehensive method works

based on pyramidal image structure and textual analysis. For the original image, a pyramidal tree structure (images at different resolutions) is created. At next phase, the bounding boxes are extracted from the image at lowest possible resolution and candidate regions are selected removing the bounding boxes' overlap areas. Then the regions are classified with horizontal analysis and are further investigated with textual and statistical analysis of the high resolution sample to recognize different regions of the document such as text, image and drawing/table.

The proposed method was examined on different colorful/gray documents achieved from different sources. Experiments show more accurate 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 97.3% in the worst circumstances.

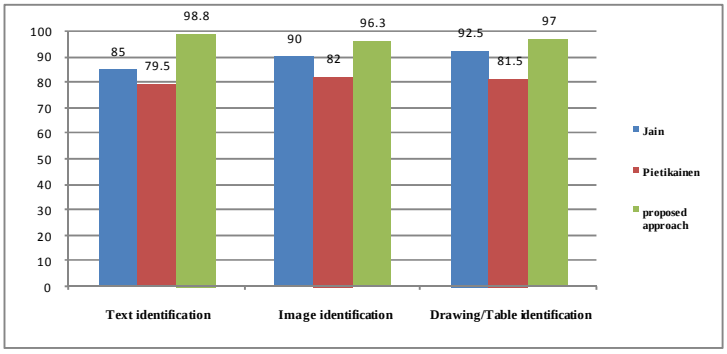

Figure. 8. The comparative results of the proposed method taking other artworks.

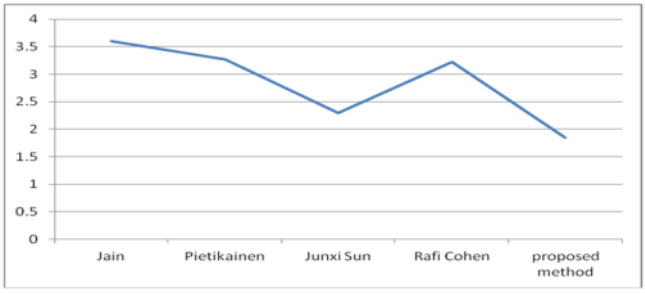

Figure. 9. The comparative results of the proposed method taking other artworks based on process time

Table 1. The results achieved for the proposed method.

	Total number of regions	**The number of regions found**	**The rate of regions found (%)**	**The number of correct regions found**	**The rate of correct regions found (%)**	**The number of unfound regions**	**The rate of unfound regions (%)**	**The number of incorrect regions found**	**The rate of incorrect regions found (%)**	**Max. overall error rate (%)**
Total region	448	440	98.2	435	97.1	13	*2.9*	5	*1.1*	*2.9*
Text region	382	379	99.2	376	98.4	6	*1*	3	*0.8*	*1*
Image region	43	42	97.7	41	95.4	2	*4.6*	1	*2.3*	*4.6*
Drawing/table	23	23	100	22	95.6	1	*4.3*	1	*4.3*	*4.3*

REFERENCES

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

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

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

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

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

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

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

[8] A. Jain & Y. Zhong, ªDocument segmentation Using Texture Analysis", (1996), Pattern Recognition, vol. 29, pp. 743-770.

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

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

[11] Junxi Sun, Dongbing Gu, Hua Cai, Guangwen Liu and Guangqiu Chen, "Bayesian Document Segmentation Based on Complex Wavelet Domain Hidden Markov Tree Models", Proceedings of the 2008 IEEE International Conference on Information and Automation June 20 -23, 2008, Zhangjiajie, China, pp. 493-498.

[12] Rafi Cohen, Abedelkadir Asi, Klara Kedem, Jihad El-Sana, Itshak Dinstein, "Robust text and drawing segmentation algorithm for historical documents", Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing,(2013), pp. 110-117.

[13] M. Viswanathan, "Analysis of scanned documents - a syntactic approach,". (1992), in Structured Document Image Analysis, Springer-Verlag, pp. 115–136.

[14] Mitchel P. E & Yana H., "Newspaper layout analysis incorporating connected component separation", (2004), Image and Vision Computing, vol. 22, pp. 307–317.

Authors

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.

Parisa Sheykhi Hesarlo was born in shahindezh, Azarbayjane Gharbi, Iran, in 1992. She is B.Sc. student in Pnu univesity in Computer Engineering field.