

## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

Computer Science Department - University of Technology

Received 12 June 2015 ; Accepted 29 February 2016

### Abstract

Image preprocessing has assumed an essential part in Arabic Handwriting Recognition System (AHRs). It has several stages which affect the accuracy of the recognition system. An efficient preprocessing framework for Arabic handwriting recognition system has been proposed in this paper. Since preprocessing stage in AHRs is imperative to reduce the dimensionality of image to remove the undesirable information then increase the processing speed of the AHRs. Besides, it provide correction process for the dataset elements to make it work probably through several processes that will discuss in this paper. Exploratory results with artificial and real life images shows that, the proposed method gives an efficient results that help to get a high recognition rate in the AHRs.

**Keywords:** Preprocessing, AHRs, Arabic text, Thresholding, Thinning, Normalization.

تجهيز إطار عمل كفوء لنظام التعرف على الكتابة اليدوية للغة العربية

علياء كريم عبد الحسن      مصطفى سلام كاظم

قسم علوم الحاسبات - الجامعة التكنولوجية

### الخلاصة

مرحلة تجهيز الصورة لها دور اساسي في نظام التعرف على النص المكتوب يدويا للغة العربية (AHRs). حيث لديها العديد من المراحل التي تؤثر على دقة نظام التعرف. تم اقتراح إطار تجهيزها كفاءة في اللغة العربية نظام التعرف على الكتابة اليدوية في هذا البحث. حيث ان مرحلة التجهيز AHRs أمر لا بد منه للحد من بعدية الصورة لإزالة المعلومات غير المرغوب فيها ثم زيادة سرعة تجهيز AHRs. الى جانب ذلك، فإنه يوفر عملية تصحيح للعناصر مجموعة البيانات والعمل على انجاحه وربما من خلال العديد من العمليات التي ستناقش في هذا البحث. نتائج استكشافية مع صور الحياة الاصطناعية وحقيقية تبين أن الطريقة المقترحة يعطي نتائج فعالة من شأنها أن تساعد على الحصول على معدل تعرف عالي في AHRs. **كلمات مفتاحية:** تجهيز ، AHRs ، النص العربي ، مستوى العتبة ، ترقيق ، التسوية

### Introduction

Recognition of handwriting has various practical applications in the areas such as postal address reading for mail sorting purposes, cheque recognition and word spotting on a handwritten text page, etc. For recognize handwritten words or characters there are several strategies in the computational pattern recognition such as artificial neural networks and statistical approaches like K-Nearest Neighbor KNN. Naturally, handwriting is cursive due to several factors which are the writer's style, quality of paper and geometric factors controlled by the writing condition its very unsteady in shape and quality of tracing. Preprocessing is the first step in Handwriting Recognition systems it is helpful to reduce the variability of handwriting by correct these factors and it will help to enhance the accuracy of segmentation and recognition methods[1].The most important step in handwriting recognition system is preprocessing based on the assumption that extracting and distinguish the objects from the background based on the gray levels. Besides, removing the image noise and the black space in the background is the next step in the preprocessing. The next step in preprocessing is to correct the rotation of the text in image to make a stability for the image which provide clear result for the recognition process epically for the feature extraction step [2].Furthermore, the image normalization reduces the image size to prepare the last edition of the image and send it to the next step of the recognition process [2].

### Related Works

Several researchers have considered the use of such preprocessing for AHRS. Subhash and Neeta [3] proposed new adaptive binarization approach for handwritten document to convert it from grayscale into binary based on the intensity of the pixel value. Furthermore, in [4] the researchers proposed a new algorithm for skew detection and correction based on the vertical and horizontal pixels. In [5] researchers propose a method for image binarization [6]. The algorithm is based on the fact that a document image includes very few pixels of useful information (foreground) compared to the size of the image (foreground + background). It is given the amount of black pixels in relation with all the pixels of the image concerning the cleaned version of the mentioned image in black and white. While in [7] three main stages are used for image thinning which are conditional contour selection, pixel removing, and one pixel width stage. These stage normalize the image by removing every two neighbor's pixels and make it as a single pixel. Most of the researches in handwriting recognition field face big issue in character segmentation which affect the recognition accuracy and make it low because of the differences in the handwriting styles of the writers. Therefore, working with the word handwriting level is more efficient to avoid the critical issue of the segmentation. In this work represent a proposed framework for image preprocessing stage based on several methods, for Arabic Handwriting Recognition System AHRS.

### Characteristics of the Arabic Writing

Arabic language is a widely used language as more than one billion people use Arabic in either their daily activities or religion-related activities [8]. Arabic language is cursive and written from right to left. It has 28 basic letters and eight diacritics [8]. Each Arabic letter can has different shape according to its position in the word. Moreover, there are different fonts that make Arabic character shape changed dramatically [9]. The table below shows the 28 letters and their various forms. Each letter has multiple forms depending on its position in the word. Each letter is drawn in an isolated form when it is written alone, and is drawn in up to three other forms when it is written connected to other letters in the word. For example, the letter Ain

**An Efficient Preprocessing Framework for Arabic Handwriting Recognition System**

**Alia Karim Abdul Hassan      Mustafa S. Kadhm**

has four forms: Isolated form (ع) and Initial, Medial, and Final forms (ع ع ع), respectively from right to left.

Furthermore, letters Hamza, Teh, and Alef have other forms, as shown in the table1 within a word, every letter can connect from the right with the previous letter. However, there are six letters that do not connect from the left with the next letter [8].

**Table 1: Arabic Letters and Their Forms**

Character	Isolate	Initial	Medial	Final	Character	Isolate	Initial	Medial	Final
Alif	ا	ا	ا	ا	Dhad	ض	ض	ض	ض
Ba	ب	ب	ب	ب	Taa	ط	ط	ط	ط
Ta	ت	ت	ت	ت	Dha	ظ	ظ	ظ	ظ
Tha	ث	ث	ث	ث	Ain	ع	ع	ع	ع
Jeem	ج	ج	ج	ج	Ghain	غ	غ	غ	غ
Ha	ح	ح	ح	ح	Fa	ف	ف	ف	ف
Kha	خ	خ	خ	خ	Qaf	ق	ق	ق	ق
Dal	د	د	د	د	Kaf	ك	ك	ك	ك
Thal	ذ	ذ	ذ	ذ	Lam	ل	ل	ل	ل
Rai	ر	ر	ر	ر	Meem	م	م	م	م
Zai	ز	ز	ز	ز	Noon	ن	ن	ن	ن
Seen	س	س	س	س	Ha	ه	ه	ه	ه
Sheen	ش	ش	ش	ش	Waw	و	و	و	و
Sad	ص	ص	ص	ص	Ya	ي	ي	ي	ي

**The Proposed Framework**

The proposed framework involves several steps which are image binarization, noise removal, remove image background, skew detection and correction then image thinning and normalization. Figure (1) illustrates the main steps of the proposed framework.

An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan      Mustafa S. Kadhm

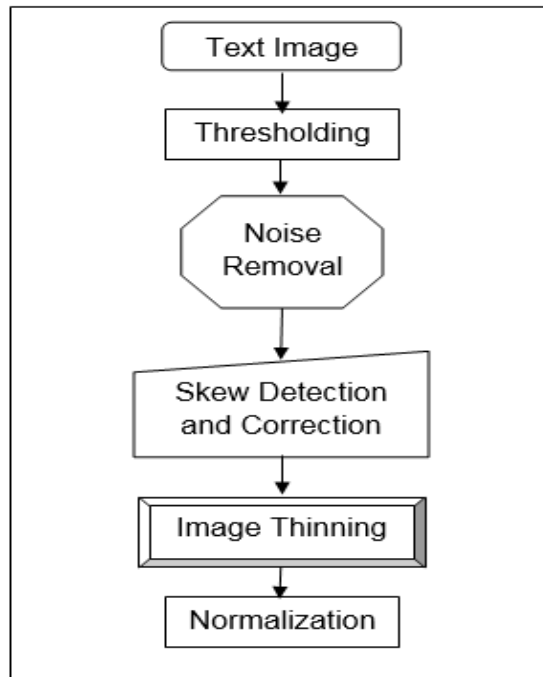


Figure1: The proposed Framework

1. Database

Here an Arabic word image database has been proposed. The dataset collected from different people with different ages and education background. All the participants received white papers and write down the most Arabic words. Figure2 shows the collection of our dataset samples

الشيخ	للفضل	طلب	السيادة
مع	يرجى	المير	الكلمة
السيد	اليها	ولغايت	ولحدة

Figure2: Arabic handwriting words database

## 2. Image Thresholding and Noise Removal

The input to the AHRS is a RGB text image which has the Arabic word. Before any processing could take place, it was then necessary to convert RGB images into grayscale representations of the handwriting. The first step in the proposed method is the image thresholding which convert the input image from grayscale into binary to reduce the image dimensionality which reduce the processing time. A thresholding method based on Fuzzy C-Means clustering (FCM) that proposed in [10] has been used to convert the input grayscale image into binary in this paper. After that, median filter has been used to remove undesired information from the binary image as shown in figure (3).

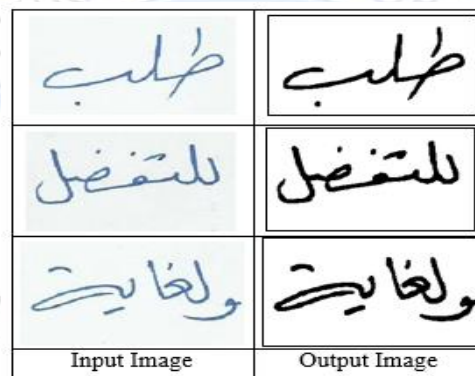


Figure3: Image Thresholding

## 3. Remove the White Space

The second step of the propose method is removing the unwanted white space in the image background. The white space representing by (0) value which can affect the feature extraction result and make it not efficient. The proposed approach for the removing the black space is based on using the BoundingBox tool in Matlab. First, the number of (0) values are calculated from the all image borders until the text as shown in figure (4).

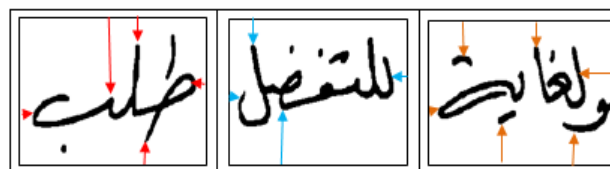


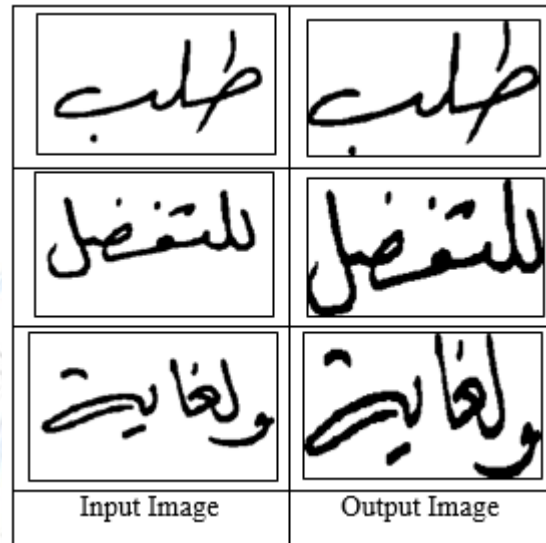
Figure4: Calculating of (0) values

## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

After calculating the number of (0) values from the borders into the text, the minimum distance of the results are save it for the all borders directions. The second step is drawing a BoundingBox around the saved distances points and crop the image from these points. The result of the remove white space method is illustrates in figure (5).



**Figure5: the output of remove the white space method**

#### 4. Skew Detection and Correction

Skew/rotation is a big challenge that face the handwriting recognition. The handwriting dataset has groups of same text images for different writers. Each of these images has different style and rotation depends on the writer style. Therefore, there is a need for a method to fix this problem and make all the dataset images be in the same skew. To start with detecting and correcting the image skew, several steps has been take place.

##### Algorithm of skew detection and correction

Step1: Load Image.

Step2: Finding the Horizontal projection profile of the input image.

Step3: Finding the histogram of the horizontal projection profile.

Step4: Display the histogram of horizontal projection.

An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan      Mustafa S. Kadhm

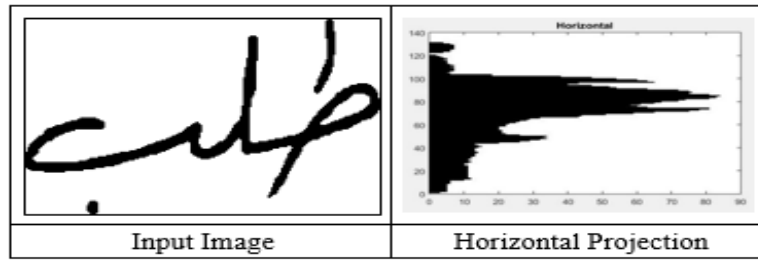


Figure6: Horizontal projection Profile

The main purpose of the skew detection is to determine the skew and the baseline of the image. The baseline is a vertical reference position for the characters and subwords in a handwritten text line image. The baseline is represent the line of the notebook that the writers used for writing the text. An example of the baseline can be shown in figure7.

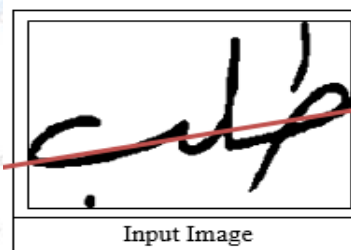


Figure7: Baseline of text image

The image skew can be determine through calculating the maximum peak in the histogram of the horizontal and vertical projection profile. However, to correct the image skew the image is rotated by positive and negative angle then after each rotation the maximum peak is calculated and compare it with the other peaks until get the higher one to make the baseline of the text image in the correct direction as shown in figure(8).

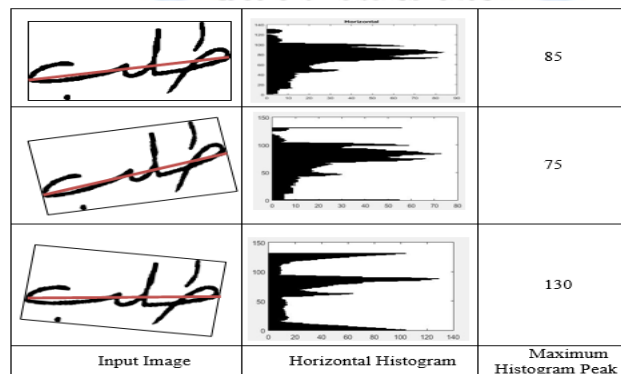


Figure8: Skew correction method



## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

According to the results in figure (8), the maximum peak of the histogram after rotate the image in positive and negative direction is 130. Furthermore, all the images in the dataset will have the same skew after applying the above method.

### 5. Image Thinning

The process to reduce image size to compact size and find the medial axis which defines as a set of pixels  $S$  where these pixels have an equal distance from the boundary pixels around it, and the output of this process is skeleton for the handwritten word, this process must save the geometry and the connections between the characters and the location of original character [11, 12], based on border pixels removing recursively taking into account saving the geometry, location and connections.

Skeleton representation advantages:

- Good way to represent the structural relations between components in the pattern.
- Wide-range used [12] for character, word, signature and handprint recognition systems.

Figure (9) illustrate the results of applying skeleton image method that proposed in [12].

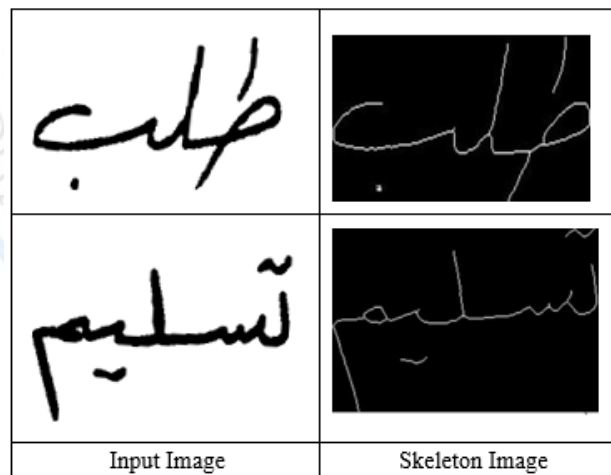
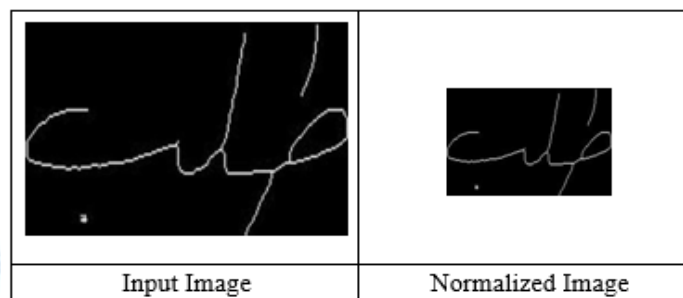


Figure9: Image Thinning

## 6. Size Normalization

The proposed Arabic dataset has various image sizes. It is important to make all the images in the dataset in the same size and make the recognition process faster. According to [13] the 128\*128 is the best size for recognition. All the dataset normalize into size 128\*128 an example in figure (10) for this normalization.

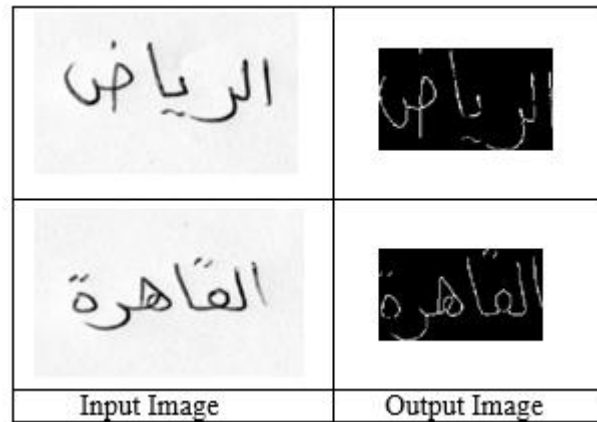


**Figure10: Image normalization**

After the normalization step the images are ready for the next step in recognition system which are feature extraction then classification.

## Experimental Evaluation and Discussion

The proposed framework is implemented using Matlab R2015a version, under windows7 64-bit Operating System, with RAM 6GB, CPU 2.50GHz core i5 and it achieved fast and effective results. Besides, the proposed framework has been applied for several images with different size and it gives a good results. However, the proposed framework has applied for standard database which is IESK-arDB[14] .The IESK-arDB is an off-line handwritten database. It contains 280 pages of a 14th century historical manuscripts, more than 4000 handwritten word images, and 6000 segmented character images. The word database vocabulary covers most of Arabic part of speech nouns, verbs, country/city names, security terms, and words used for writing bank amounts. The images from IESK-arDB has been tested also as shown in figure 11.



**Figure11: Image before and after applying the proposed framework**

The main benefits of the proposed framework is the representing of the text images with less number of foreground and background pixels. The number of the foreground and background image pixels can be obtained by applying the following algorithm:

#### **Algorithm to count the image pixels**

Step1: Load Image.

Step2: Scan the image from left to right and gave a label to each object using (BoundingBox).

Step3: Measure properties of image regions.

Step4: Initialize a counter=0.

Step5: Start a loop (in case there are more than one object in the image).

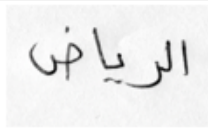

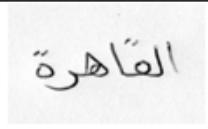

Step6: crop the mask according to (BoundingBox).

Step7: Counting the foreground and background pixels.

Step8: end of the loop

Step9: Print the result (number of the image pixels).

After applying the code1 on the images in figure(10) the results shows that, the number of pixel after implementing the proposed framework less than the pixels in the original images and in the same time it keep representing the text in images without any loose of the desired pixels as shown in figure(12).

Images before and after preprocessing	Number of foreground pixels	Number of background pixels
	16324	13984
	417	15967
	19677	18433
	430	15980

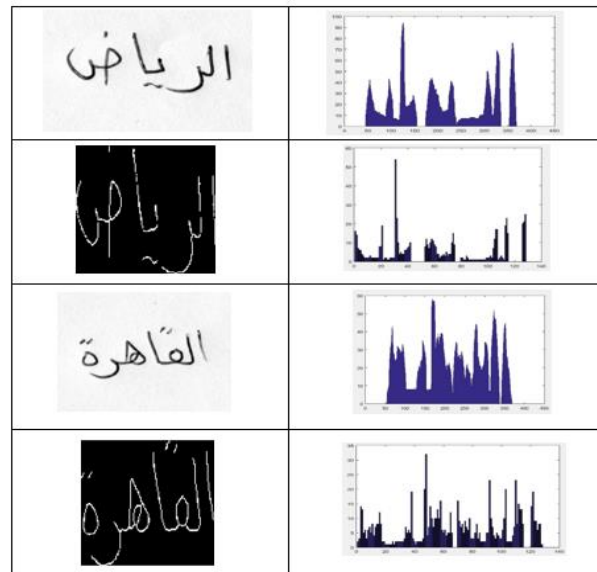
**Figure12: Number of foreground and background**

After test several images by the proposed framework, the average of the reduced pixels has been obtained. The medium range of the number of foreground pixels of the input images is 17200 and the number of background pixels is 14000. However, the medium range of the number of foreground pixels of the output images is 400 and the number of background pixels is 400. Therefore, the average of the reduced pixels in the foreground is 43% and the average of the reduced pixels in the background is 35%. Moreover, after obtaining the histograms of the input images after before any process then after applying the proposed framework, the histograms shows that the output images still have the same representation of the input images with less pixels. Besides, there are no missing important pixels which may distort the output images. Figure(13) illustrate the histograms of the input and output images.

An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm



**Figure13: Images histograms before and after applying the proposed framework**

Moreover, another major that can apply to check effected of the proposed framework is the entropy. Image entropy is a quantity which is used to describe the 'business' of an image. An image that is perfectly flat will have an entropy of zero. On the other hand, high entropy images such as an image of heavily cratered areas on the moon have a great deal of contrast from one pixel to the next. Figure14 shows the image entropy , before and after applying the proposed framework.

	4.726
	0.568
	4.492
	0.753
Image	Entropy

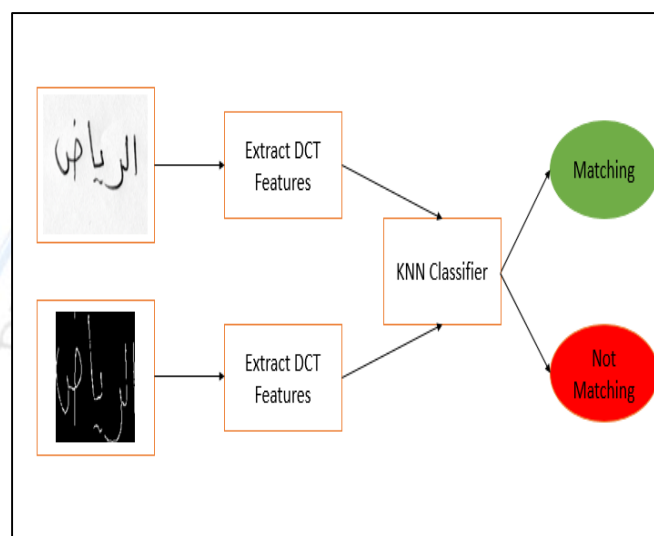
**Figure14: Images entropy**

## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

The last major used to evaluate the framework is the image matching technique. Discrete Cosine Transform (DCT) [15] has been applied for the input and output images of the frame work to extraction the features of the images. Each image will have a 120 feature vector to represent the extracted features. Besides that, K Nearest Neighbor (KNN) [16] is applied for matching between the images. The matching accuracy between the input and output images was 100%. Figure15 illustrate the matching technique.



**Figure15: Images Matching**

Most of the existing work using a common thresholding methods [17]. By applying the existing methods and the proposed method, it shows that the number of noise and misclassified pixels in our proposed method are less than the other methods. The output image of the propose method is more clear, readable as shown in figure16 and will be important for the next stage of handwriting recognition system, since the unclear image with noise will affect the accuracy result of the recognition process[18].

An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan      Mustafa S. Kadhm



**Figure16: Comparative outputs of various thresholding methods and the proposed method**

Furthermore, we have implemented the proposed method in MATLAB on a set of 500 handwritten text Images of IESK-arDB database. The proposed method could determine the exact amount of skew in each image efficiently. We have also implemented the linear regression and bounding box methods, and Hough transform. The comparative analysis is summarized in Table 2.

**Table 2: Results Comparison**

Compare Methods	Skew angle	Accuracy (%)
<b>Bounding Box Method</b>	0 – 25 Degrees	78.40
<b>Linear Regression</b>	0 – 360 Degrees	85.20
<b>Hough Transformation</b>	0 – 180 Degrees	98.30
<b>Proposed Method</b>	0 – 360 Degrees	98.55

The results of the proposed framework make the feature extraction method work simple and fast, because there is no need for complex process to obtain the feature vector since there are few pixels to represent each image. This concept lead to make the recognition process fast and efficient in terms of memory processing and recognition accuracy.

**Conclusion**

An efficient framework for image preprocessing stage in the Arabic Handwriting Recognition System (AHRS) has been proposed in this paper. Several methods discussed start from image binarization, noise removal, remove image background, skew detection and correction then image thinning and normalization. Experiments, an efficient results has been obtained from the

## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

proposed framework which has very good affect for the recognition processing speed and accuracy of AHRS.

### References

1. Kadhm, Mustafa S., and Asst Prof Dr Alia Karim Abdul. "Handwriting Word Recognition Based on SVM Classifier." *International Journal of Advanced Computer Science & Applications* 1.6: 64-68.
2. Asst.Prof.Dr. Alia Karim Abdul Hassan, Mustafa S. Kadhm. "Handwriting Word Recognition Based on Neural Networks." *International Journal of Applied Engineering Research (IJAER)* 10.22 (2015): 43120-43124.
3. Subhash Panwar, Neeta Nain."A Simple and Novel Adaptive Binarization Approach for Handwritten Documents". Malaviya National Institute of Technology, Jaipur 2013
4. M.L.M Karunanayaka, C.A Marasinghe, N.D Kodikara. "Thresholding, Noise Reduction and Skew correction of Sinhala Handwritten Words". University of Colombo School of Computing No.35, Reid Avenue, Colombo07,Sri Lanka MVA2005 IAPR
5. Anis Mezghani, Slim Kanoun,Souhir Bouaziz, Maher Khemakhem and Haikal El Abed. "Baseline Estimation in Arabic Handwritten Text-Line Evaluation on AHTID/MW Database". University of Sfax, National School of Engineers (ENIS), Sfax, Tunisia 2013
6. Ergina Kavallieratou. "A Binarization Algorithm specialized on Document Images and Photos". Dept. of Information and Communication Systems Engineering University of the Aegean.2005
7. Abu-Ain, W., Abdullah, S.N.H.S., Bataineh, B., Abu-Ain, T. & Omar, K., "Skeletonization Algorithm for Binary Images", in *International Conference on Electrical Engineering and Informatics (ICEEI 2013)*, UKM, Bangi, Selangor, Malaysia. pp. 690-694, 2013.
8. Gheith A. A.; K. S. Younis and M. Z. Khedher (2008) "Handwritten Arabic Character Recognition Using Multiple Classifiers Based On Letter Form",*Computer Engineering Department, University of Jordan, Amman 11942, Jordan, in Proc. 5<sup>th</sup> IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications, Innsbruck, Austria.*
9. Mostafa M. G. (2004) "An Adaptive Algorithm for the Automatic Segmentation of Printed Arabic Text ",*Computer Science Department, Faculty of Computer Science, King Abdul Aziz University, Al-Madinah Al-Munawwarah, P.O. Box 344, Saudi Arabia.*



## An Efficient Preprocessing Framework for Arabic Handwriting Recognition System

Alia Karim Abdul Hassan

Mustafa S. Kadhm

10. Dr. Alia Karim Abdul Hassan, Mustafa Salam Kadhm. "An Efficient Image Thresholding Method for Arabic Handwriting Recognition System." Engineering and Technology Journal 34.1 (2016): 26-34.
11. PETROS A. MARAGOS, RONALD W. SCHAFFER, "Morphological Skeleton Representation and Coding of Binary Images," IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-34, NO. 5, pp. 1228–1244, October 1986.
12. Lam, L., Lee, S., and Suan, C.Y., "Thinning Methodologies – A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(9): pp. 869–885, September 1992.
13. Malik Waqas Sagheer. "Novel Word Recognition and Word Spotting Systems for Offline Urdu Handwriting". Computer Science and Software Engineering Concordia University Montreal, Quebec, Canada.2010
14. Elzobi, Moftah, Ayoub Al-Hamadi, Zaher Al Aghbari, and Dinges, Laslo (2012) <http://www.iesk-ardb.ovgu.de/>
15. SA Khayam. "The Discrete Cosine Transform. (DCT): Theory and Application". Tech. Report. Michigan State University March 10th, 2003.
16. R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed, Wiley Interscience, 2001.
17. Liana M.: "Offline Arabic Handwriting Recognition: A survey". IEEE transaction .Pattern analysis and machine intelligent (2011)
18. Mustafa S.Kadhm, Asst. Prof. Dr. Alia Karim Abdul Hassan. "ACRS: Arabic Character Recognition System Based on Multi Features Extraction Methods." International Journal of Scientific and Engineering Research 6.10 (2015): 665-661.