An Efficient Image Thresholding Method for Arabic Handwriting Recognition System

Dr. Alia Karim Abdul Hassan
Computer Science Department, University of Technology/ Baghdad
Email:hassanalia2000@yahoo.com
Mustafa Salam Kadhm
Computer Science Department, University of Technology/ Baghdad.
Email: must.salam@yahoo.com

Received on:26/5/2015 & Accepted on:12/11/2015

ABSTRACT

Image preprocessing has assumed an essential part of handwriting recognition system. The main primary stage of the image preprocessing is thresholding. An effective thresholding method is based on Fuzzy C-Means clustering (FCM) for Arabic Handwriting Recognition system (AHR) has been proposed in this paper. Since thresholding stage in AHR is imperative to reduce the dimensionality of image to remove the undesirable information (noise) then increase the processing speed of the AHR system. The algorithm is performing by feeding the intensity of the pixel value of the image pixels into the FCM clustering algorithm. Exploratory results with artificial and real life images show that the proposed method gives better accuracy and good efficiency than the current methods.

Keywords: Thresholding, Preprocessing, FCM clustering, Arabic text.

طريقة صورة مستوى العتبة كفوءة لنظام التعرف على الكتابة اليدوية للغة العربية

الخلاصة

عملية تجهيز الصورة تعتبر جزء اساسي في نظام التعرف على الكتابة اليدوية. المرحلة الاساسية في عملية تجهيز الصورة تعتبر طريقة مستوى العتبة.في ورقة البحث هذه ، نقدم عملية العتبة فعالة على أساس Fuzzy C-Means clustering (FCM) خوارزمية لنظام التعرف على الكتابة اليدوية العربية Arabic Handwriting Recognition (AHW). وذلك ان مرحلة العتبة في AHW مهمة جدا لتقليل أبعاد الصورة لإزالة البيانات غير المرغوب فيها لزيادة سرعة المعالجة للمراحل المقبلة. وضعت خوارزمية من خلال دمج نتائج العتبة العالمية في خوارزمية التجميع FCM. وتم مقارنة أداءالخوارزمية المقترحة مع خوارزميات العتبة الموجودة. النتائج التجريبية مع الصور الاصطناعية و الحقيقية تشير إلى الخوارزمية المقترحة هي اكثر كفاءة وفاعلية.

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 correcting these factors and it will help to enhance the accuracy of segmentation and recognition methods. Thresholding or binarization and noise removal are the crucial landmarks of the preprocessing in handwriting systems.

The most important step in handwriting recognition system is Thresholding based on the assumption that extracting and distinguish the objects from the background based on the gray levels. By applying the thresholding method, the output will be a binary image which represents the foreground by level 0 (black) and 1 for (white) representing the background.Generally, two main types of threshold selection which are local and global. The global methods segment the entire image by using histogram of the gray level with a single threshold. However, in the local methods the threshold indicates for not to the whole image, but for sub images that divided from the original one [1].

Besides, Noise described in the image processing research as any undesirable information containing digital image. Noise appear by converts an optical image into continuous electronical signal which is the essential source of introducing the noises into the electronic images and noises are introduced into the image of the transmission process. Several works take in account removing the noise by transform filters. Hadeel and Ali[2] frame let transform to remove the noise from images , also Jabir and Arshed [3] used complex discrete wavelet transform for removing undesired data from image but both methods are complex in term of time consuming.

However, the proposed method indicates the threshold based on the intensity of the pixels in the image the clustering technique to get best threshold results. Furthermore, median filter has been used in the proposed method to remove the noise in faster way. In this work a proposed method for thresholding based on Fuzzy C-mean Clustering FCM, for Arabic Handwriting Recognition Applications.

Related Work

There are numerous methods for image thresholding which already used by some researchers. The most common thresholding method has been proposed by Otsu [4]. Otsu's method works better where clear separation between foreground and background exists or where image illumination is not variable. However, real life images possess especially in handwriting images various kinds of degradations (e.g. illumination contrast, skewed, stains, and noise) that weaken thresholding proposed by Otsu's.

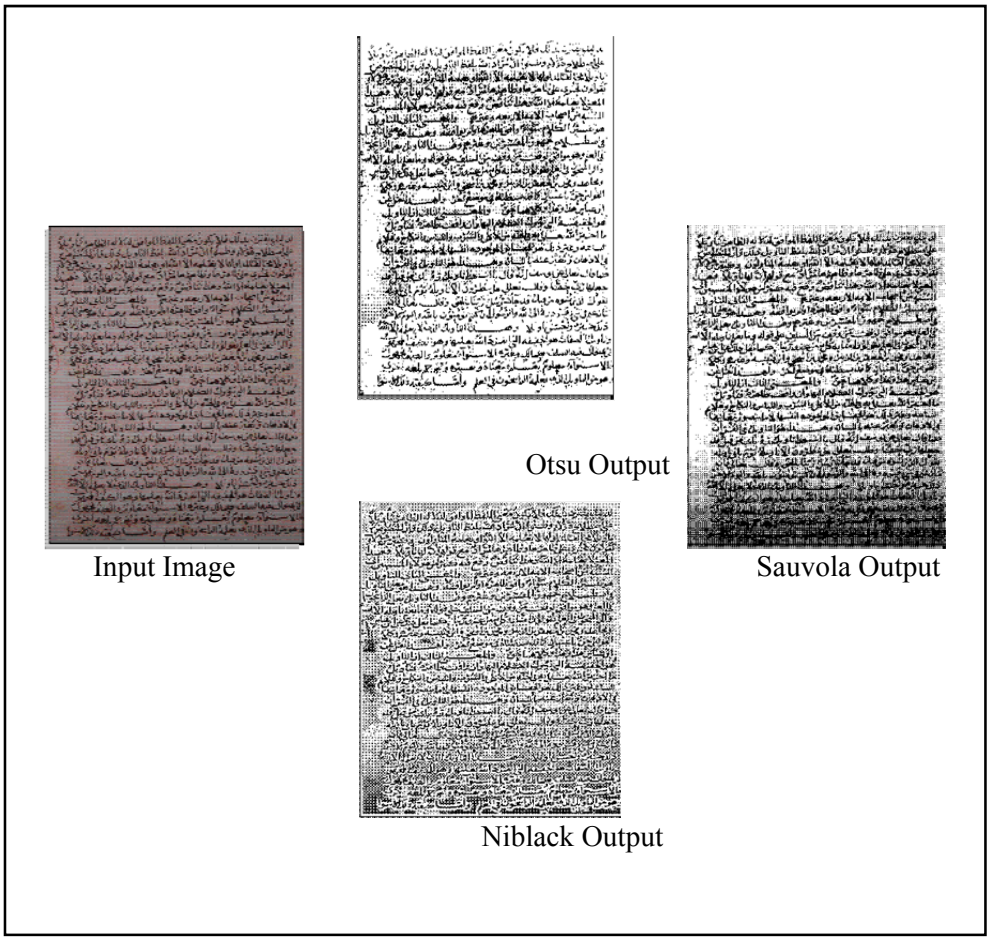
Niblack method [5] is based on the calculation of the local mean and of local standard deviation. It slide a rectangular window over the whole gray image to calculate the pixel wise threshold. The threshold T(X, Y) is indicated by the formula

$$T(X, Y) = m(X, Y) + K * s(X, Y) \quad \dots (1)$$

Where

m(X, Y) and s(X, Y) is the standard deviation values and average values of all the pixels in the rectangle window respectively and K is a constant (value between 0and 1). The size of the neighborhood consider the main problem of the method because it supposed to be small enough to keep the local details, but at the same time large to remove the noises as shown in figure(1).

Sauvola's algorithm [6] claims to improve Niblack's method by using the dynamic range of image gray-value standard deviation to computing the threshold. This method outperforms Niblack's algorithm in images where the text pixels have near 0 gray-value and the background pixels have near 255 gray. However, the results degrade significantly in images shown in figure (1) where the gray values of text and background pixels are close to each other.



Figure(1)Images results of various thresholding methods

Fuzzy C-Means Clustering

There are two main processes for Fuzzy c-means clustering which are: the mentioning of points to these centers using Euclidian distance and the calculation of cluster centers [7].To make the cluster centers are stable the process is repeated. For each item of the data for the clusters FCM assigns a membership value within a range of 0 to 1 [8]. So it combines the fuzzy set's [9] concepts of partial membership and forms overlapping clusters to support it [10]. A fuzzification parameter *m* is needed in range [1, n] that indicates the degree of fuzzinessin the clusters [11].FCM is depends on minimization the below objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=0}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad \ldots (2)$$

where

m is any real number bigger than 1, *C* is the number of clusters, *N* is the number of data, x_i is the *i*th of d-dimensional measured data, c_j is the d-dimension center of the cluster, u_{ij} is the degree of membership of x_i in the cluster *j*, and ||*|| is any norm expressing the similarity between any measured data and the center.

Through an iterative optimization of the above function Fuzzy partitioning is carried out, with the update of membership u_{ij} and the cluster centers c_j by:

$$\mathrm{u_{ij}} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|xi - cj\|}{\|xi - ck\|} \right)^{\frac{2}{m-1}}} \quad \ldots (3)$$

Where
$||x_i - c_j||$ is the Distance between point *i* and current cluster center *j*, $||x_i - c_k||$ is the Distance between point *i* and other cluster centers *k*.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}}{\sum_{i=1}^{N} u_{ij}^{m}} \frac{x_i}{} \quad \ldots (4)$$

It is stop when, $max_{ij}\left\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\right\} < \varepsilon$ where ε is between (0) and (1), for a termination criterion, whereas *k* are the iteration stages.
However, for selecting the optimal threshold for the given histogram function h(z) , kittler and Illingworth minimum error thresholding function have used. Assume object pixels are distributed (histogram) according to $p_o(x)$ and the background pixels by $p_B(x)$. The error is misclassifying object pixels is:

$$\int_{-\infty}^{t} p_O(x)dx \quad \ldots(5)$$

And misclassifying background pixels as object pixels is

$$\int_{t}^{\infty} p_B(x)dx \quad \ldots(6)$$

Let θ be the fraction of pixels in the object. Then the total error is

$$E(t) = \theta \int_{-\infty}^{t} p_O(x)dx + (1-\theta)\int_{t}^{\infty} p_B(x)dx \quad \ldots (7)$$

To find the minimum

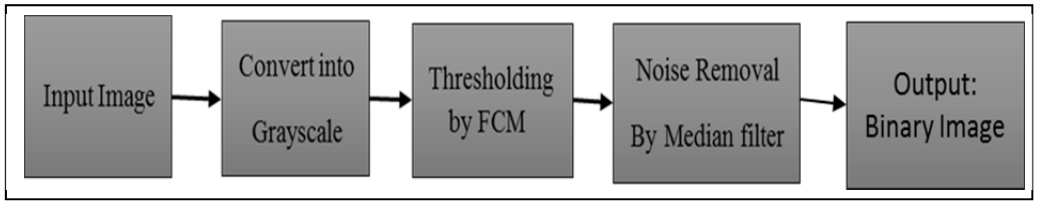
$$0 = \frac{\partial E}{\partial t} = \theta\, p_O(t) - (1-\theta)\, p_B(t)$$

or

$$\theta\, p_O(t) = (1-\theta)\, p_B(t) \quad \ldots (8)$$

PROPOSED THRESHOLDING METHOD

This section presents the proposed method to describe the thresholding process. The proposed method has three steps. First, convert input image into grayscale image. Second, use FCM to convert the grayscale image into binary image. Last step, remove unwanted noises using median filter from the image.Figure2 shows the proposed method steps.



Figure(2) Steps of the proposed method

After converting the image into grayscale the intensity value of pixels (I_{max} and I_{min}) has been calculated then find the mean and different between the max and min intensity .Furthermore , the result will be as input to the FCM to attract the nearest similar pixels by clustering operation.Algorithm1 describes the main steps of the proposed method.

Algorithm1: Arabic Handwriting Image Thresholding (AHIT)
Input: Color image(Type: .jpg , .bmp ; Size: any ; Resolutions: 128*128 or more)
Output: Binary image

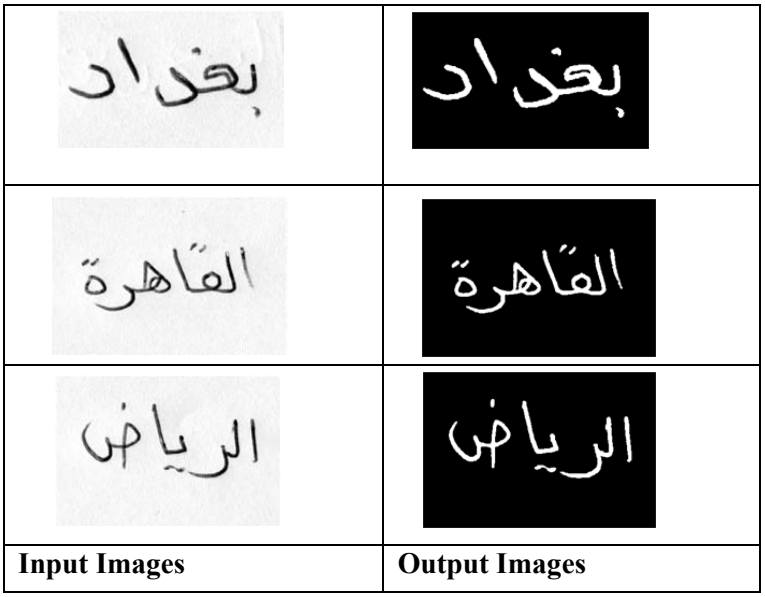
Step1: Read an RGB image
Step2: Convert RGB image into grayscale
Step3: Calculate the $\mathbf{I_{Max}}$ and $\mathbf{I_{Min}}$ intensity of pixel through:
$\mathbf{I_{Max}}$ = max (grayimage(:));$\mathbf{I_{Min}}$ = min(grayimage(:));
Step4: Find the $\mathbf{I_{Mea}}$ of the intensity through
$\mathbf{I_{Mea}}$ = mean (grayimage(:));
Step5: Find the different between the $\mathbf{I_{Max}}$ and $\mathbf{I_{Min}}$ intensity
$\mathbf{I_{diffr}= I_{Max} - I_{Min}}$
Step6: IF the $\mathbf{I_{Mea}}$ of the intensity bigger than the different between the $\mathbf{I_{Max}}$
And $\mathbf{I_{Min}}$,then→ save **I=120 , Else I=0**
Step7: Feed the **FCM** with the kept pixel value (**I**) from the previous step
Step8: Use **median filter** to remove the noise
Step9: Display the output image

Experimental Results and Discussions

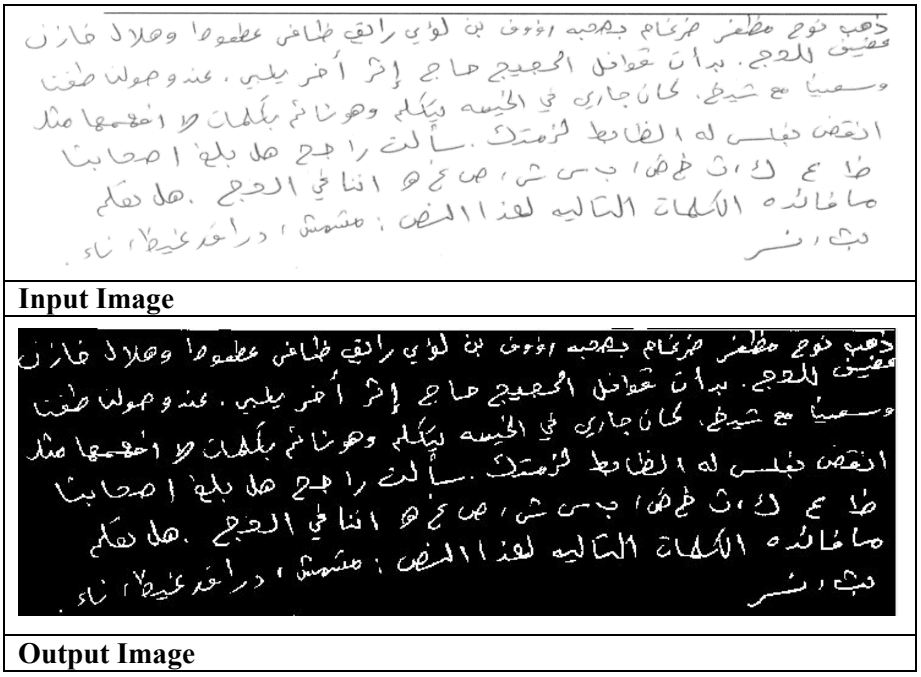
The proposed method is implemented using Matlab R2015a version, under windows7 64-bit Operating System, with RAM 6GB, CPU 2.50GHz core i5 and it achieved fast and effective results. Besides, the proposed method has been tested for several handwriting dataset and it gives a very good results.

The intensity of pixel value determined the black and white pixels in the image then the FCM will use the black/white as input, and select several black/white pixels to use it as centroid for the clustering operation. Moreover, the proposed method provide better results comparing with the current methods, and it has big advantage which is ,the user no need to select the number of clusters because FCM specify it automatically.

Ideally, after applying the proposed method on several handwriting images databases, the proposed method gives a best result comparing to the existing methods. Figures (3, 4) illustrates the result of applying our method on IESK-arDB database [12] and KHATT [13] database and how clear the output binary images will be. In the Figure (3,4) the background is used in the color black to make the result more clear.



Figure(3) The output of apply proposed method on IESK-arDB database

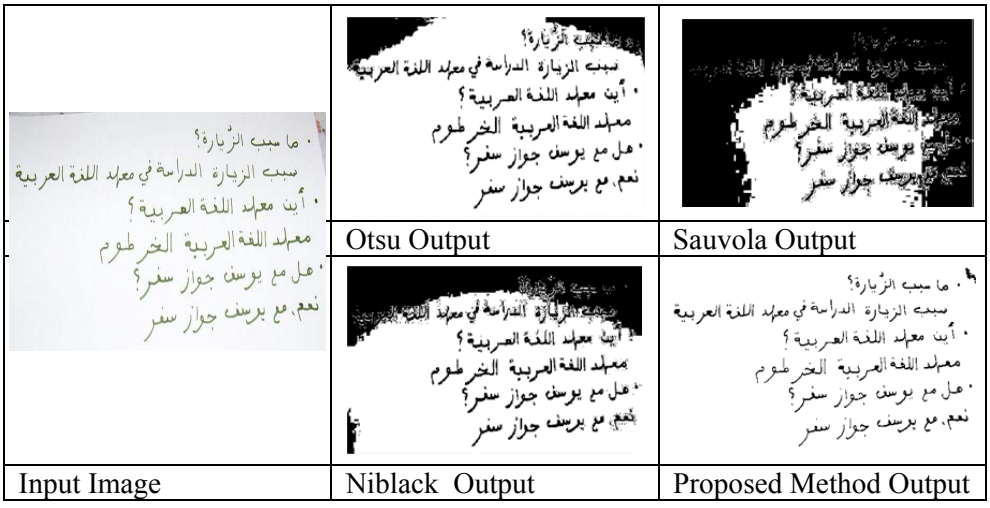


Input Image

Output Image

Figure (4) The output of apply proposed method on KHATT database

After 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 figure5 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[14].



Otsu Output | Sauvola Output

Input Image | Niblack Output | Proposed Method Output

Figure (5) Comparative outputs of various thresholding methods and the proposed method

After testing several values between (0 and 200) to get best variation range .Therefore, 0 and 120 was the best offset variation range of text. In addition, if difference between foreground and background intensity, that is $I_{diff}>I_{mean}$ (average value), then the possible variation range of text content in also in large range and so offset (I) is set to be 120 and for small difference, the

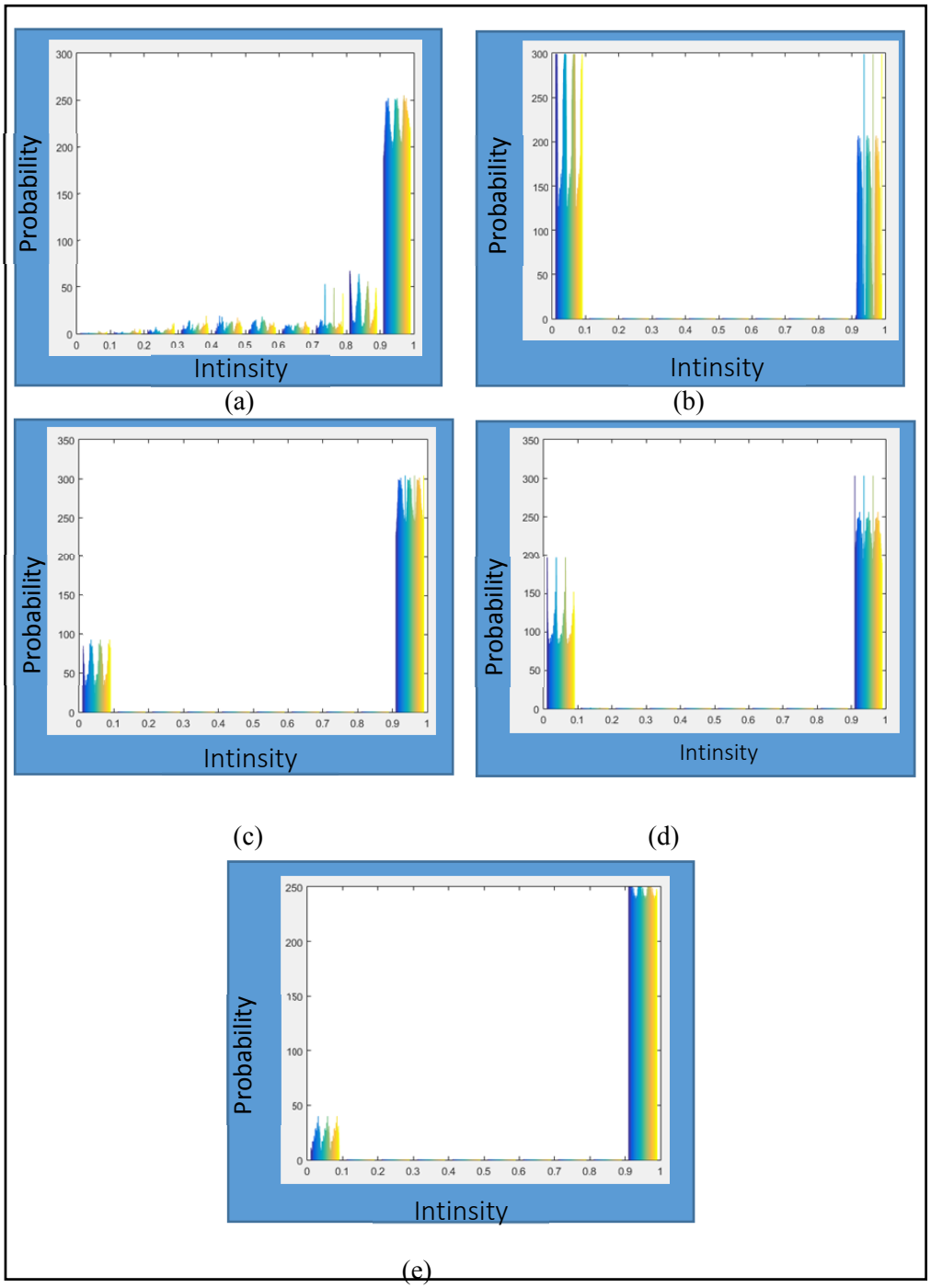
variation range of the text content will also be small. Thus, the offset (I) is set to be 0 a small value.

Last step of the proposed method is removing the unwanted noise. Median Filter (medfilt) command in Matlab has been used for removing undesired noise and detecting the misclassified pixels for white color (0) background and black color (1) foreground in the existing methods and the proposed one as shown in Table1 below.

Table (1). The misclassified pixels for white (0) background black (1) foreground existing and proposed method

The all methods misclassified pixels	Sauvola	Otsu	Niblack	Proposed method
Foreground	120	20	60	5
Background	70	30	40	15
Total	190	50	100	20

Furthermore, the proposed method has been select the best distinguish point (thresholding point) that isolate the foreground from the background of the result image which avoid the overlapping between the front and back color. Image histogram (imhist) command has been used for view the histograms of the output images which illustrates in Figure (6).



Figure(6). Histograms of thresholding results: (a) Original image (b)Sauvola's method (c) Otsu's method (d) Niblack's method (e) proposed method.

The histograms of the output images of the different threshold methods and the proposed one in figure(6(e))shows that, the intensity of the white color (0) for background and the black color (1) for the foreground color of the exist methods have a lot of overlapping between the two colors black/white. Since the perfect result of the thresholding method supposed to make the histogram more closely to either black color or the white color. According to the input image

the black color has the highest intensity and the white color has the lowest intensity that make the statistical measurement in the histogram look very high for the black and very low to the white. Therefore, the proposed method gave the best result to biniraize the image clearly.

CONCLUSION

In this a paper, a proposed method based on using FCM for thresholding of Arabic Handwriting images has been presented. The results and performance have been compared with other methods and are quite satisfactory. Experiments, our proposed method is gave best results in terms number of noises and misclassified pixels in the result images. The proposed method gave more clear and free noise binary image comparing to existing methods. Experimental results show that the offset values of 0 for lower intensity variations and120 for large variations could cater to general pen pressure and color variations, and hence gives very good results.

REFERENCES

[1] R.Garnett , T. Huegerich, C. Chui and W. He,.A New Framework for Removing Gaussian and Impulse Noises.,viewed8March 2015,at http://www.cs.umsl.edu/~chui/publ/GHCHnoise.pdf.

[2] Hadeel and Ali ,Image Denoising Using Frame let Transform, Eng. & Tech. Journal, Vol.28, No.13, 2010

[3] Hadeel , Jabir and Arshad . Complex Discrete Wavelet Transform-Based Image Denoising.Eng. & Tech. Journal, Vol.29, No.5, 2011

[4] Otsu, N. A threshold selection method from gray level histograms. IEEE Trans. on Systems, Man and Cybernetics, Vol. 9, pp.62-66,(1979).

[5] W. Niblack, "An Introduction to Digital Image Processing.", Englewood Cliffs, New Jersey: Prentice-Hall, 1986.

\[6] J.Sauvola, T,Seppanen, S.Haapakoski, M.Pietikainen, "Adaptive Document Binarization", 4th Int. Conf. On Document Analysis and Recognition, Ulm, Germany, pp.147-152 (1997).

[7] Fayyad, U., Uthurusamy, R.: Data mining and knowledge discovery in databases. ACMCommun. 39, 24–27 (1996)

[8] Au, W.H., Chan, K.C.C.: Classification with Degree of Membership: A Fuzzy Approach.In: Proceedings IEEE International Conference on Data Mining, ICDM 2001 (2001)

[9] Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainty and Information. Prentice Hall, Engle wood Cliffs (1988)

[10] Pal, K., Mitra, P.: Data Mining in Soft Computing Framework: A Survey. IEEE transactions on neural networks 13(1) (January 2002)

[11] Cox, E.: Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. Elsevier, Amsterdam (2005)

[12]Elzobi, Moftah, Ayoub Al-Hamadi, Zaher Al Aghbari, and Dinges, Laslo, 'IESK_arDB database' viewed 9March 2015, at http://www.iesk-ardb.ovgu.de/

[13]Sabri A. Mahmoud : KHATT (KFUPM Handwritten Arabic TexT) database, King Fahd University, viewed 5March 2015, athttp://khatt.ideas2serve.net/

[14] Niall, Gunter, Innsbruck: Optical Character Recognition. Informatics Research Institute (IRIS) at University of Salford(2011)