World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:11, 2013

Ultra High Speed Approach for Document Skew Detection and Correction Based On Centre of Gravity

Seyyed Yasser Hashemi

***Abstract*—**Skew detection and correction (SDC) has a direct effect in efficiency and exactitude of documents' segmentation and analysis and thus is considered as a very important step in documents' analysis field. Skew is a major problem in documents' analysis for every language. For Arabic/Persian document scripts this problem is more severe because of special features of these languages. In this paper an efficient and fast algorithm for Document Skew Detection (DSD) based on the concept of segmentation and Center of Gravity (COG) is proposed. This algorithm is examined for 150 Arabic/Persian and English documents and SDC process are done successfully for 93 percent of documents with error rate of less than 1°. This algorithm shows better results for English documents compared to Arabic/Persian documents. The proposed method is also represents favorable results for handwritten, printed and also complicated documents such as newspapers and journals even with very low quality and resolution.

***Keywords*—**Arabic/Persian document, Baseline, Centre of gravity, Document segmentation, Skew detection and correction.

I. Introduction

OVER the past three decades, several different works for DSD method are reported [1]-[3]. Skew occurs due to careless scan or careless writing in the handwritten documents. This situation may affect the efficiency of document analysis algorithms, since most of these methods are designed with the assumptions that they are to apply to non-skewed documents. Therefore, documents' skew detection and correction is necessary for efficient document analysis.

The existing methodologies can be broadly classified into categories, viz., projection profile based [4]-[6], Hough Transform based [7]-[16], morphological based [17]-[19] and nearest neighbor clustering based [20]-[24].

A projection profile of a document image is a one-dimensional array with a number of locations equal to the number of black pixels in the corresponding row of the image. Reference [6] used the sum of squared differences in adjacent cells of the projection profile created over a range of angles to estimate the skew angle of document. Reference [25] improved the time computation of [6] by obtaining projection profile of centers of gravity of connected components. Reference [26] in a similar method to [25] used bottom points of connected components to create projection profile over a range of angles. This method was faster than [6] and works well within ±15°. Another alternative method to reduce the time computation of [6] was proposed by [25], in which the image was down-sampled before calculating the projection profile at different degrees. The angle that maximizes the variance of the number of black pixels in a scan line is chosen as the skew angle of the document image .

Akiyama et al. in [1] divided an image into swaths of fixed heights then a vertical projection profile was calculated within each swath. This method is limited only to documents with vertically printed text liner (i.e. Chinese script) and fails to detect the skew angles above 5°. However, all projection profile based methods are sensitive to noise and are time consuming. The projection profile based methodologies are capable of detecting skew angles only up to ±15°.

Another class of skew detection algorithms is based on Hough transformation. The Hough transform maps points from (x,y) domain to curves in (r,θ) domain, where r is the perpendicular distance of the line from the origin and θ is the angle between perpendicular line and horizontal axis in (x,y) plane. Crossing curves in (r,θ) domain are result of collinear pixels in (x, y) plane. Reference [15] varied value of θ between 0 and 180 degrees for each black pixel in a document image and calculated value of r in Hough space. The maximum value in (r,θ) space is considered as skew angle of document image. The method [15] was time consuming due to mapping operation from (x,y) plane to (r,θ) plane for all black pixels, especially for images containing non-text dominant area (i.e. pictures, graphs etc). The artwork reported in [13] improved computational time of [15] by applying Hough transform only on bottom pixels of connected components belonging to dominant text area. This method shows accuracy of 0.5° to the original skew angle of the image. A method superior to [13] was proposed by [12] in which, each connected component was enclosed by conducting BAG (block adjacency graph) for it, later the document was divided into segments and the angles between bounding boxes in each segment which are located in range of [-45°,+45°] are accumulated in a histogram. The pixels co-ordinates corresponding to the peak of the histogram are subjected to Hough transform [12] show the average error of 0.06° and is faster than [13]. Reference [9] improved the run time of [15] in a different way from [13] by considering every twentieth black pixel vertically and horizontally in order to convert from

Seyyed Yasser Hashemi is with the Department of Computer Engineering, miyandoab Branch, Islamic Azad University, miyandoab, Iran (e-mail: hashemi.uni@gmail.com).

(x,y) domain to (r,θ) domain.

Reference [10] segmented the image into paragraph-like blocks and then applied Sobel operator to detect the segmented block edges. Later Hough transform was applied on the detected edges to estimate the skew angle. This method is accurate within ±10°. Generally Hough transform based methods are robust enough to estimate skew angles between -15 to +15 degrees. But they are computationally expensive and sensitive to noise. Applying morphological operation on document images has been used in few attempts [17] smeared the image text horizontally and produced a connected component for each line of document. Later, a line fitted to the pixels in each connected component and the skew angle was obtained from the histogram of all fitted lines. In another approach, [19] detected the number of ascender and descender pixels by applying a set of detector masks on a dilated and eroded image, the founded number was used to estimate the orientation of document image. Thinning and morphological dilation and erosion on blotched image was applied by [17] to estimate a wide range of skew angles particularly in Persian documents.

The other method of skew detection is based on nearest neighbor clustering. Reference [20] first proposed a nearest neighbor based method, in which a histogram of angles between nearest neighbors of connected components belonging to document text lines was obtained. The peak of histogram was chosen as skew angle of document; however, [20] fails to estimate small skew angles (near 0°) and large skew angles (near 45°). Reference [23] generalized [20] by considering five neighbors of each connected component. The angles of such neighbors are accumulated in a histogram and a gross estimate of skew angle is obtained by determining the angle that has the maximum value in histogram. The gross estimate shows accuracy of 0.5° to the skew angle of document. Reference [22] introduced size restriction to detect the nearest neighbors of a connected component. The median value of angles of all nearest neighbors is chosen as skew angle. Experimental result shows that [22] outperforms [20] in all cases. In fact, the nearest neighbor clustering based methods are fast and able to estimate skew angles within ±45° but the accuracy is highly dependent on structure of script characters. Characters with vertical expansion on both sides of base line (i.e. j, y, p and q in English script) tend to reduce the accuracy of these methods due to detection of nearest neighbors, which are not preserving the skew angle of document. Nevertheless, the effect of such characters can be neglected in English Documents since the number of vertically expanded characters is few, but in scripts that are rich in such characters (i.e. South Indian and Persian scripts), the nearest neighbor clustering fails to detect the skew angle.

The Persian script (or called Farsi) has 32 characters and it is cursive in nature. The words in Persian script are written from right to left where as the numbers from left to right. Most of the characters are vertically expanded on both sides of the base line and each character can have up to four forms as below.

Detached -The character as it appears by itself, no character joining to it either before or after.

Initial -The character as it appears when preceded by a joining character.

Medial -The character as it appears when there are joining characters both before and after the letter.

Terminal -The character as it appears when it is preceded but not followed by joining character

Although the skew problem in English documents have been largely resolved, but for images with complicated charts and graphs and different amounts of skew in consecutive paragraphs the problem still has not been fully resolved. In Arabic/Persian documents this problem becomes excessively acute because most algorithms used in English documents do not provide good results for Arabic/Persian documents due to abovementioned different structure of text in these languages.

An efficient and very fast method for Arabic/Persian DSD based on COG is proposed in this paper.

This paper is organized as follows. The proposed algorithm is described precisely in the following Section. Section III includes the experimental results of the proposed algorithm. Discussions are given in Section IV. Finally, this paper is concluded in Section V.

II. New Skew Detection Algorithm

A. Overview

The proposed algorithm uses the concept of COG of the polygon to determine Arabic/Persian document skew. The skew angle detection is generally done in two steps. First step is the Baseline Identification (BI) and the other is the Skew Angle Correction (SAC). The BI is generally the most important step of the whole process. Baseline is the line that the COG of the word hangs to it. 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), identifying its COG and identify the baseline of document. The angle between the baseline and horizontal lines specifies the skew angle.

B. Connected Component Analysis

In order to determine the ARD, four corners of the main document in the skewed document must be recognized. For this purpose, the Connected Component (CC) analysis should be done. This algorithm uses a bottom-up approach to extract CCs from the document image. The first step in CCs extraction is to locate rectangular regions called Rects [26]. A Rect is thought to be a rectangular region of loosely connected black pixels. More specifically, a Rect has at least one black pixel in every 9-pixel square. A Rect 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. For implementation, Rects are stored in a linked list. This type of

structure is preferable 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 the image in raster order. The top left corner of a Rect is defined to be the first 9-pixel square found containing at least one black pixel. The right edge of this Rect is located by searching toward 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 [26]. 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 Rect in the Rect list. Each Rect is subsequently added to a single CCs or used to define a new CCs. If a Rect is found to be adjacent to more than one CCs, 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. An example of a skewed document image and the result from segmentation is shown in Figs. 1 and 2, respectively.

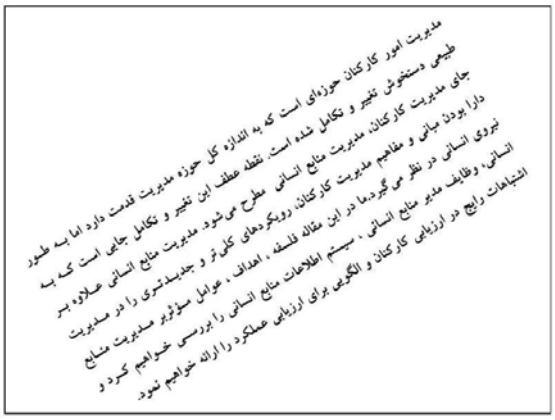


Fig. 1 Skewed document image

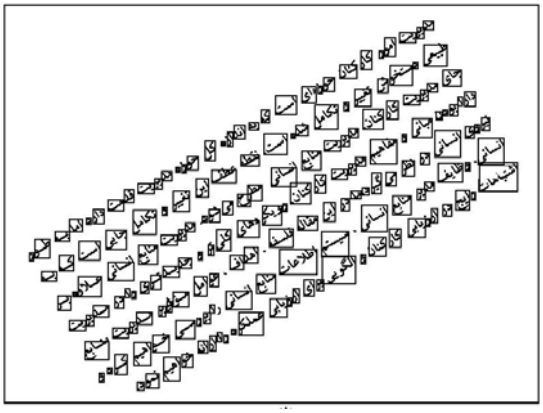


Fig. 2 Document segmented to connected component

Noise is the most important challenge for document analysis algorithms so a noise filtering will be performed on the result of CCs analysis in order to remove the noises resulted in scanning step.

C. Determine the Actual Region of Document

After segmentation, coordinates of four corners of ARD in skewed document denoted with (xi,yi) in Fig. 3 must be determined. 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. C0, C1, C2, and C3 are four corners of the original document. Selected CCs determine the coordinate of the ARD in skewed document. Fig. 3 shows the coordinate of the ARD in skewed document that is segmented in Fig. 2.

Therefore, the coordinate of the ARD in skewed document are identified in three steps as follows:

- Document segmentation to CC.
- Filtering CCs to remove small CC which is considered as noise.
- To identify four CC that have the more distance from the C0, C1, C2 and C3 corners as the coordinates of ARD in skewed document.

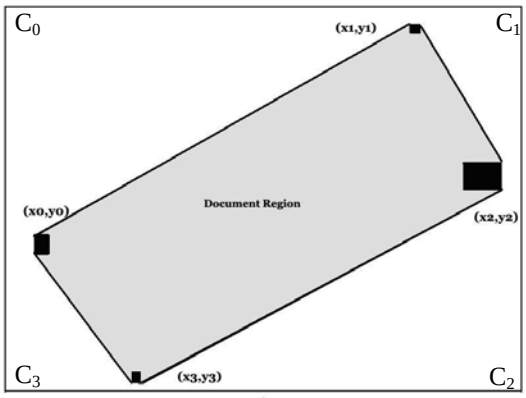


Fig. 3 The ARD in skewed document

D. Determine the COG

The ARD would be a polygon shape. In this step, the COG of the Polygon that obtained from identifying ARD is calculated. The COG is also known as the "center of mass". The position of the centroid assuming the polygon to be made of a material of uniform density is given below. In Fig. 4, a hesagonal is shown and COG is calculated using (1):

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

$$COG_Y = \frac{1}{6A}\sum_{i=0}^{N-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (2)$$

'A' is the polygon's area that is calculated from the following equation.

$$A = \frac{1}{2}\sum (x_i y_{i+1} - x_{i+1} y_i) \quad (3)$$

ARD and its corners' coordinates and the COG point are

shown in Fig. 5. The calculated COG is the COG of de-skewed document.

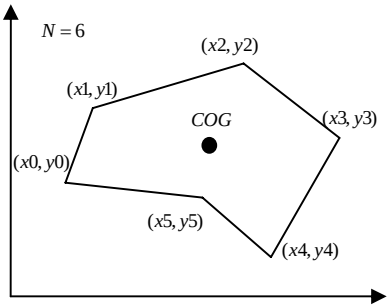


Fig. 4 COG of hesagonal

E. Calculate the Skew Angle

In this step, after identifying the actual coordinates of the document and its COG, the amount of document skew angle will be calculated. For document skew angle detection, we will draw a line (baseline) from COG to the center of the line that connect the two upper and lower left corner of the ARD (midpoint). The angle between baseline and the horizontal line that passes through the midpoint reveals the document skew angle. COG and baseline of document is shown in Fig. 5 and the document skew angle is shown in Fig. 6. A skewed document is inputted into the proposed algorithm and a de-skewed document is outputted. The steps for this algorithm are:

- Document segmentation (CCs identification)
- Identification of the actual region of document
- Finding the center of gravity.
- Baseline identification
- Calculation of the amount of document skew angle
- Rotation of the document (see Fig. 5 which is the rotated version of Fig. 1)

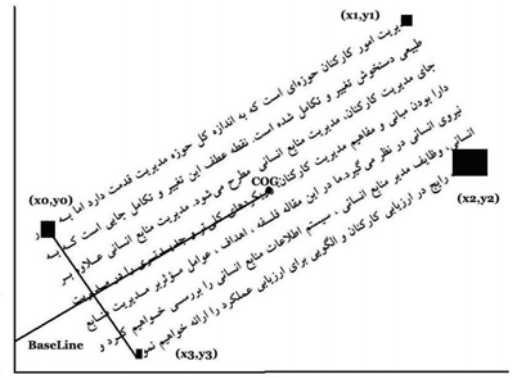


Fig. 5 Document's COG

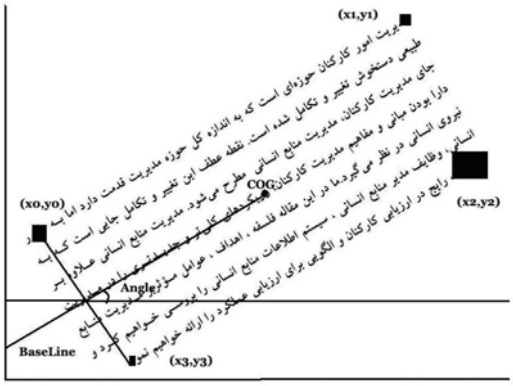


Fig. 6 Calculate the amount of document skew angle

Fig. 7 Deskewed document with Proposed method

III. Results

The proposed method has been implemented on duo core 2.0 GHZ in 2010. We have considered different skewed documents from different sources like journals, textbooks, newspapers and also handwritten documents. For experimentation purpose 150 documents are considered, 50 of them are handwritten documents. The result that obtained by using the proposed method are reported in Table I. Fig. 9 shows the images of the corrected document after applying the proposed method on document images shown in Fig. 8.

IV. Discussion

SDC algorithms can be considered in terms of simplicity, generality and the feasibility. The proposed method is simple and works with both the handwritten and the printed documents using the same procedures while the Hough Transforms and Projection Profile cannot work efficiently for both the handwritten and the printed documents when using the same procedures. The proposed method can also find the skewed angle of deviation of the different kinds of printed documents such as magazines or books and other printed documents, as shown in Figs. 8 (a) and (b), as well as handwritten documents, as shown in Fig. 8 (c). This method is also capable to work with documents with different

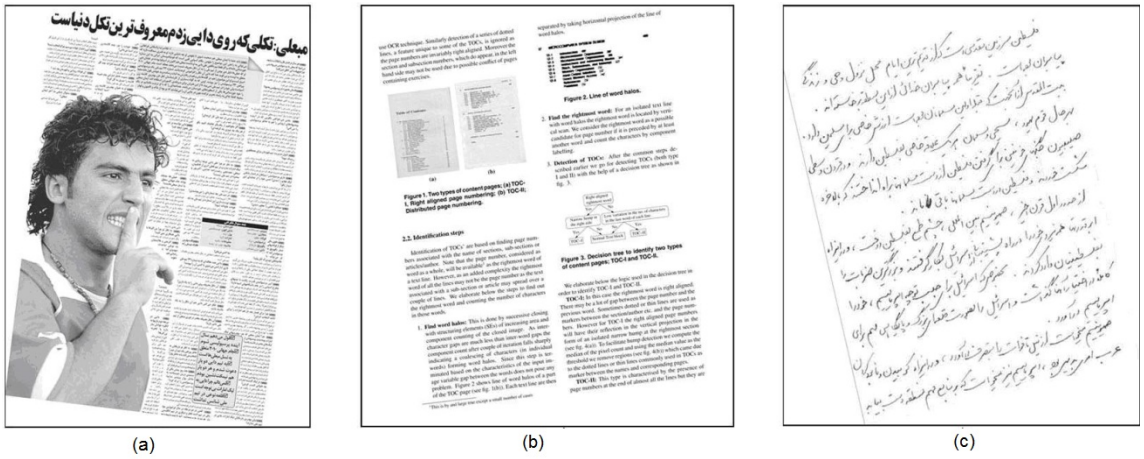
resolutions, as shown in Fig. 9. While on the other hand, other methods usually are designed to find a deviation skewed angle of a certain type of documents with constant resolutions. The noise does not affect the proposed method's performance, while it has very high influence in other methods such as the Hough Transform and Projection Profile.

V. Conclusion

In this paper a novel and accurate method to estimate skew angle is presented. The proposed efficient, simple and fast method works based on COG and document segmentation. With document segmentation and appropriate choice of four corners of connected component as the coordinates of the ARD in skewed document and calculating the COG and determining the baseline, the document's skew angle is calculated. This method works well for different documents with different degrees of skew and complexity as well as for English and Arabic/Persian printed and handwritten documents. On the other hand document's low quality does not affect performance of the proposed method. This method was applied to 150 different documents (90 Arabic/Persian and 40 English and 20 hybrid of English and Arabic/Persian) and the rate of accuracy with error rate of less than 1 degree is 93%.

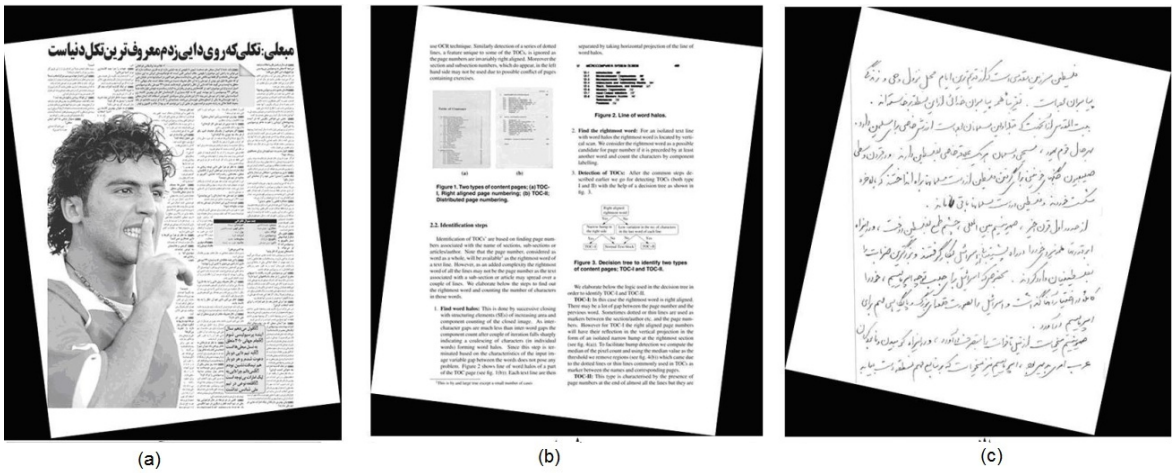
TABLE I
The Results Obtained for Proposed Method Using 150 Documents

Document Type	*Number of documents*	*number of correct detection within error tolerance*		*Percentage of correct detection within error tolerance*	
		<.5	*<1*	*<.5*	*<1*
Arabic/Persian document	20	18	19	90	95
Complex Arabic/Persian document	70	61	64	87	91
English document	10	9	10	90	100
Complex English document	30	24	26	86	90
Hybrid document	20	16	18	85	90



(a) (b) (c)

Fig. 8 Examples of the document (a) Arabic/Persian Handwritten documents (b) Journal; (c) The Arabic/Persian newspaper



(a) (b) (c)

Fig. 9 Corrected documents image after applying the proposed method

REFERENCES

[1] Akiyama T, Hagita N. Automated entry systems for printed documents. *Pattern Recognition* 1990; 23: 11; 1141–1154.

[2] Bloomberg D S, Kopec G E, Dasari L. Measuring document image skew and orientation. *Document Recognition II (SPIE vol.2422)*1995;302–316.

[3] Chevillat P, Schindler H R. Arrangement for determining the optimum scan angle for printed documents. *International Business Machines Corp (U.S Patent 4)*1982; 338–588.

[4] Ishitani Y. Document skew detection based on local region complexity. *The Second International Conference on Document Analysis and Recognition* 1996; 49–52.

[5] Pavlidis T, Zhou J. Page segmentation and classification. *Graphical models and image processing*. 1992; 54: 6;484–496.

[6] Postl W. Detection of linear oblique structures and skew scan in digitized documents. *8th International. Conference. on Pattern Recognition(Paris)*1986,687–689.

[7] Amin A, Fischer S. A document skew detection method using the Hough transform. Pattern *Analysis and Applications*. 2000;3: 3;243–253.

[8] Farrow G S D. Ireton M A, Xydeas C S. Detecting the skew angle in document images. *Signal Processing: Image Communication*, 1994;**6**:6;101–114.

[9] Farrow G S D, Xydeas C S. Detecting skew in digitized images. *Int Computers Ltd., London, European Patent App* 1992;485–491.

[10] Ham Y K, Chung H K, Kim I K, Park R H. Automated analysis of mixed documents consisting of printed Korean alphanumeric texts and graphic images. *Optical Engineering*. 1994;33:6;1845–1853.

[11] Hind S C, Fisher J L, D'Amato D P D. A document skew detection method using run-length encoding and the Hough transform. *Proceedings of the 10th International Conference on Pattern recognition(Atlantic City, New Jersey)*1990;464–468.

[12] Kwag H K, Kim S H, Jeong S H, Lee G S. Efficient skew estimation and correction algorithm for document images. *Image and vision computing*. 2001; 20:1;25–35.

[13] Le D S, Thoma G R, Wechsler H. Automated page orientation and skew angle determination for binary document images. *Pattern Recognition*. 1994;27:10;1325–1344.

[14] Lee Y. Method of detecting the skew angle of a printed business form. *Eastman Kodak Company, U.S. Patent 5,054,098;*1991.

[15] Srihari S N, Govindaraju V. Analysis of textual images using the Hough transform. *Machine Vision Applications*. 1989;2:3;141–153.

[16] Yu B, Jain A K. A robust and fast skew detection algorithm for generic documents. *Pattern Recognition*. 1996;29:10;1599–1630.

[17] Ashkan M Y, Guru D S, Punitha P. Skew estimation in Persian documents: A novel approach. *Proceeding of International Conference on Computer Graphics, Imaging and Visualization (CGiV'06)*2006;64–70.

[18] Chen S, Haralick R M. An automatic algorithm for text skew estimation in document images using recursive morphological transform. *IEEE International Conference on Image Processing (ICIP94)*1994;139–143.

[19] Dasari L, Bloomberg D S. Rapid detection of page orientation Xerox Corporation. *U.S. Patent 5276742*;1994.

[20] Hashizume A, Yeh P S, Rosenfeld A. A method of detecting the orientation of aligned components. *Pattern Recognition Letters.* 1986;4:2;125–132.

[21] Liu J, Lee C M, Shu R B. An efficient method for the skew normalization of a document image. *11th International Conference on Pattern Recognition (IAPR)* 1992;152–155

[22] Yue L Y, Chew L T. A Nearest neighbor chain based approach to skew estimation in document images. *Pattern Recognition Letters*. 2003;24:14;2315–2323.

[23] O'Gorman L. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1993;15:11;1162–1173.

[24] Okamoto M, Twaakyondo H M, Nishizawa H. Skew detection, skew normalization and segmentation of document images using segmented block code. *Journal of the Faculty of Engineering*. 1988;1:1;9–18.

[25] Bloomberg D S, Kopec G. Method and apparatus for identification and correction of document skew. *Xerox Corporation, U.S. Patent 5,187,753*; 1993.

[26] Mitchel P E, Yana H. Newspaper layout analysis incorporating connected component separation. Image and Vision Computing. 2004;22:4;307–317.

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.