Skew Detection and Correction of Mushaf Al-Quran Script using Hough Transform

Salem Saleh Bafjaish[1], Mohd Sanusi Azmi[2],
Mohammed Nasser Al-Mhiqani[3],
Amirul Ramzani Radzid[4]
Faculty of Information and Communications Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Hairulnizam Mahdin[5]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—Document skew detection and correction is mainly one of base preprocessing steps in the document analysis. Correction of the skewed scanned images is critical because it has a direct impact on image quality. In this paper, the authors proposed a method for skew detection and correction for Mushaf Al-Quran image pages based on Hough transform method. The technique uses Hough transform lines detection for calculating the skew angulation. It works for different version of Mushaf Al-Quran image pages which has skewed text zones. Moreover, it can detect and correct the skew angle in the range between 20 degrees. Experiment conducted on different Mushaf Al-Quran image pages shows the accuracy of the method.**

Keywords*—Skew detection; skew correction; Hough transform; preprocessing; binarization; image analysis

I. Introduction

Document Image processing is one of the fields that are rapidly growing faster in nowadays. It aims to convert paper-based documents to forms that are proper for storage. It can be defined as the method that is used to perform some operation on specified image such as (Digitization, Storage, compression, Re-printing) [1]. Besides that, there are different aspects that image processing could be the base such as, electronic engineering and computer science too. One of the problems in this field is that, the text in a document may be rotated when scanning which leads to produce a skewed text in the printed document as in Fig. 1.

Fig. 1. Al-Quran Surah Al-Fatiha with Skew Angle -8°.

As a result of that, the quality of the document is decreased and that will lead to multiple problems in analysis the image as well as reduce performance of optical character recognition (OCR) [18]. This paper focuses on skew detection and correction for Mushaf Al-Quran image pages. By comparison to other language scripts, skew detection and correction for Mushaf Al-Quran script is quite different as it has diacritical marks as well as the handwritten style is different too as compared to normal Arabic scrips. Hough transform is a simple feature extraction technique that is widely used in computer vision, image analysis and image processing as well. It can simply use to find lines in image by linear transform to detect straight lines [9].

II. Related Work

Many studies of skew correction are published but for different languages such as English, Urdu, Chinese. However, in the document, text can be written on serval text lines. A various methods are used for skew detection and correction based on different algorithm like, Projection profile, nearest neighbor clustering, Fourier transform, cross correlation and others. Skew can be defined as the angle that deviates from x-axis. Furthermore, accurate skew detection and correction helps other processes of OCR to be more successful. In [2] a novel method was proposed to recognize Arab / Jawi and roman digit by OCR. In [3] skew in documents can be classified into three class namely global skew, multiple skew and no-uniform text line skew. In [4] document analysis depends on preprocessing stage , the much better the image is preprocessed, a much better result of analysis the image is. Furthermore, it increases the quality and the accuracies in the OCR systems. In [5] skew detection and correction can be the first step in the process of the document analysis as well as understanding processing steps as it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. Currently, a lot research in Arabic documents bus less work is intensively been explored for Mushaf Al-Quran. Initially method to estimate the skew angle in a paper as in Fig. 2 is to draw a line through the text characters, and then the angle of the drawn line with the horizontal edges of the original paper is the skew angle.

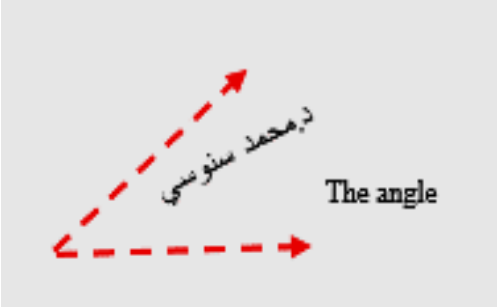


Fig. 2. Basic Skew Angle.

Generally, all ordinary pages have the skew angle of zero. However, the skew angle occurs due to different reasons. The main purpose of skew detection and correction is required to improve the quality of the scanned documents. In [6] O'Gorman paper, all these techniques can be categories into three groups as projection profile, Hough transform and nearest neighbor clustering. In [3] an evaluation for the most frequently skew detection techniques cited in their paper as (i) Projection Profile Analysis (PP), (ii) Hough Transform (HT) and (iii) Nearest Neighbor(NN). A comparison between the three techniques, the comparison started the weakness and strengths of each method as well as to compare the performance for both of them in term of the speed and the accuracy. Their evaluation showed that nearest neighbor techniques is the fasted one among them according to the speed but in other hand, its accuracy estimation evaluation is poor comparing to the other techniques. Furthermore, project profile technique gave the best estimation for the angle when it comes to the accuracy, in opposite its time is the longest to be executed. In [7] an efficiency discussion of two techniques Principal Component analysis (PCA) and Hough transform is presented to overcome problems that spoils the scanned documents. In [8] projection profile method is proposed for skew detection for handwritten signature, they used horizontal projection for detecting the skew angle and correct it using rotation transformation. In [9] a method proposed for detecting the skew and correct it for the handwritten Devanagari script using the technique Hough transform. The proposed method is to detect the skew and correct it at the word level as Devanagari script as in Fig. 3 is a little difficult comparing to other scripts because of the style of the writing as well as the writing style differs from one person to the other one [9].

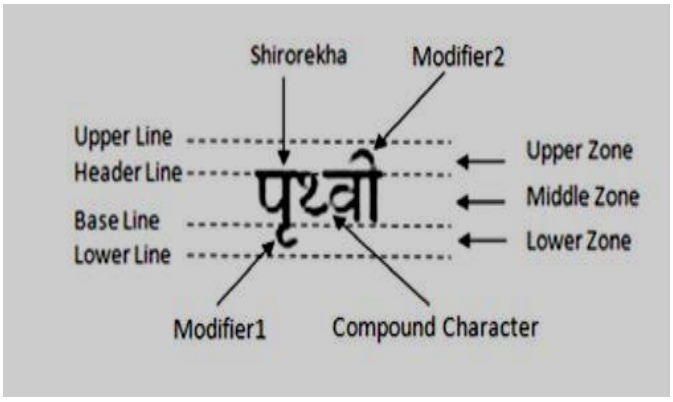


Fig. 3. Devanagari Simple Character [9].

The proposed method consists of preprocessing stage followed by word extraction stage is made in the image in order to extract the words, lastly Hough transform algorithm is applied in order to detect the skew of the word. In [4] proposed a novel skew detection and correction approach for scanned documents contains of two stages, first find the angles of the lines in the image with the respect of x-axis and second find the exact skew angle from the angles that are extracted from lines in the first stage. In [11] a proposed a simple and fast algorithm that determine the skew angle of the image as well as the slant angle of the text characters using the gradient orientation histogram. Additionally, the angle can be obtained using searching for a peak in the image histogram, the image can be corrected by a rotation at such an angle. In [1] proposed a new technique that detect the skew and correct it for the Arabic printed scripts based on connected component analysis and pixel projection. Moreover, the proposed technique take the advantage of the sharp writing line property for Arabic language that is obtained from histogram projection of the image for skew detection. In [12] an image moments are used for skew detection and correction. An image moment is the calculation of the weighted average (Moment) of the pixels 'intensities of the image. So, moments are employed to find the primary axis of every object in the document instead of applying the Hough transform. Finally, by using a feature that depends on the size of the object, the weighted average angle is estimated. In [13] skew detection method that uses run-length and Hough transform algorithm is presented. The proposed method reduce the amount of data in the image through using black horizontal and vertical run-lengths histograms which also reduces computational calculation of Hough transform and increase the speed of skew detection.

III. Mushaf Al-Quran Scripts Challenges

Mushaf Al-Quran is the holy book millions of Muslims around the world. It can be in two versions digital or printed form, although is in Arabic, but the way it is written is different from any Arabic/Jawi based document as it has "diacritics". In [14] a proposed method for identifying types of Arabic calligraphy in Malay accent script that is written in Jawi. Fig. 4 illustrates most of Arabic script challenges as in [15], [16] have presented Arabic scripts challenges as the following:

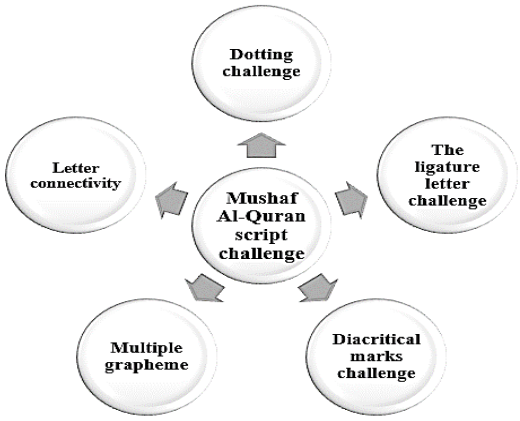


Fig. 4. Arabic Script Challenges.

1) The connectivity challenge: Arabic text can be only scripted cursively, that means all graphemes are connected together, this happens whether the text was handwritten or font written as in Fig. 5.

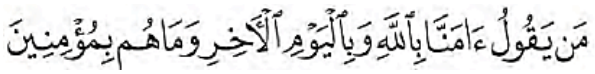


Fig. 5. Al-Quran Surah Al-Baqarah 8.

2) The dotting challenge: Dots in Arabic scripts are used to differ between the characters sharing similar graphemes. Accordingly, if a dot is missed with the process of skew detection, then that will affect the meaning of the text. Fig. 6 illustrates the dotting challenge.

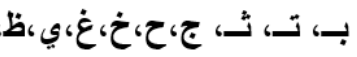


Fig. 6. Dotting Letters.

3) The multiple grapheme cases challenge: In Arabic orthography it's very due to have the connectivity in text which means that same letter can be different in the way how it's written based on the position of it in the Arabic word. Fig. 7 illustrates the letter ع with different writing styles.

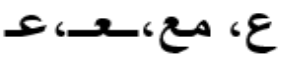


Fig. 7. Multiple Grapheme.

4) The ligatures challenge: Character in Arabic script can be compounded together at certain positions of the Arabic word. Ligatures can be found at almost all the Arabic fonts. Fig. 8 illustrates ligatures challenge.

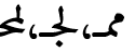


Fig. 8. Ligature Letters.

5) The diacritics challenge: The usage of diacritical marks helps to resolve linguistic ambiguity of the text. [14] However, in some case they goes vertical while the main text is going straight on line (horizontal) from right to left. Therefore, that makes some confusion for skew detection step in OCR. Fig. 9 illustrates a segment of Mushaf Al-Quran text with diacritics marks.

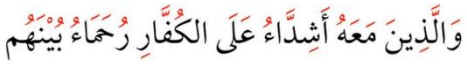


Fig. 9. Al-Quran Surah Al-Hujuratt 29.

IV. Proposed Method

The proposed methodology for Mushaf Al-Quran skew detection and correction is described here. The proposed method consists of six stages namely as convert to grayscale image, binary image, foreground image, Hough transform method to detect lines, calculate skew angle and finally rotate image as in Fig. 10.

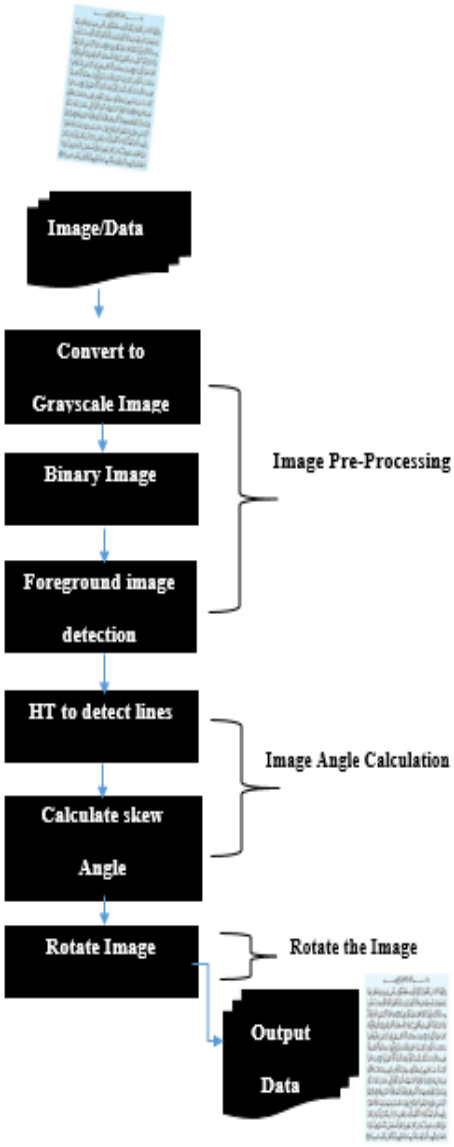


Fig. 10. Proposed Method.

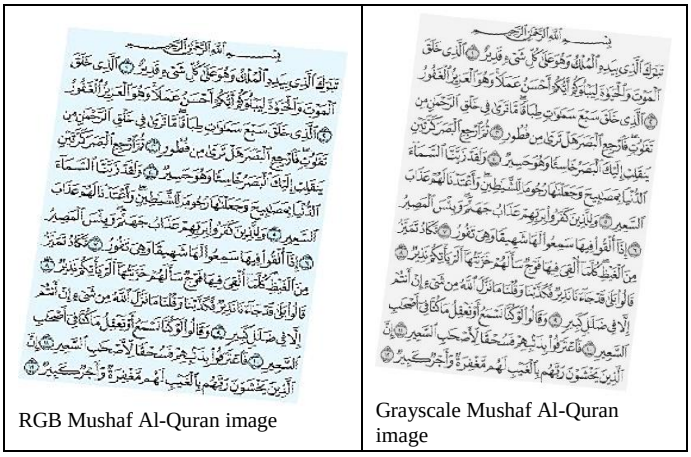
A. Grayscale Image

As in Fig. 11 some of Mushaf Al-Quran pages comes with different colors, so there is a need to re therefore there is a need for the conversion to grayscale image to get high performance of skew detection. There are some reasons for converting color Mushaf Al-Quran images to grayscale images as the following:

1) Reduce color: in color images, sometimes information of the images doesn't help to identify the important areas on the images and other features such as lines on the images.

2) Grayscale (8bits) images makes it easy for implementing binary algorithms because there are only two shades of colors in grayscale images which are white and black whereas color images has blue- green-red.

3) Algorithms applied to grayscale images are much faster than the once applied to RGB images.

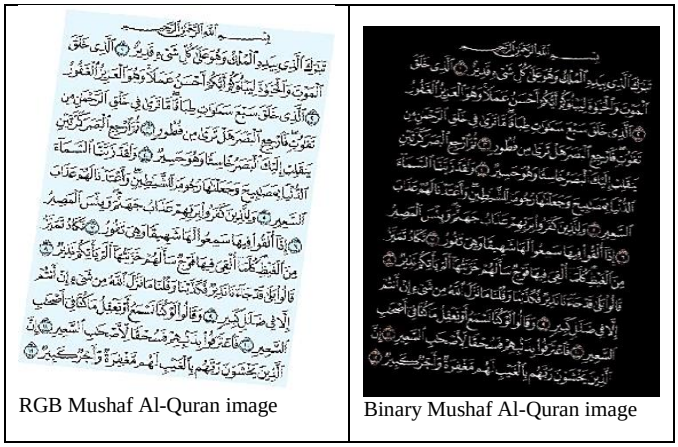

RGB Mushaf Al-Quran image

Grayscale Mushaf Al-Quran image

Fig. 11. Al-Quran Surah AlFajr (Different between color and grayscale image of Mushaf Al-Quran).

B. Binarization

In [17] an amendment has been made by applying Otsu's method for to improve noise and prepare images for the new proposed extraction feature method. In [10] also Otsu's method is applied for Arabic characters dynamically in order to choose the discriminant threshold on the image. Therefore in this paper Otsu's method is used too. Once an image is formed in grayscale form, next preprocessing step is applied on the image is the Binarization. Binary images are the images that have only values for each pixel, the two possible values are black and white. However, in this step a binary image is created from the original image to help to detect only the important areas and parts of Mushaf Al-Quran images. Fig. 12 illustrates the difference between normal image and binary image.

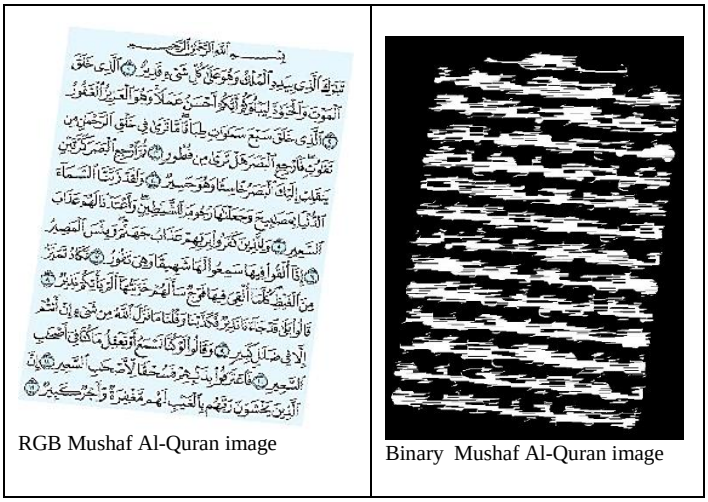

RGB Mushaf Al-Quran image

Binary Mushaf Al-Quran image

Fig. 12. Al-Quran Surah AlFajr (Different between color and binary image of Mushaf Al-Quran).

C. Foreground Mushaf Al-Quran Image Detection

Once a binary image is created, a morphology is applied to detect areas that have text in Mushaf Al –Quran images and then convert gotten text to lines using this morphology in the direction of x (close morphology is used here). This morphology produces image contains lines which will be used in the next stage for the angle calculation. Fig. 13 illustrates the difference between color image and foreground image.


RGB Mushaf Al-Quran image

Binary Mushaf Al-Quran image

Fig. 13. Al-Quran Surah AlFajr (Color and Foreground Image Detection).

D. Angle Calculation

This is the most important stage where skew angle is calculated. Text in previous stage is converted to connected lines, so line detection comes second. The connected words in the previous stage can be considered as straight lines which helps to apply Hough transform method for line detection. To make it clear, this stage can be achieved in by two important steps as the following

1) Line detection: using (1) helps to detect straight lines in the images using the equation.

$$r = xcon\ \theta + y\ sin\theta \quad (1)$$

Hough transform is one of the most used feature extraction technique in computer vision, image analysis and digital image. It was introduced by Paul Hough 1962. So based on the Fig. 14 for each point (X_0, Y_0) there are other set of points which can create lines as (X_1, Y_1), this set of points that create line can be defined with equation (1).The two values (r, θ) in the above formula represents the lines (connected words in Mushaf Al Quran images that gotten from the previous stage) that goes within (X_0, Y_0). In other words, the line in the image space represented as a point in the parameter space as in Fig. 15.

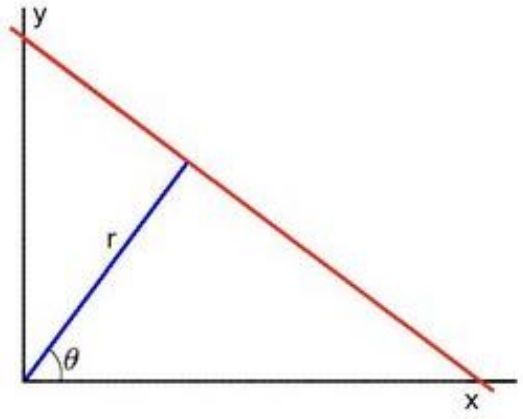


Fig. 14. Hough Transform Space.

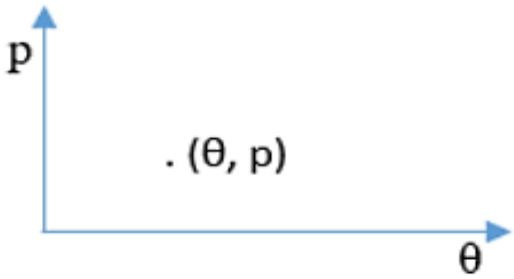


Fig. 15. Hough Transform Space.

Likewise, for linear Hough transform, two dimensional array are used for detecting lines in the image space in which each line is represented with two values of (θ, p) respectively. Further, straight lines are represented with peak strong point in the accumulator array in the image space as in Fig. 15 above. Once all peak points are detected, then it's easy to find line segments by end points of the peak values. The more intersections in the image space leads to the longest line among lines and that is the required line to calculate the skew angle.

2) Angle Calculation: skew angle is calculated using the longest straight line. In other words, it can be calculated with the deviation of the line with horizontal axis.

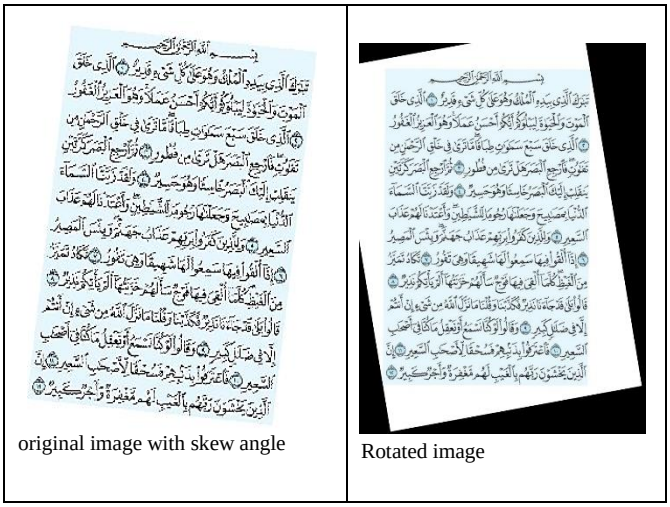
3) Rotate image: Once the skew angle is detected, it becomes easy for rotate it in order to correct the skew. However, several methods are used for skew correction like (direct method, indirect method contour oriented projection based and others), rotation of the image is done through Affine Transformation (2).

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (2)$$

Where (x, y) are the coordinates of the skew detected line, (θ) is the angle detected by Hough transform method. The above equation is for couter-clockwise.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(-\theta) & -\sin(\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3)$$

Equation (2) in which the rotate the calculated skew angle to horizontal angle. The line is rotated with θ angle, if the detected angle is positive, the angle is corrected to the negative angle with the same value and the vice versa. Fig. 16 illustrates the last stage of the proposed method "Rotate image".



original image with skew angle

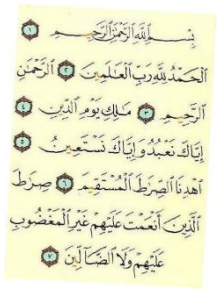
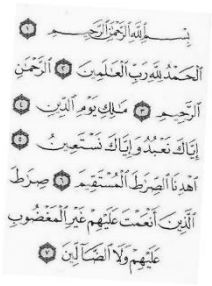
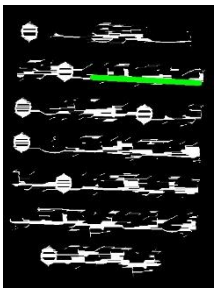
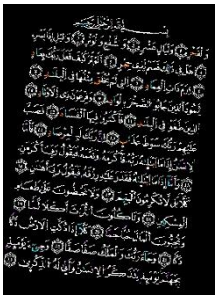
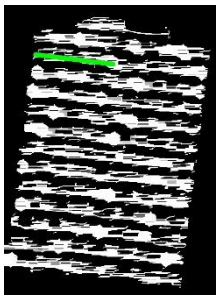
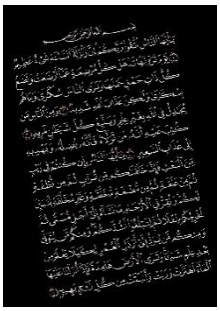
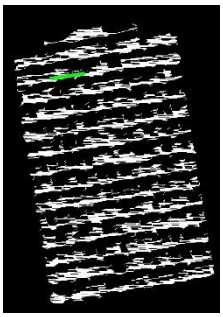
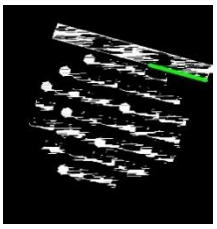
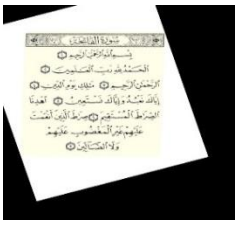
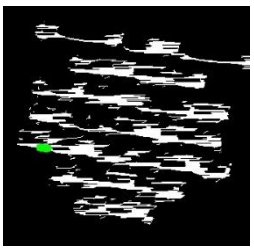
Rotated image

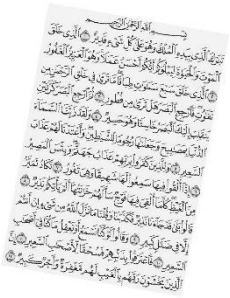
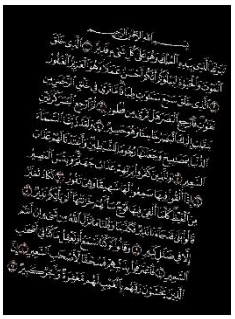
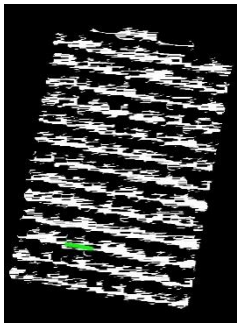
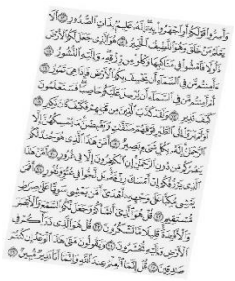
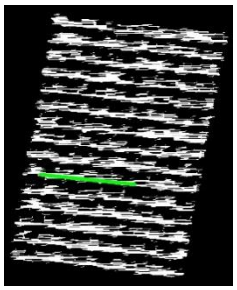
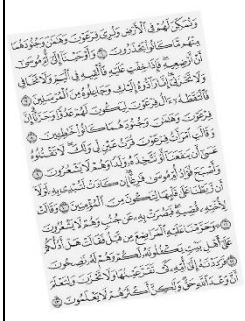
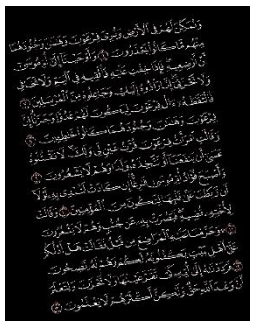
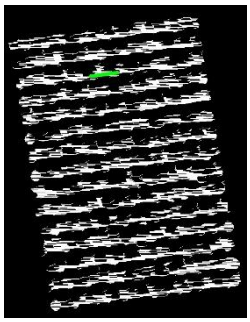
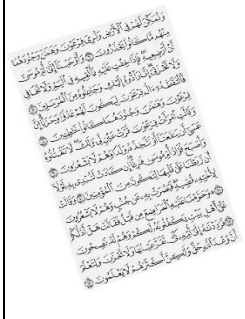
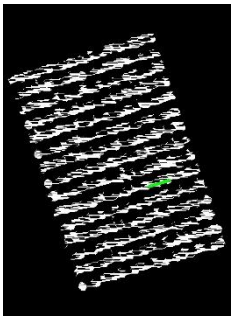
Fig. 16. Image Rotation.

V. Experimental Results and Discussion

In this paper, a proposed method was tasted on Mushaf Al-Quran images that have skewed text. In addition, 50 Mushaf Al-Quran images were tasted by our proposed method. The proposed method has been implemented using Java programming language, Test environment used a PC with Intel i5- 2430M CPU @2.40 GHz with 12GB of memory, also Opencv function was implemented with Java code for detecting skew lines in the image as well as for measuring skew angle. In addition, the documents image of Mushaf Al-Quran were self-obtained from the source (https://www.nourelquran.com/quranforall/fahd/index.php [19] and they were manipulated with different skew angle using software ImageJ. The accuracy for skew correction was about 90% for the images been tasted. Mostly, The Mushaf Al-Quran images that are colorful or have high resolution have lower accuracy in skew correction conversely with the images that have lower resolution in which the proposed method works perfectly. Therefore, Mushaf Al-Quran images have to be pre-processed before applying Hough transform method on, as in Fig. 11, converting input image to grayscale image helps to increase skew detection and process. Binary image also is a good way to increase skew detection as it was explained in proposed method section. The proposed method detecting and correcting skew angles through six stages namely, convert to grayscale, binary image, foreground image detection, HT transform method, calculate skew angle, rotate image. Table I shows a sample Mushaf Al-Quran images before and with skew correction at different angles.

TABLE I. SAMPLE RESULTS FOR SKEW DETECTION AND CORRECTION USING THE PROPOSED METHOD

Deg°	Original image	Grayscale image	Binary image	Line detection	Straighten image
3					
7					
-9					
17					
6					

10					
9					
-8					
-17					

VI. Conclusion and Future Work

In this paper, a methodology for Mushaf Al-Quran skew detection and correction was presented. Moreover, the proposed method was based on Hough Transform algorithm which simply used by different handwritten script skew correction. This method is tasted on handwritten Mushaf Al-Quran images that have skew text in. Furthermore, the proposed method consists of six stages are combined together to deliver the final result. To conclude, this method can be improved for further research to be more accuracy. A possible future work is to enhance the proposed technique with respect to processing time as well as to skew angle estimation.

Acknowledgment

The authors thank the Ministry of Education for funding this study through the following grants: FRGS/1/2017/ICT02/FTMK-CACT/F00345. Gratitude is also due to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

[1] I. Ahmad, 'A technique for skew detection of printed arabic documents', Proc. - 10th Int. Conf. Comput. Graph. Imaging, Vis. CGIV 2013, pp. 62–67, 2013.

[2] M. S. Azmi, K. Omar, M. F. Nasrudin, B. Idrus, and K. Wan Mohd Ghazali, 'Digit recognition for Arabic/Jawi and Roman using features from triangle geometry', AIP Conf. Proc., vol. 1522, pp. 526–537, 2013.

[3] A. Al-Khatatneh, S. A. Pitchay, and M. Al-Qudah, 'A Review of Skew Detection Techniques for Document', Proc. - UKSim-AMSS 17th Int. Conf. Comput. Model. Simulation, UKSim 2015, pp. 316–321, 2016.

[4] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, T. Breuel, and A. Dengel, 'A Novel Skew Detection and Correction Approach for Scanned Documents', DAS. IAPR Int. Work. Doc. Anal. Syst. (DAS-12), April 11-14, Santorini, Greece, no. 4, pp. 1–2, 2016.

[5] A. M. Al-Shatnawi, 'A skew detection and correction technique for Arabic script text-line based on subwords bounding', 2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014, pp. 324–328, 2014.

[6] L. O'Gorman, 'The Document Spectrum for Page Layout Analysis', IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993.

[7] R. N. Verma and L. G. Malik, 'Review of illumination and skew correction techniques for scanned documents', Procedia Comput. Sci., vol. 45, no. C, pp. 322–327, 2015.

[8] L. B. Mahanta and A. Deka, 'Skew and Slant Angles of Handwritten Signature', pp. 2030–2034, 2013.

[9] T. A. Jundale and R. S. Hegadi, 'Skew Detection and Correction of Devanagari Script Using Hough Transform', Procedia Comput. Sci., vol. 45, pp. 305–311, 2015.

[10] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, 'Arabic calligraphy classification using triangle model for Digital Jawi Paleography analysis', Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011, pp. 704–708, 2011.

[11] C. Sun and D. Si, 'Skew and slant correction for document images using gradient direction', Proc. Fourth Int. Conf. Doc. Anal. Recognit., vol. 1, pp. 142–146, 1997.

[12] G. Kapogiannopoulos and N. Kalouptsidis, 'A fast high precision algorithm for the estimation of skew angle using moments', Proceeding of IASTED, 2002.

[13] S. C. Hinds, J. L. Fisher, and D. P. D. Amato, 'A DOCUMENT SKEW DETECTION METHOD USING RUN-LENGTH ENCODING AND THE HOUGH TRANSFORM', pp. 464–468, 1990.

[14] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, 'Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks', Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011, no. July, pp. 704–708, 2011.

[15] M. Attia, 'Arabic Orthography vs . Arabic OCR', vol. 1, 2000.

[16] A. A. Aburas and M. E. Gumah, 'Arabic Handwriting Recognition : Challenges and Solutions Electrical and Computer Engineering Dept International Islamic University Malaysia Department of Information Technology , University Technology PETRONAS 2 . Pervious related Research Work', 2008.

[17] M. S. Azmi, K. Omar, M. F. Nasrudin, and A. K. Muda, 'Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis', vol. 5, pp. 696–703, 2013.

[18] F. Kleber, M. Diem, and R. Sablatnig, "Robust Skew Estimation of Handwritten and Printed Documents Based on Grayvalue Images." In Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 3020-3025. IEEE, 2014.

[19] Nourelquran.com. (2018). مجمع ذ سخة | الا ملك سورة | الا قران ذور موقع الا شريف الا مصحف لا ط باعة فهد الا ملك. [online] Available at: https://www.nourelquran.com/quranforall/fahd/sora/67.html [Accessed 20 Aug. 2018].