Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (12): 2265-2271, 2008
ISSN 1812-5654

On Skew Estimation of Persian/Arabic Printed Documents

A.E. Sharif and N. Movahhedinia
Department of Computer, Faculty of Engineering,
The University of Isfahan, Isfahan, Iran

Abstract: In this study, we propose two methods especially designed to detect skew in Persian/Arabic prints. The first one is based on vertical Black Line Segments (BLSs) which are the baselines of horizontal projection profiles of vertical strips. This method is fast and presents good accuracy against practical skews. The second method, which is based on vertical White Line Segments (WLSs), offers superior performance in recognizing large skew angles.

Key words: Document image analysis, text line extraction, projection profile, cursive language, least square line

INTRODUCTION

The first step in processing a document by computer is to create the digitized version of it, which is usually done by a scanner device. However, human fault (or document feeder impairment) in placing the document on the platen may lead to some skew in the document image. Being so, a proper measurement against the skew should be considered before document reorganization. The skew detection procedure may be performed at one of the following levels:

- **Page/Text blocks level:** In most machine printed documents, entire page is skewed. Therefore, skew detection may be done before page segmentation. In some documents such as advertisements, however, text blocks are printed in different orientation intentionally. Hence, skew correction should be postponed until blocks are classified. Almost all methods using this strategy use the fact that text lines in a document image remain parallel even in presence of skew.
- **Group level:** In this case, the slant should be corrected for each group (a connected component, a character, a word or part of a word). It is generally applicable to documents printed in a cursive language or handwritten ones. It is suitable, albeit slow, to detect local skews.
- **Feature level:** Document recognizers, using this strategy, utilize some rotation invariant features for recognition, so it is not necessary to correct the skews. However, it is rarely used in practice because the number of rotation invariant features is not high enough to lead to strong results.

The proposed methods estimate skew at page level. Researchers have introduced many methods for this purpose. These methods are usually categorized based on the techniques they utilize, such as projection profiles, Hough Transform, cross correlation or gradient analysis (Cattoni *et al*., 1998; Hull, 1998; O'Gorman, 1993; Okun *et al*., 1999). Considering the main idea behind them, we classify these methods into four following classes:

I Methods that choose a set of representative points (mostly related to text lines) and identify explicitly or implicitly few parallel straight lines using some line fitting methods (Antonacopoulos, 1997; Hashizume *et al*., 1986; Hinds *et al*., 1990; Nakano *et al*., 1990; Smith, 1995; Yu and Jain, 1996).
II Methods that extract some orientation sensitive features and use them to calculate the skew (Li *et al*., 2007; Chou *et al*., 2007; Dong *et al*., 2005; Amin and Wu, 2005; Kapoor *et al*., 2004; Kapogiannopoulos and Kalouptsidis, 2002; Rundle, 1974; Sauvola and Pietikäinen, 1995; Sun and Si, 1989).
III Methods that define some target criteria which are maximized or minimized to alleviate the skew angle (Baird, 1987; Ishitani, 1993; Postl, 1986; Srihari and Govindaraju, 1989).
IV Other methods that are not placed in the above classes such as ones that use hardware to detect page properties (Aghajan *et al*., 1994; Yamada, 1989).

Each of the above classes has its advantages and disadvantages. Methods of class I usually use bounding blocks of connected components to select representative points. As locating connected components is time-consuming, they are not generally time efficient,

Corresponding Author: Naser Movahhedinia, Department of Computer Engineering, Faculty of Engineering,
The University of Isfahan, Isfahan, Iran Tel: +98 311 7934103 Fax: +98 311 7932670

especially when Hough Transform is used for line fitting. However, they usually achieve accurate skew estimation. Calculating orientation sensitive features may not be cost effective, albeit, class II methods are typically faster than class I methods. One disadvantage of this class is inaccuracy of the results in most of the cases. Class III requires a search mechanism to find minimum or maximum of the target criteria which is generally time-consuming.

Detecting skew in a document image prepared by a cursive language is so complex that many existing methods may not present adequate performance. In such documents, words may consist of scattered pieces that are located above or below base lines. Moreover, the width and the height of connected components are considerably different. Furthermore, documents written in these languages usually are less structured than others. Class I methods have usually worst results in case of cursive text, as they do classification (in text and non-text blocks) based on the size of bounding boxes or other features that are not appropriate for cursive languages.

Basically, skewed cursive language documents are treated in two manners: as generic documents or as cursive handwritten ones. Considering these documents, as handwritten and detecting skew at non-page level is not favorable, since the structured-ness of these documents, such as presence of a base line for each text line, is ignored. As such, there is a need to design new skew detection algorithms for such documents.

Persian and Arabic are from the most used cursive languages and are similar in many aspects such as writing styles. Researcher may refer to Hadjar and Ingold (2003) for detailed information about Arabic language.

Although many papers have been published on skew estimation, few of them are related to Arabic or Persian languages. Amin (2000) has introduced a method that estimates skew at group level. At first, connected components are extracted from the document image and then they are grouped based on some criteria. Afterwards, each group is divided into vertical segments of approximately the width of one connected component and only the bottom rectangle in each segment is stored. At last, Hough Transform is utilized to detect skew angle.

Sarfraz *et al.* (2003) have proposed a procedure named Drift Correction, which first determines the rotation angle of the text by computing the tangents of all the line segments that can be constructed between any pair of black pixels in the image. The angle that has the highest number of occurrences is assumed to be angle of skew for the image. Sarfaraz *et al.* (2005) have utilized Haar wavelet to decompose image into detail sub images and have used multi-scale properties of the image along with Principal component analysis to estimate the orientation of principal axis of clustered data.

To deal with document printed in these two languages, we propose two methods which belong to class I, but are quite fast and accurate. These methods use the horizontal projection profiles of the strips rather than features extracted from connected components.

THE FIRST PROPOSED METHOD

First proposed method is based on the vertical black line segments. This section delivers the details of this method.

A well-known technique in document image analysis is splitting document page into narrow horizontal/vertical strips and then calculating vertical/horizontal projection profiles of these strips, respectively (Akiyama and Hagita, 1990; Min *et al.*, 1996; Pavlidis and Zou, 1991; Schlang, 1985). Here, we employ vertical strips and their horizontal projection profiles (HPPs). The strips should be wide enough to contain useful information. Research show that the width of two small-sized or one medium-sized character (roughly 24 pixels in 100 ppi images) yields acceptable results. The advantage of using HPPs of these strips rather than analyzing HPPs for entire page is that the parts related to the text lines do not attach to the adjacent lines, if strips are not too wide and the skew is not severe.

The baselines of HPPs of strips, which we refer to them as vertical Black Line Segments (BLSs), are the building blocks of our method. In Fig. 1, a typical Persian document and its extracted BLSs are shown.

Filtering the BLSs: We need to identify the BLSs corresponding to the text part of the document image. For this purpose, the conditions are checked for each BLS:

$$T_{BLS} < D \quad (1)$$

$$|T - T_{BLS}| = 0.5T \quad (2)$$

where, T_{BLS}, D and T denote the thickness of text line corresponding to a BLS, the dominant interline distance and the dominant text line thickness, respectively. If all the conditions are satisfied, the BLS is considered as related to the text parts, otherwise it is discarded:

As an estimate of T, the peak of the histogram of the length of the BLSs is considered. D is obtained approximately using the histogram of the distances between consequent peaks of HPPs corresponding to the BLSs. It is worthy of mention that text written in Arabic language has a salient peak related to baselines in their HPPs (even in presence of a little skew).

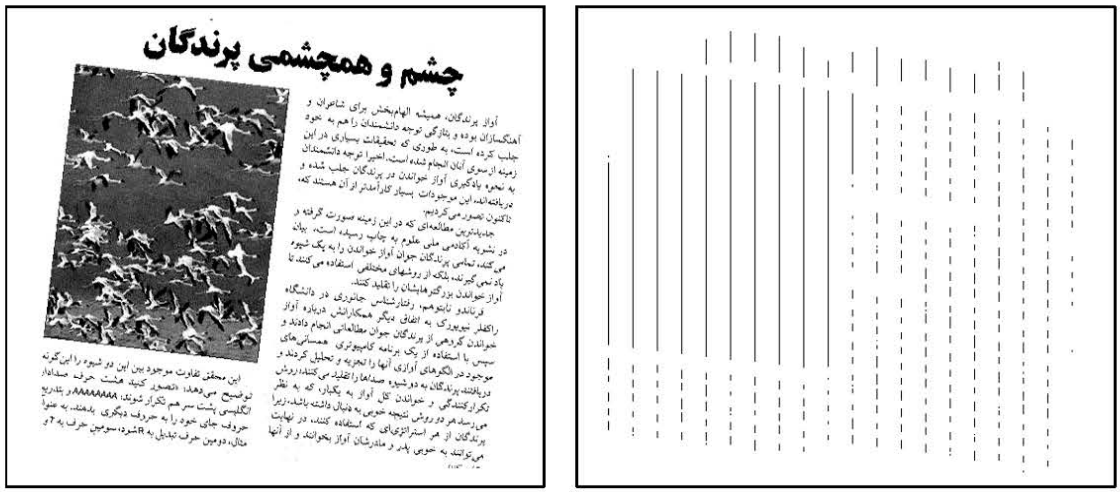

Fig. 1: A typical Persian document and its black line segments

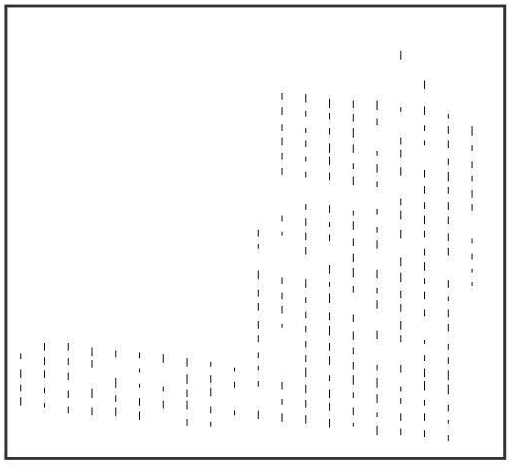

Fig. 2: The BLSs of Fig. 1 after filtering

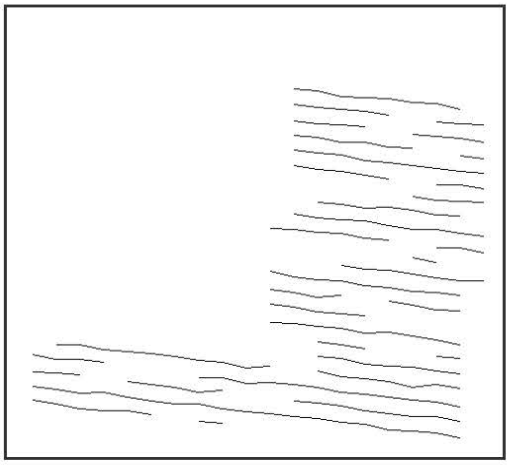

Fig. 3: The extracted text lines for the document of Fig. 1

Applying the mentioned rules, those BLSs would be left over that are more likely related to the text lines. Figure 2 is obtained from BLSs of Fig. 1, using this method.

Assigning the BLSs to text lines: The next step is to identify the text lines from the determined BLSs. This is performed by assigning an incremental number to each BLS as its text line number and then considering the BLSs with the same number as a text line. The idea behind this procedure is that: if two BLSs in two adjacent strips are the nearest ones to each other, they are related to a text line. Note that the word nearest is ambiguous. In this research Manhattan distance of the middle points of two BLSs used as a measure of their nearness. Assume A and B are two BLSs in two adjacent strips. Suppose that A_{mid} and B_{mid} are the middle points of A and B, respectively, then the distance between A and B is calculated as:

$$\text{Distance}(A, B) = |X_{Amid}-X_{Bmid}| + |Y_{Amid}-Y_{Bmid}| \quad (3)$$

Since, A and B are located in adjacent strips, the first term of the above equation is equal to the strips width which is a constant number. As this function is used only for comparison, one can ignore the first term and obtain:

$$\text{Distance}(A, B) = |_{YAmid}-Y_{Bmid}| \quad (4)$$

The above procedure links the BLSs with the same text line numbers together to obtain the extracted text lines, as shown in Fig. 3.

Extracting representative points: In the next stage, we need to select a number of representative points for each text line. The midpoints of the BLSs related to the same text line may be chosen for this purpose. However, a better choice is the intersection points of diameters of trapezoids constructed with consequent BLSs, which are related to the same text line.

Once the representative points are identified, Least Square Error method is used to fit straight lines to the points related to each text line. The majority vote of the gradient of these lines identifies the skew angle. Since the gradients are not necessarily an integer number, quantization is unavoidable. Experiments show that by considering 0.5° as the step size of quantization, the results are acceptable.

THE SECOND METHOD

Although the BLS method is fast, the size of skews detected by this method does not satisfy the need of some applications. This limitation arises from two reasons:

- The way BLSs are built: If the skew is increased, we should make the strips more narrow to prevent joining HPPs. But as mentioned earlier, the strips can not be too narrow.
- The manner text lines are extracted: Try to extract text lines and use them to determine the skew, any error in text line identification would lead to significant error.

To overcome these limitations this research propose a second method which is called WLS. In this method we white pixels are used rather than black ones to detect representative points. Furthermore, this study introduces a new method to detect straight-lines from the representative points.

The WLS method: The WLS method is on the basis of the following facts in Persian/Arabic text:

- Many parts of the text are placed on baselines. These parts are related to some letters (mostly, initial or medial letters) or to the linkage between them.
- The baselines are parallel and they remain parallel even in presence of skew. Therefore the vertical distances between two text lines are almost constant and are equal to Dominant Interline Distance (DID).
- DID is much larger than pen thickness. So, if we use vertical White Line Segments (WLSs) as new building blocks, it's no need to find a way to get rid of dots and diacritic. Note that in BLS method HPPs use to avoid this problem.

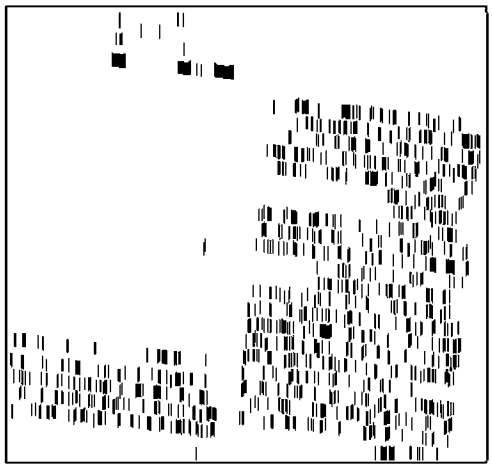

Fig. 4: The white line segments (WLSs) for the document of Fig. 1

DID is calculated as the most occurrence length of WLSs. Moreover, to remove false WLSs, the ones which are very shorter or longer than DID are ignored (Fig. 4).

It is worthy to notice that both BLS and WLS methods result in blank blocks for non-text parts of a document image. So, one may use these two methods to categorize the text and non-text image blocks.

Extracting representative points: Similar to BLS method, the representative points are the set of pixels that are generally placed on the baselines. To define the representative points, the midpoints between the bottom points of WLSs related to a text line and the top points of WLSs related to the line below are selected.

To determine whether the line B is below a selected text line A, the following conditions are checked:

$$Y_{top\_point\_of\_B}\ Y_{bottom\_point\_of\_A} \quad (5)$$

$$\text{Euclidean\_Distance(bottom\_point\_of\_A, top\_point\_of\_B)} \le \text{DID} \quad (6)$$

The first condition guarantees that B is related to one of text lines below A. The second condition restricts the distance between A and B to ensure that B is exactly related to the next text line of A.

To attain higher speed, the number of representative points related to a WLS can be reduced to one. Experiments show that such reduction still leads to acceptable results (Fig. 5).

Skew detection: To detect skew from the selected points, it is proposed a procedure which we call Slop Seeker (SS).

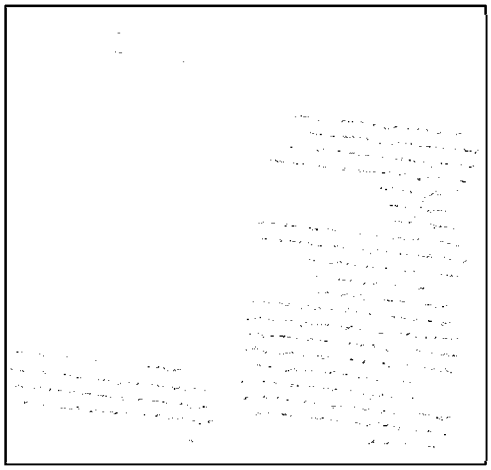

Fig. 5: The representative points taken from WLS of the document in Fig. 1

This procedure is as accurate as Hough Transform, yet faster than that. To formulate the procedure, suppose that $P_1 \ldots P_n$ denote the n representative points and let $L_{i,j}$ denote the line crossing P_i and P_j and $\theta_{i,j}$ be the slope of it. $\theta_{i,j}$ can be calculated by the following formula:

$$\theta_{i,j} = \tan^{-1}((Y_{Pj}-Y_{Pi})/(X_{Pj}-X_{Pi})) \quad (7)$$

Consider $C_{i,\alpha}$ to be the number of lines crossing P_i and their quantized values of slops are all equal to the integer $|\alpha|$. Let M stand for maximum degree that can be detected by our method. Now, to intensify the score of collinear slope values, the energy function of $C_{i,\alpha}$ is calculated as:

$$D_a = \sum_{i=1}^{n} C_{i,\alpha}^{\ 2} \quad 1 \le i \le n,\ 1 \le \alpha \le M \quad (8)$$

The maximum value of that energy function gives the estimate of skew:

$$\text{Skew} = \theta \text{ where } D_\theta = \max(D_\theta) \quad (9)$$

EXPERIMENTAL RESULTS

To evaluate performance of the proposed methods, 455 document images were considered from different sources such as magazines, newspapers, advertisements and books. To reduce the run time, the images are down-sampled to 100 ppi. Then, the images were rotated from −45 to +45° with 0.5° step size. The performances of BLS and WLS have been evaluated against the imposed skew values.

Table 1: Experimental results for WLS schemes

Method	Skew range	Average run time (msec)	Max error (degree)
Black Line Segments (BLS)	(-15°,0°), (0°,15°)	17	±1.0
WLS and Slop Seeker	(-15°,0°), (0°,15°)	38	±0.5
WLS and Hough Transform	(-15°,0°), (0°,15°)	49	±0.5
WLS and Slop Seeker	(-30°,-15°), (15°,30°)	44	±0.5
WLS and Hough Transform	(-30°,-15°), (15°,30°)	82	±0.5
WLS and Slop Seeker	(-45°,-30°), (30°,45°)	54	±1.0
WLS and Hough Transform	(-45°,-30°), (30°,45°)	121	±1.0

Furthermore, to have a comparison between the proposed Slop Seeker method and Hough Transform, both of them have been employed with conjunction to WLS method (Table 1). The experiments show that BLS, although faster, fails against skews larger than 15°. Both methods of WLS and Slop Seeker and WLS and Hough Transform present satisfactory performance, appropriate to be applied to documents containing mixed text, graphic or line drawing. However, Slop Seeker performs in much higher speed especially for large skew estimations.

Comparing to a the recent Arabic skew estimation method, WLS method detects skews up to 45° with 1° error, which is much superior to Sarfaraz *et al.* (2005) algorithm which detects 10.2627° skew in Arabic document with -0.2627° error.

CONCLUSION

In this study, two methods for skew estimation of printed documents in Persian or Arabic languages are proposed. The first method (BLS), which is based on detecting black line strips, is fast and simple, yet accurate against practical skews, however, fails in case of skews larger than 15 degrees. The second method is based on White Line Strips (WLS) of the printed document and presents good performance against skews up to 45°. Moreover, we proposed Slop Seeker, an algorithm to estimate the skew, out of white line strips. Though as accurate as Hough Transform, Slop Seeker performs faster, especially for large skews. Both of the proposed methods can be applied to the document images with low resolution as 100 ppi. These methods can be used to categorize the blocks of a document image to text and non-text classes, as well.

REFERENCES

Aghajan, H.K., B.H. Khalaj and T. Kailath, 1994. Estimation of skew angle in text-image analysis by SLIDE: Subspace-based line detection. Mach. Vision Applied, 7: 267-276.

Akiyama, T. and N. Hagita, 1990. Automatic entry system for printed documents. Pattern Recog., 23: 1141-1154.

Amin, A., 2000. Recognition of printed Arabic text based on global features and decision tree learning techniques. Pattern Recog., 33: 1309-1323.

Amin, A. and S. Wu, 2005. Robust skew detection in mixed text/graphics documents. Proceeding of 8th International Conference on Document Analysis and Recognition, 29 Aug.-1 Sept. 2005, pp: 247-251.

Antonacopoulos, A., 1997. Local skew angle estimation from background space in text regions. Proceeding of 4th International Conference on Document Analysis and Recognition, 18-20 Aug., 1997, Ulm, Germany, pp: 684-688.

Baird, H.S., 1987. The skew angle of printed documents. Proceeding of SPSE 40th Symposium on Hybrid Imaging Systems, May, 1987, Rochester, New York, USA., pp: 204-208.

Cattoni, R., T. Coianiz, S. Messelodi and C.M. Modena, 1998. Geometric layout analysis techniques for document Image understanding: A review. Technical Report, IRST, Trento, Italy, pp: 1-68.

Chou, C.H., S.Y. Chu and F. Chang, 2007. Estimation of skew angles for scanned documents based on piecewise covering by parallelograms. Pattern Recog., 40: 443-455.

Dong, J., P. Dominique, A. Krzyyzak and C.Y. Suen, 2005. Cursive word skew/slant corrections based on Radon transform. Proceeding of 8th International Conference on Document Analysis and Recognition, (ICARD, 05) 2005, pp: 478-483.

Hadjar, K. and R. Ingold, 2003. Arabic newspaper page segmentation. Proceeding of 7th International Conference on Document Analysis and Recognition, August 3-6, 2003 Edinburgh, pp: 895-899.

Hashizume, A., P.S. Yeh and A. Rosenfeld, 1986. A method of detecting the orientation of aligned components. Pattern Recog. Lett., 4: 125-132.

Hinds, S.C., J.L. Fisher, D.P. D'Amato, 1990. A document skew detection method using run-length encoding and the Hough transform. Proceeding of 10th International Conference on Pattern Recognition, June 16-21, 1990, Atlantic City, N.J. USA, pp: 464-468.

Hull, J.J., 1998. Document Image Skew Detection: Survey and Anotated Bibliography. In: Document Analysis Systems II, Hull, J.J. and S.L. Taylor (Eds.). World Scientific, Singapore, pp: 40-64.

Ishitani, Y., 1993. Document skew detection based on local region complexity. Proceeding of 2nd International Conference on Document Analysis and Recognition, 20-22, October, 1993, Tsukuba, Japan, pp: 49-52.

Kapogiannopoulos, G. and N. Kalouptsidis, 2002. A fast high precision algorithm for the estimation of skew angle using moments. In: IASTED International Conference Signal Processing, Pattern Recognition and Applications, SPPRA, June 25, Crete, Greece, pp: 275-279.

Kapoor, R., D. Bagai and T.S. Kamal, 2004. A new algorithm for skew detection and correction. Pattern Recog. Lett., 25: 1215-1229.

Li, S., Q. Shen and J. Sun, 2007. Skew detection using wavelet decomposition and projection profile analysis. Pattern Recog. Lett., 28: 555-562.

Min, Y., S.B. Cho and Y. Lee, 1996. A data reduction method for efficient document skew estimation based on hough transformation. Proceeding of 13th International Conference on Pattern Recognition, 25-29 Aug. 1996, IEEE Press, Vienna, Austria, pp: 732-736.

Nakano, Y., Y. Shima, H. Fujisawa, J. Higashino and M. Fujinawa, 1990. An algorithm for the skew normalization of document images. Proceeding of 10th International Conference on Pattern Recognition, 16-21 June, 1990, Atlantic City, N.J. USA., pp: 8-13.

Okun, O., M. Pietikäinen and J. Sauvola, 1999. Document skew estimation without angle range restriction. IJDAR, 2: 132-144.

Pavlidis T. and J. Zou, 1991. Page segmentation by white streams. Proceeding of 1st International Conference on Document Analysis and Recognition, Sep. 30-Oct. 2, 1991 France, pp: 945-953.

Postl, W., 1986. Detection of linear oblique structures and skew scan in digitized documents. Proceeding of 8th International Conference on Pattern Recognition, 1986, Paris, France, pp: 687-689.

Rundle, A., 1974. Optimum Scan Angle Determining Means. International Business Machines, Inc., U.S. Patent 3,831,146.

Sarfraz, M., S. Nazim and A. Al-Khuraidly, 2003. Offline Arabic text recognition system. Proceeding of International Conference on Geometric Modeling and Graphics, 16-18 July, 2003, pp: 30-35.

Sarfraz, M., A. Zidouri, S.A. Shahab, 2005. A novel approach for skew estimation of document images in OCR system. Proceeding of International Conference on Computer Graphics, Imaging and Vision: New Trends, 26-29 July, 2005 pp: 30-35.

Sauvola, J. and M. Pietikäinen, 1995. Skew angle detection using texture direction analysis. Proceeding of 9th Scandinavian Conference on Image Analysis, 1995. Uppsala, Sweden, pp: 1099-1106.

Schlang, A., 1985. Text line bounding system. Litton Systems Inc., U.S. Patent 4,558,461.

Smith, R., 1995. A simple and efficient skew detection algorithm via text row accumulation. Proceeding of 3rd International Conference on Document Analysis and Recognition, (ICDAR, 95) 1995 Montreal, Canada, pp: 1145-1151.

Srihari, S.N. and V. Govindraju, 1989. Analysis of textual image using the Hough transform. Mach. Vision Applied, 2: 141-153.

Sun, C. and D. Si, 1989. Skew and slant correction for document images using gradient direction. Proceeding of 4th International Conference on Document Analysis and Recognition, 18-20 August, 1997, Ulm, Germany, pp: 142-146.

Yamada, M., 1989. Image processing system. Canon, U.S. Patent 4,802,229.

Yu, B. and A. Jain, 1996. A robust and fast skew detection algorithm for generic documents. Pattern Recog., 29: 1599-1629.