Journal of Computer Science 5 (5): 363-368, 2009
ISSN 1549-3636

Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity

Atallah Mahmoud Al-Shatnawi and Khairuddin Omar
Department of System Science and Management, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, Selangor, Malaysia

Abstract: Problem statement: Skew detection and correction is the first step process in the document analysis and understanding processing steps. Correction the skewed scanned document image is very important, because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. The noises and the deviation in the document resolution or types are still the main two challenges facing the Arabic skew detection and correction methods. **Approach:** The proposed method work involved inscribing the text in the document by an arbitrary polygon and derivation of the baseline from polygon's centroid. **Results:** The proposed method was implemented on 150 different scanned Arabic documents, from different sources like journals, textbooks, newspapers and the like in addition to handwritten document, with different resolutions and different fonts and it was obtained an accuracy ratio of 87%. **Conclusion:** The proposed method was efficient, simple and fast, it was not affected by noise and it was proved their suitability to work with documents with different fonts and documents with different resolutions.

Key words: Arabic document, skew detection, skew correction, centre of gravity

INTRODUCTION

The goal of the character recognition systems is to transform the input data (pattern of data), such as text written document on manuscript, text typed on document or online writing into a digital format. This can be manipulated by word processing software[4,13]. Recognition can be done offline or online. In offline recognition, papers, manuscripts or documents are scanned or captured and finally are manipulated by OCR system. In online recognition application takes place during the writing process[2,7,9,19].

Arabic language is generally considered universal as its letters are the basis for various other languages like Urdu, Farsi, Jawi and many others languages[3]. Arabic character recognition system is considered quite complex as compared to Latin and Chinese because the text is written cursively and also the complexity of the alphabets representation in Arabic[19]. The Arabic OCR system goes through five stages: Image acquisition, Preprocessing, Segmentation, Feature Extraction and Classification (Recognition)[2,7,9,16]. These stages work together to improve OCR systems recognition ratio moreover to reduce the recognition time[13,14].

Arabic OCR preprocessing stage should contain smoothing, noise removal, image decomposition, skew detection and correction, edge detection and baseline detection, the document skew detection and correction is that research's focus.

The skew correction is considered mandatory in preprocessing Arabic OCR system stage, because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. As skew is generally introduced into the image while scanning and leaving it as it is without correction will give wrong results during document analysis and recognition[10]. Figure 1 shows one of the skewed document images.

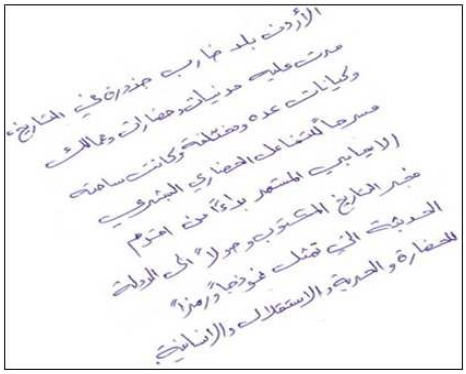

Fig. 1: Shows one of the skewed document images

Corresponding Author: Atallah Mahmoud Al-Shatnawi, Department of System Science and Management, Faculty of Information Science and Technology, University Kebangsaan Malaysia, Selangor, Malaysia

Both Hough transform, Cross Correlation, Projection Profile, Fourier transform and K Nearest Neighbor (K-NN) clustering were used for skew detection and correction the scanned images.

Hough transform[15] is also used for skew detection. The points in the Cartesian coordinate system are described as a summation of sinusoidal distribution:

$$p = x\cos\theta + y\sin\theta \quad (1)$$

The skew angle is calculated on the basis that at the skew angle the density of Transform spaces is maximum. After mapping (x, y) into (p, θ), the count of points where a sinusoidal curve intersects another sinusoidal curve with a different (p, θ) value increases the probability that a line determining the skew angle. A lot of research has been carried out on using Hough transform for determining skew angle. In[8], a selected area of the document is chosen for skew angle determination. The reason for choosing a particular area is to reduce the input data to process as the computational complexity in this process is quite high. A few other methods like hierarchical Hough transform[18] whose computational complexity is also high, another method with less computational complexity using hierarchical Hough transform[8,12] have already been proposed.

Cross correlation[17] is a computation intensive approach which is quite accurate. The lines of text in a document are considered as vertical lines which are spaced with a uniform distance of d between them. The skewed document vertical lines subtend an angle with the horizontal. So pixels in these vertical parallel lines are translated due to skewing. This translation concept is used for finding the correlation as:

$$R(x_0, s) = \sum_y L_1(x_0, y)\, L_2(x_0+d, y+s) \quad (2)$$

where, L1 and L2 are the parallel vertical lines.

The main drawback is that in real scanned document d is not constant and often needs to be backtracked.

Horizontal projection profile[6] is generally a histogram of the number of dark pixels in horizontal scan lines of a document. The troughs and Peaks are calculated as for a script with horizontal text lines the projection have peaks at text line positions and troughs at locations between successive text lines. The difference between the trough and peak is calculated at every angle and the maximum difference gives the skew angle. Method discussed in[1] is a unique method in which the document is partitioned into vertical strips and the Horizontal projection profile is applied to each strip individually which are later correlated to determine skew angle. This method is particularly good in determining small skew angles (less than 100).

The Fourier Transform method[10] works on the basic principle that skew angle is the one at which density of spectrum is largest for the document. However, the computational complexity is high.

Another computationally intensive method is the clustering method. In[5], the skew angle for all the connected words in the document is found out and a histogram for the determined skew angles is realized. The maximum clustered skew angle in histogram is the skew angle of the document. In another method[11], the centers of the nearest neighbors of the connected words in the document are vectorised and later correlated to determine the skew angle.

Revealing the importance of this stage the author of this study mainly concentrates on the Skew angle determination sub process of this stage and implements a novel algorithm in this research.

MATERIALS AND METHODS

The text image which is scanned at an angle can be rotated to a normal position following a series of steps. As a starting point an Arabic text image which is scanned in an abnormal direction is considered. The steps to be followed for getting back to the normal position include:

- Base line identification
- Skew angle correction

The baseline identification is generally the most important step of the whole process. Baseline is the Line along which the center of gravity of the word hangs. A novel Approach is used in this algorithm where in the whole word is inscribed in a polygon with at least two dimensions. The centre of gravity of the polygon is considered as a single line extending a certain angle with the horizontal. The angle is measured which gives the angle by which the word or document is rotated and also signifies the direction and angle by which is should be rotated for it to be a text in readable and normal form.

Centroid: The centroid is also known as the "centre of gravity" or 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. 2, a 6 side polygon is shown and the Center Of Gravity (COG) is calculated using the equation:

$$c_x = \frac{1}{6A}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$
$$c_y = \frac{1}{6A}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (3)$$

After calculating the COG point, a line is drawn joining the origin and the COG which is the required line of gravity.

Using a rectangle as the polygon is described here. The centroid for a rectangle is the point joining the diagonals which is considered as line of gravity for the polygon. Consider the document in Fig. 3 below which is skewed. Inscribe the text document in a rectangle by considering the farthest pixel in the four directions and determine the centroid. The line joining the centroid and the origin is the required baseline which gives the skew angle.

Although a rectangle is used as polygon for explanation in this case, the implemented algorithm uses an ellipse as polygon because of the simplification that the baseline for the ellipse is the major axis itself which reduces the computation.

The proposed algorithms steps:

Input: Text Image, like the samples showing in Fig. 4.

Output: Skew corrected text image, like the samples showing in Fig. 5.

Step 1: Determine the farthest points in all the four directions. Figure 6 shows the scanned image farthest points.

Step 2: Find the centroid using these four points, so the previous four points representing the polygon corners and the polygon center (COG), can be calculated by using the equations number three. Figure 7 shows the COG.

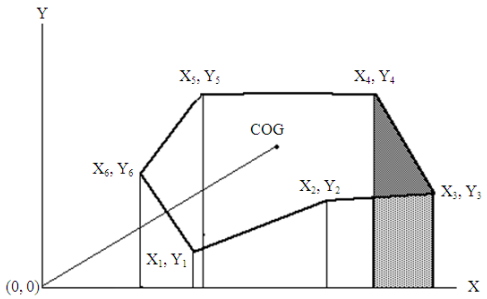

Fig. 2: Polygon

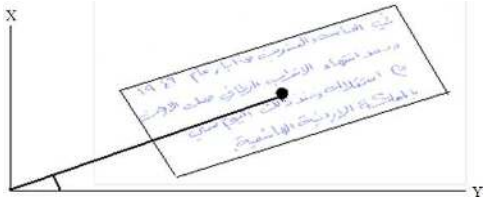

Fig. 3: Word inscribed in ellipse

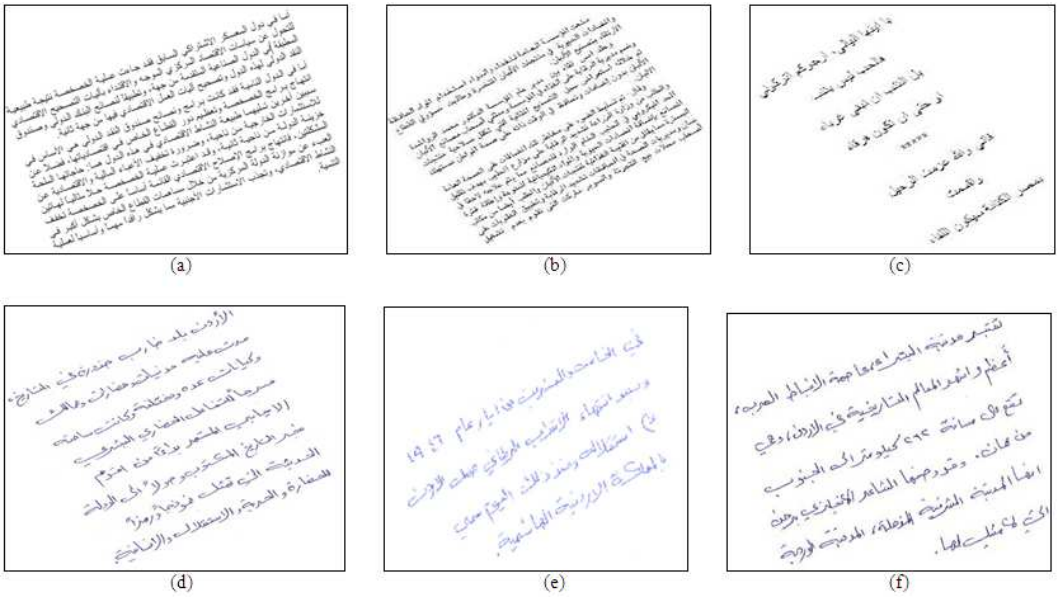

Fig. 4: Samples of the text input images: (a and c): Box pages; (b): Journal text; (d, e and f): Handwritten documents

Fig. 5: Corrected text image after applying the proposed method

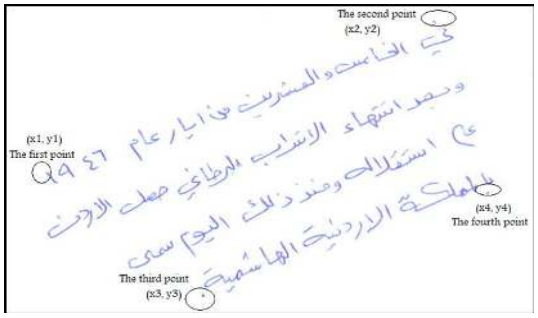

Fig. 6: The scanned images farthest points

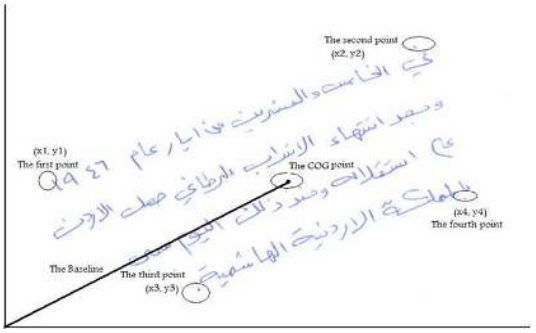

Fig. 7: The scanned images Baseline and the COG

Step 3: To get the baseline, Join the centroid to the origin, Fig. 7 shows the scanned image baseline and the COG too.

Step 4: Find the angle of the so formed baseline which gives the skew angle. Figure 8 shows the skewed angle detection.

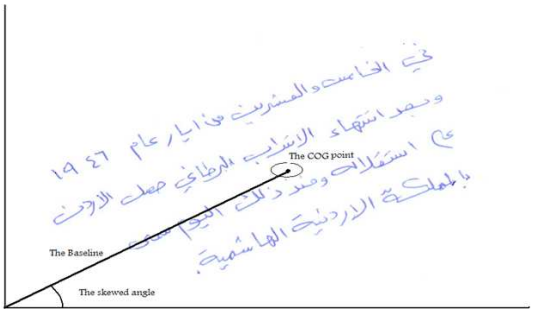

Fig. 8: The skewed angle detection

Step 5: Rotate the document in the reverse direction (clockwise direction) by the skew angle.

RESULTS

The method has been implemented in the math lab programming language on duo core 2.0 GHZ in 2009. We have considered different skewed documents from different sources like journals, textbooks, newspapers and the like in addition to handwriting document. For experimentation purpose 150 documents are considered, half of them are handwriting documents, samples of text documents are shown in Fig. 4. The correct skewed angle and the result skewed angle obtained by using the proposed method, in addition to the computing time taken by the proposed method for these documents are reported in Table 1. Figure 5 shows the corrected text image after applying the proposed method of images reported in Fig. 4.

Table 1: The correct skewed angle, proposed method skewed angle and the computing time

Image number	The correct angle (degrees)	The skewed angle detected by the proposed method	Computing time (sec)
a	19	19	0.3478
b	33	34	0.4822
c	42	41	0.3734
d	27	25	0.4826
e	20	19	0.3623
f	18	17	0.3501

DISCUSSION

In the skew detection and correction algorithms, the simplicity, generality and the applicability of the proposed method can be discussed, so the proposed method works with both the handwritten and the printed documents, by using the same procedures, while in other hand each of the Hough transform and Projection Profile can work with both types of documents, whether printed or handwritten, but the results remain poor unless it is supported by some of the conditions for working with each method on its own. The proposed method also is proved that working with the printed documents is better than the handwritten documents, because of the nature of the Arabic language handwritten and this can be solved by surrounding the text of the document with eight points, two points in each direction.

Also the proposed method can 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 Fig. 4 a and b, as well as handwritten documents, as shown in Fig. 4 d-f. And it is able to work with documents with different resolutions, as shown in Fig. 5. While on the other hand, other methods usually are designed to find a deviation skewed angle of a certain type of documents with resolutions constant. The noise does not affect the proposed method performances, while it has very high influences in them, basing on the Hough transform and Projection Profile.

CONCLUSION

An efficient, simple and fast a novel and accurate method to estimate skew angle is presented in this study. The proposed methods work based on Centre of Gravity. After surrounding the text of the document with four points which represent the four corners for the production of parallelogram. Proposed method has proved it can find the angle of deviation of the different kinds of printed documents such as magazines or books and other printed documents, as well as handwritten documents. Proposed method also proved their suitability to work with documents with noise and documents with different resolutions. Proposed method was implemented on 150 different documents and the rate of accuracy was 87% and it is proved that working with the printed documents is better than the handwritten documents.

ACKNOWLEDGMENT

This study was supported by a grant from University Kebangsaan Malaysia, Selangor, Malaysia. We also appreciate the excellent cooperation and support of Mohammad Nasrudin and Majdi Tahat.

REFERENCES

1. Akiyama, T. and N. Hagita, 1990. Automated entry system for printed documents. Patt. Recog., 23: 1141-1158. DOI: 10.1016/0031-3203(90)90112-X
2. Al-Badr, B. and S. Mahmoud, 1995. Survey and bibliography of Arabic optical text recognition. Signal Process., 41: 49-77. 1995. DOI: 10.1016/0165-1684(94)00090-M
3. AL-Shatnawi, A. and K. Omar, 2008. Methods of Arabic baseline detection-the state of art. Int. J. Comput. Sci. Network Secur., 8: 137-142. http://paper.ijcsns.org/07_book/200810/20081021.pdf
4. Argner, V. and h. El Abed, 2008. Databases and Competitions: Strategies to Improve Arabic Recognition Systems. pp: 82-103. DOI: 10.1007/978-3-540-78199-8
5. Hashizume, A., P.S. Yeh and A. Cosenfeld, 1986. A method of detecting the orientation of aligned components. Patt. Recog. Lett., 4: 125-132. http://cat.inist.fr/?aModele=afficheN&cpsidt=7989370
6. Hou, H.S., 1983. Digital Document Processing. Wisely New York, ISBN: 0471862479.
7. Khorsheed, M.S., 2002. Off-line Arabic character recognition-a review. Patt. Anal. Appl., 5: 31-45. DOI: 10.1007/s100440200004
8. Le, D.S., G.R. Thoma and H. Wechsler, 1994. Automatic page orientation and skew angle detection for binary document images. Patt. Recog., 27: 1325-1344. http://archive.nlm.nih.gov/pubs/doc_class/prword.php
9. Liana, M. and G. Venu, 2006. Offline Arabic handwriting recognition: A survey. IEEE Trans. Patt. Anal. Mach. Intell., 28: 712-724. DOI: 10.1109/TPAMI.2006.102

10. Omar, K., A. Ramli, R. Mahmod and M. Sulaiman, 2002. Skew detection and correction of jawi images using gradient direction. J. Technol., 37: 117-126. http://www.penerbit.utm.my/onlinejournal/37/D/JT37D13.pdf
11. O'Gorman, L., 1993. The document spectrum for page layout analysis. IEEE Trans. Patt. Anal. Mach. Intell., 11: 1162-1173. DOI: 10.1109/34.244677
12. Pal, U. and B.B. Chaudhuri, 1996. An improved document skew angle estimation technique. Patt. Recog. Lett., 17: 899-904. DOI: 10.1016/0167-8655(96)00042-6
13. Sarhan, A.M., and O.I. Al Helalat, 2007. Arabic character recognition using artificial neural networks and statistical analysis. Proc. World Acad. Sci. Eng. Technol., 21: 32-36. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.4838&rep=rep1&type=pdf
14. Safabakhsh, R. and P. Adibi, 2005. Nastaaligh handwritten word recognition using continuous-density variable-duration HMM. Arabian J. Sci. Eng., 30: 95-118. http://www.kfupm.edu.sa/publications/ajse/Articles/301B_07P.pdf
15. Srihari, S.N. and V. Govindaraju, 1989. Analysis of textual images using the Hough transform. Mach. Vis. Appl., 2: 141-153. DOI: 10.1007/BF01212455
16. Nawaz, S.N., M. Sarfraz, A. Zidouri and W.G. Al-Khatib, 2003. An approach to offline Arabic character recognition using neural networks. Proceeding of the 10th IEEE International Conference on Electronics, Circuits and Systems, Dec. 14-17, pp: 1328-1331. http://eprints.kfupm.edu.sa/3803/
17. Yan, H.,1993. Skew correction of document images using interline cross correlation. Comput. Vis. Graph. Image Process., 55: 538-543. DOI: 10.1006/cgip.1993.1041
18. Yu, B. and A.K. Jain, 1996. A robust and fast skew detection algorithm for generic documents. Patt. Recog., 29: 1599-1629. DOI: 10.1016/0031-3203(96)00020-9
19. Zeki, A.M., 2005. The segmentation problem on Arabic character recognition-the state of the art. Proceeding of the 1st International Conference on Information and Communication Technology, Aug. 27-28, IEEE Xplore Press, USA., pp: 11-26. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1598538