ISSN:0975-9646

A Novel Approach to Correction of a Skew at Document Level Using an Arabic Script

Amani Ali Ahmed Ali[#1], Suresha M[*2]

[#]Deptartment of MCA & Computer Science, Kuvempu University
Shankaraghatta, Shimoga, India
[*]Assistant Professor, Dept of MCA & Computer Science, Kuvempu University
Shankaraghatta, Shimoga, India

***Abstract*— The performance of an Arabic character system without accurate skew correction will not be satisfied for most of the scanned documents. This paper proposed a method for skew correction for Arabic document script. The proposed approach with high accuracy can detect skew. The experimental result shows that the proposed method is efficient compared to well-known existing methods.**

***Keywords*— Arabic Handwritten, skew detection, skew correction.**

I. INTRODUCTION

Offline Arabic character recognition is a very important technique for different fields of computer applications in recent years. The conversion of the paper based documents into electronic versions for storage, retrieval, automatic processing, and transmission. Online library and data searching become more common during the life. Documents and data imports become a huge problem that authors have to face. By manual data input takes too much time, therefore improving the optical Arabic character recognition rate and speed provides efficient works. Recognition of Arabic character is considered a hard problem because of the large categories, the complex structure, and the widely variable and many similar shapes of Arabic character.

Arabic language is generally considered universal as its letters are the basis for various other languages like Farsi, Urdu and many others languages. Also 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.

The optical Arabic character recognition system goes through five stages: Image acquisition, Preprocessing, Segmentation, Feature Extraction and Classification Recognition (Liana M. and Govindaraju V., 2006) and (Nawaz S.N., etc., 2003) these stages work together to improve the accuracy of Arabic character recognition systems and reduce the recognition time (Sarhan, A.M., and Al Helalat O.I., 2007).

Arabic character recognition 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 detection, correction and transformation are required in the stage of Arabic character recognition system preprocessing, because of the skew angle while writing or, scanning it takes the characters in input image skew. The skew characters will seriously effect on the accuracy and reliability of the segmentation, feature extraction stages and recognition results. The frequently used skew detection methods are KNN, Hough, Robust, MST cluster, and etc. (Cheng F. H., 1989) and (Touj S., etc., 2005). 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 (Omar, K., 2002).

In this paper some experiments are performed to detect and correct the skew character image. The technique is particularly useful for computing a global description of a feature, given local measurements. The technique for line detection is that each input measurement indicates its contribution to a globally consistent solution. In case of the parameter space is not than two-dimensional, this kind of transformation has a satisfied effect. Because of excellent characteristics of this technique, such as sensitive to local defects, random noise robust and parallel processing. This method is widely used in image processing, pattern recognition, and computer vision.

II. LITERATURE REVIEW

The skew-detection system can be divided into the sections shown in Figure 1. The processing starts with data acquisition and ends up with the skew-correction process.

Skew estimation and correction are necessary preprocessing steps of document layout analysis and optical character recognition approaches (Chaudhuri B.B. and Pal U., 1997). Skewness refers to the tilt in the bit mapped image of the scanned document for optical character Recognition. It is usually happened if the document is not well aligned on the scanner, thus yielding a skewed (rotated) digital image. Most of the optical character recognition algorithms are sensitive to the skew or orientation of the input document image making it necessary to develop algorithms to perform skew detection and correction automatically (Lehal G.S. and Dhir R., 1999).

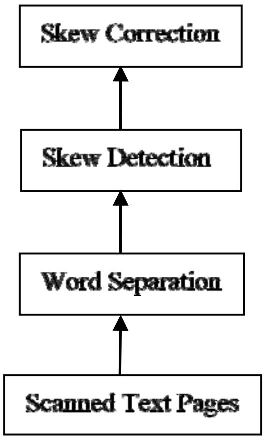


Fig. 1 Character recognition flow diagram.

Cross Correlation, Fourier transform, Projection Profile and K Nearest Neighbour (K-NN) clustering were used for skew detection and correction The cross correlation method (Cao Y., etc., 2003) is deepened on the correlation between two vertical lines in a scanned image. Correlation matrix can be produced since the pixels in the two parallel lines are translated due to skew. Cross correlation (Yan H., 1993) is a computation intensive approach which is quite accurate. In a document image the lines of text are considered as vertical lines those are spaced with a uniform distance between them. Pixels in these vertical parallel lines are translated due to skewing because the skewed document vertical lines subtend an angle with the horizontal. This translation concept is used for finding the correlation. The drawback is that in real scanned document distance d is not constant and often needs to be backtracked.

(Yan H., 1993) presented a method based on the cross-correlation between two lines in the document with a fixed distance. The correlation functions for all pairs of lines in the image are accumulated. To determining the skew angle the shift for which the accumulated cross-correlation function takes the maximum. The document is rotated in the opposite direction for skew angle.

Horizontal projection profile (Hou, H.S., 1983) in horizontal scan lines of a document image is a histogram of the number of dark pixels. For a script with horizontal text lines the troughs and Peaks are calculated the projection have troughs at locations between successive text and lines peaks at text line positions. Calculated at every angle and the maximum difference gives the skew angle is the difference between the trough and peak. In (Akiyama, T. and N. Hagita, 1990) the 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) and a unique method where the document is partitioned into Horizontal and vertical strips.

In (Omar, K., etc., 2002) the Fourier Transform method works on the fundamental principle where skew angle is the density of spectrum is largest for the document.

In (Chen Y,Wang J, 2000) proposed method based on Fourier transform. Where the direction for which the density of the Fourier space is the largest gives the skew angle. This requires the computation of the Fourier transform, which can be time consuming for a large document.

However, the computational complexity is high. Another computationally intensive method is the clustering method. In (Hashizume A., 1986) 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 (O'Gorman, L., 1993) the centers of the nearest neighbours are vectorised of the connected words in the document and later correlated to determine the skew angle.

In (Mahmoud A, etc., 2009) still the main two challenges facing the Arabic skew detection and correction the noises and the deviation in the document resolution or types. The proposed method work involved overwrite the text in the document by an arbitrary polygon and derivation of the baseline from polygon's centroid. The steps followed for getting back to the normal position include:

- Base line identification
- Skew angle correction

Generally, the most important step of the whole process is the baseline identification. Baseline is the line along which the center of gravity of the word hangs. In this algorithm a novel approach is used where in the whole word is overwrite in a polygon with at least two dimensions where a single line extending a certain angle with the horizontal was considered the center of gravity of the polygon. The angle is measured where gives the angle by which is should be rotated for it to be a text in readable and normal form and angle by which the word or document is rotated and also signifies the direction. The centroid is also known as the "center of gravity" or the "center of mass".

In (Touj S., etc., 2005) the top line algorithm does not operate directly on the skewed image. The skewed document first is converted to a segment file or a thin segment file, and then the algorithm operates on one of these files to find skew angle.

A series of projection profiles are obtained in projection profile method in a number of angles close to the expected orientation, and the differences is calculated for each of the profiles. The profile which gives maximum differences corresponds to the projection with the best alignment to the text line, this projection angle is called the skew angle (Durrani N., 2007).

III. METHODOLOGY

Close gaps and fill small holes, morphological closing operation is applied on the image as a preprocessing stage to prepare the image for subsequent stages. Proposed method is mainly applied on binary image (i.e. edge image). Therefore, gray image preprocessing should be done before, such as image filtering and edge detection. The result of preprocessing directly impacts the skew result. Gaussian noise and Impulse noise are two wide known common image noises. In this paper used a method depended on multiple median extractions binary filtering. This method covers pixel intensity similarity and spatial neighbourhood correlation. To protects the edge of the testing image that has Gaussian noise and Impulse noise the reference pixel value, which selected by pseudo-median filter is used. In contrast to past noise filter methods which aim to multiple median extractions binary filtering, one kind of noise, can process documents with mixed noise, and it had a good filtering result. Using the method of repeat binary filtering instead of Canny operator's Gaussian filtering process and adaptive filtering process, it avoids the fuzzy edge of the images which caused by filtering process, and gets a better edge detection result.

A corner can be considered as intersection between two contours, or is the sudden changes pixel in a gray value to detect and correct inclination.

$$E(u,v) \approx [u,v]SM\begin{bmatrix} u \\ v \end{bmatrix} \tag{1}$$

$$SM = \Sigma_{x,y}\, w(x,y)\begin{bmatrix} Ix^2(x,y) & Ix(x,y)Iy(x,y) \\ Ix(x,y)Iy(x,y) & Iy^2(x,y) \end{bmatrix} \tag{2}$$

E: The edge

SM: Symmetric matrix

Ix, Iy: derivatives in x and y directions respectively.

$w(x,y)$ the window function for smoothing, generally the Gaussian window is used.

$$C = \det(SM) - k(trace(M))^2 \tag{3}$$

$$\begin{aligned} \det(SM) &= \lambda 1 * \lambda 2 \\ &= Ix^2(x,y) * Iy^2(x,y) - Ix(x,y)Iy(x,y)^2 \end{aligned} \tag{4}$$

$$trace(SM) = \lambda 1 + \lambda 2 = Ix^2(x,y) + Iy^2(x,y) \tag{5}$$

The value of k between 0.04 and 0.06, The authors determine that the best result for k value in this methodology is 0.05. Take an image NXN as the example, the value of ρ is range from $\left[-N, \sqrt{2N}\right]$the value of θ is from $[0,\pi]$.

The θ space separated into M parameter sector, and thought a non-zero (x,y) possibly belongs to each θparameter sector of the document space, it may belongs to any straight line. In every θparameter sector, according the equation (6) to calculate the value of ρ, then falling into the corresponding ρ parameter sector, finally check each ρ parameter sector voting number(the length of the straight line), if it more than the setting threshold ,it will be considered to be a straight line.

$$\rho = x\cos\theta + y\sin\theta \tag{6}$$

Skew angle detection Select randomly two edge pixels and calculate the skew angle: The algorithm concrete realization. step is as follows: $\theta = arctg\left(\frac{y_k - y_i}{x_k - x_i}\right)$. Put the resulting angle into a vector defining the skew angle and an integer accumulator. Repeat the pixel selection and angle deduction process until the accumulator reaches a predefined threshold. The highest peak defines directly the skew angle.

The algorithm concrete realization step is as follows:

Step 1: Calculate and find corners by the above procedures.

Step 2: Ascending order searches in the image, until a non-zero-point ρ of the image was found, take this point as the seed point.

Step 3: Take as the original point the point ρ to select a m x m zone $A\rho$, searching the none-zero points in the $A\rho$ zone, then according to the equation (6), and calculate the straight parameter pair $(\rho_i, \theta_i)$of every non-zero point ρ_i and ρ .

Step 4: Make the deviation range ρ as $\Delta\rho$, the deviation range of θ as $\Delta\theta$, voting in $A\rho$, counting the number ni of the parameter pair which falling into every parameter interval $(\rho + \Delta\rho, \theta + \Delta\theta)$; Find out the parameter zone of maximum vote nmax from $A\rho$, and get the average value of (ρ_i, θ_i), to reduce the influence of quantitative error, and get $(\overline{\rho}, \overline{\theta})$, set it as the straight line parameter which cross the point ρ of $A\rho$.

Step 5: Set Ti as a length threshold of$A\rho$, if nmax >T1, then turn to step (7) and search others points that belonging to the line. Otherwise, the line which cross point ρ is not exist, and set gray value of ρ as zero, then turn to step (2), and restart searching.

Step 6: Set the search zone into entire image, according to equation (6) to calculate the value ρ of non-zero point P, which have been found, and set the value θ as $\overline{\theta}$. If $\left|\rho - \overline{\rho}\right| < \Delta\rho$, then ρ belongs this line, then cumulate the voting number, and set the gray value as zero.

Step 7: After entire document searching, if the voting number of the line is larger than preset threshold value T2, then the line is existing, get the parameter and put into set line(n), where n is the line number detected. Otherwise the line does not exist.

Step 8: Set the gray value of ρ as zero, return step (2), and restart searching, until there is no non-zero point in document.

Skew document correction is the transformation of coordinate. Set (x,y) as the dot coordinate ofρ , transform θ degree to get documentρ' the corresponded coordinate $(x,y)$is(x',y'). then Set the values θ positive to clockwise, and negative to counterclockwise. Then the coordinate can be represented as follows:

$$x' = \mathrm{x}\cos\theta \text{ - } \mathrm{y}\sin\theta \tag{7}$$

$$y' = y\cos\theta + x\sin\theta \tag{8}$$

The document image was pixels' form, which is separate variable. documents can be transformed through the following methods. To capture images or horizontal scanned while the direction of pixel steps is up-towards, the angle of ρ is $\tan^{-1}\left(\frac{\rho - 1}{RW}\right)$, RW is width which has selected; while the direction of pixel steps is down-towards, the angle of $-\tan^{-1}\left(\frac{\rho - 1}{RW}\right)$. To vertical scanned or captured images, while the direction of pixel steps is right-towards, the angle of ρ is $\tan^{-1}\left(\frac{\rho - 1}{RH}\right)$, RH is height which has selected; while the direction of pixel steps is left -towards, the angle of ρ is $-\tan^{-1}\left(\dfrac{\rho-1}{RW}\right)$. While skew correction, clockwise skew images are tokens. The lines are represented by sub-pixel step through specifying the skew angle; therefore, the skew correction can be executing during the corresponding offset of each pixel decrease or increase. The following is the horizontal and vertical skew correction functions.

$$\theta j = Iv\left(\frac{(\mathrm{p}\text{ - }1)j}{\mathrm{RW}}\right) \tag{9}$$

$$\theta i = Iv\left(\frac{(q-1)i}{RH}\right) \tag{10}$$

The RW is the width of rect, the function $Iv(\;)$is used to get the closest integer value. Skew angle is positive while step is up-toward, or negative while down-towards.

The RH stands is the height of rect, Skew angle is positive while step is right-toward, or negative while left-towards. Generally, the new image after transform is larger than the original image. The new height and width is:

$$NH = OH + O\theta_j \tag{11}$$

$$NW = OW + O\theta_i \tag{12}$$

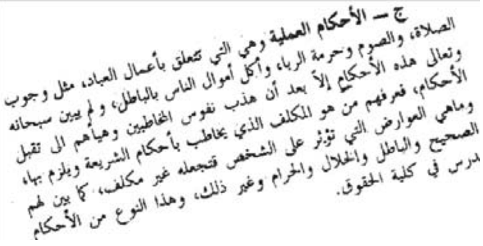
When the skew angle is positive, the original document image pixel of row j and column i corresponds row $(j-\theta_i+O\theta_i)$ of new document image and column $(i+\theta_j)$. When the skew angle is negative, the original document image pixel of row j and column i corresponds and row $(j+\theta_i)$ and column $(i-\theta_j+O\theta_j)$ of new document image.

IV. RESULTS AND DISCUSSION

The algorithm has been tested on a set of 150 different types of Arabic document images. Different handwritten text. The proposed method attempted on both handwritten and the printed documents. This method works for different types of documents with various image resolutions by several effects on processing speed of the system. A better document image analysis can improve the system accuracy and system could work more efficiently on both left and right skewed images with any of orientation of image The proposed method has been also tested on IFN/ENIT (Pechwitz, M., 2002) with different skewed angle, samples are shown in figure 2 and figure 3.

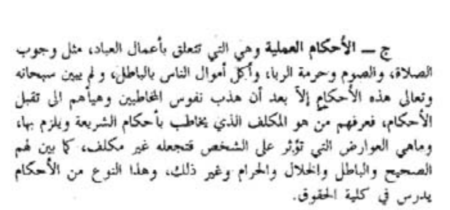
The experimental result shows that the proposed method is efficient compared to well-known existing methods. Table1 compares the experimental result by using our proposed method and the other experimental results. The experimental results show the efficacy is high compared to the result of well-known existing methods. The results show the proposed method has a good performance on noise image skew correction. and according to the experiment result, using the proposed method to correct the skew document are very efficient and exact. The proposed method has high noise immunity and adaptive.

Because the proposed method works with both the handwritten and the printed documents, by using the same procedures can be discussed in the skew detection and correction algorithms the applicability, and the generality of the proposed method can be satisfied, while each of the Hough transform and Projection Profile in other hand can work with both types of document images, whether handwritten or, printed but the results remain poor unless it is supported by some of the conditions for working with each method on its own.



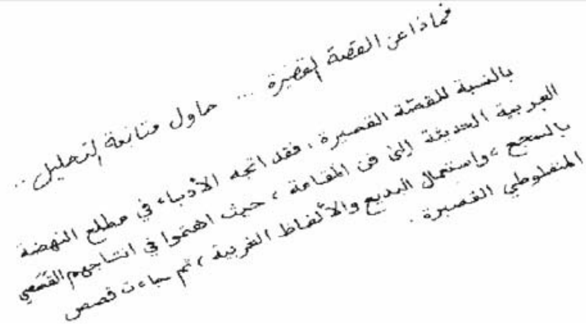
ج ـ الأحكام العملية وهي التي تتعلق بأعمال العباد، مثل وجوب
الصلاة، والصوم وحرمة الربا، وأكل أموال الناس بالباطل، و لم يبين سبحانه
وتعالى هذه الأحكام إلا بعد أن هذب نفوس المخاطبين وهيأهم الى تقبل
الأحكام، فعرفهم من هو المكلف الذي يخاطب بأحكام الشريعة ويلزم بها،
وماهي العوارض التي تؤثر على الشخص فتجعله غير مكلف، كما بين لهم
الصحيح والباطل والحلال والحرام وغير ذلك، وهذا النوع من الأحكام
يدرس في كلية الحقوق.

Fig. **2(a)** printed document image before processing.



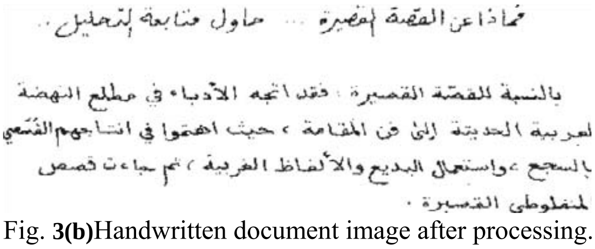
ج ـ الأحكام العملية وهي التي تتعلق بأعمال العباد، مثل وجوب
الصلاة، والصوم وحرمة الربا، وأكل أموال الناس بالباطل، و لم يبين سبحانه
وتعالى هذه الأحكام إلا بعد أن هذب نفوس المخاطبين وهيأهم الى تقبل
الأحكام، فعرفهم من هو المكلف الذي يخاطب بأحكام الشريعة ويلزم بها،
وماهي العوارض التي تؤثر على الشخص فتجعله غير مكلف، كما بين لهم
الصحيح والباطل والحلال والحرام وغير ذلك، وهذا النوع من الأحكام
يدرس في كلية الحقوق.

Fig. **2(b)** printed document image after processing



نماذا عن القصة القصيرة ... حاول متابعة لتحليل .
بالنسبة للقصة القصيرة ، فقد اتجه الأدباء في مطلع النهضة
العربية الحديثة إلى فن المقامة ، حيث اهتموا في انتاجهم القصصي
بالسجع ، واستعمال البديع والألفاظ الغريبة ، ثم جاءت قصص
المنفلوطي القصيرة .

Fig. **3(a)**Handwritten document image before processing.



نماذا عن القصة القصيرة ... حاول متابعة لتحليل .
بالنسبة للقصة القصيرة ، فقد اتجه الأدباء في مطلع النهضة
لعربية الحديثة إلى فن المقامة ، حيث اهتموا في انتاجهم القصصي
بالسجع ، واستعمال البديع والألفاظ الغريبة ، ثم جاءت قصص
لمنفلوطي القصيرة .

Fig. **3(b)**Handwritten document image after processing.

TABLE I.
COMPARISONS TO OTHER METHODS IN THE LITERATURE TESTED ON THE IFN/ENIT DATABASE.

Method	**accuracy (%)**
Skeleton-based (Pechwitz and Maergner, 2003)	88
PCA (Burrow, P., 2004)	82
Hough Projection (Pechwitz and Maergner, 2003)	83
Proposed method	98

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 document images, as shown in Fig. 2(a) and 2(b), as well as handwritten documents, as shown in Fig. 3(a) and 3(b). And it is able to work with documents with different resolutions. While other methods usually are designed to find a deviation skewed angle of a certain type of documents with resolutions constant on the other hand. In the proposed method performances, while it has very high influences in them the noise does not affect basing on the Hough transform and Projection Profile

V. Conclusion

In the Arabic skew detection and correction methods still facing the main challenges as the noises and the variation in the document image resolution. The proposed method works more accuracy on images of printed as well as handwritten documents. It was applied on different types of 150 documents and entire IFN/ENIT database is used a better document analysis can improve the system efficiency. The performance and efficiency of proposed method can be improved on noise free document images. So the proposed method presented a simple, fast and efficient document image skew detection.

To evaluate the proposed algorithm. The main process is to extract corner features points, and projected them in the proposed method to calculate the skew angle. The accuracy of 98% show the high level in performance of the proposed algorithm, which outperform the existing skew detection methods. Proposed method also proved their suitability to work with documents with noise and documents with different resolutions.

The main benefits and advantages of this algorithm are: The entire document image is reduced into a relevant lower edge image. For uniformly skewed document authors can just make a small part of document (approximately one or two lines) to detect the skew angle, which make the algorithm faster and more accurate. Besides, the advantage of the technique is that it can be applied to different applications for straight segment detection. The main drawback of this system is the short Arabic handwritten writing (IFN/ENIT database) in some cases where the lower edge points do not emphasize straight lines or the sub words are not strictly aligned.

References

[1] Akiyama, T. and N. Hagita, “Automated entry system for printed documents”, Pattern Recognition Vol 23, No 11, pp1141-1154, 1990.

[2] Burrow, P.,“Arabic Handwriting Recognition “, (M.Sc. thesis). University of Edinburgh, England, 2004.

[3] Cao Y., Wang S., Li H., “Skew Detection and correction in Documents Images Based on Straight-Line Fitting”, Pattern Recognition Letters, PRL, Vol. 24, no. 12, pp. 1871-1879. 2003.

[4] Chaudhuri B.B. and Pal U., “Skew Angle Detection of Digitized Indian Script Documents”, IEEE Transactions on pattern analysis and machine intelligence, Vol. 19, No. 2, pp. 182-186, 1997.

[5] Cheng F. H., “Recognition of Handwritten Chinese character by modified Hough Transform techniques”, Journal IEEE Transaction on Pattern Analysis and Machine intelligence, Vol. 11 Issue 4, April 1989.

[6] Chen Y,Wang J., “Skew Detection and Reconstruction Based on Maximization of Variance of Transition- Counts”, Pattern Recognition, 33(5) 195-208, 2000.

[7] Durrani N., “Typology of word and automatic word segmentation in Urdu Text Corpus”, Ph.D. Thesis, at National University of Computer and Merging Sciences Lahore, Pakistan, 2007.

[8] Hashizume A., Yeh P.S. and Rosenfeld A., “A method of detecting the orientation of aligned components”, Pattern Recognition, Letter 4: 125-132, 1986.

[9] Hou, H.S., “Digital Document Processing”, Wisely New York, ISBN: 0471862479, 1983.

[10] Kaur L., Jindal S., “Skew Detection Techniques for various Scripts”, International Journal of Scientific and Engineering Research (IJSER)", Vol. 2, No. 9, pp. 1-3, 2011.

[11] Lehal G.S. and Dhir R., “A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents”,Proceedings 5th International Conference of Document Analysis and Recognition,Banglore,pp.147-152, 1999.

[12] Liana M. and Govindaraju V., “Offline Arabic handwriting recognition: A survey”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 28, No. 5, pp. 712-724, 2006.

[13] Mahmoud A, Al-Shatnawi and Omar K, “Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity”, Journal of Computer Science 5(5), pp. 363-368 2009.

[14] Nawaz S.N., Sarfraz M., Zidouri A. and Al-Khatib W.G., “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, 2003.

[15] Omar, K., A. Ramli, Mahmod R. and Sulaiman M., “Skew detection and correction of jawi images using gradient direction”. Journal Technologi, Vol. 37, pp. 117-126, 2002.

[16] O’Gorman, L., “The document spectrum for page layout analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, pp. 1162-1173, 1993.

[17] Pechwitz, M., Snoussi S., Maergner V., Ellouze N., Amiri H., “IFN/ENIT – database of handwritten Arabic words”. In:Proceedings CIFED’02, pp. 129–136, 2002.

[18] Pechwitz M., Maergner V., “HMM-based approach for handwritten Arabic word recognition using the IFN/ENIT – database”.In: ICDAR. IEEE Computer Society, pp. 890–894, 2003.

[19] Sarhan, A.M., and Al Helalat O.I., “Arabic character recognition using artificial neural networks and statistical analysis”. International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol. 1, No. 3, pp. 506-510, 2007.

[20] Touj S., Ben Amara N., Amiri H. “Generalized Hough Transform for Arabic Printed Optical Character Recognition”, The international Arab Journal of Information Technology, Vol. 2, No. 4, pp. 326-333, October 2005.

[21] Yan H., “Skew correction of document images using interline cross correlation”, CVGIP Graphical Models and Image Processing, Vol. 55, No. 6 pp. 538-543, 1993.