ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

A New Hybrid Binarization Algorithm for Manuscripts

Ashraf Nijim[1], Muhammad AboKresha[2], Ayman El Shenawy[3], Reda Abo Alez[4]

Student, Computers and Systems Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, Egypt[1]

Lecturer, Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt[2]

Lecturer, Computers and Systems Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, gypt[3]

Professor, Computers and Systems Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, gypt[4]

Abstract: In this paper a new binarization algorithm for ancient manuscripts and historical documents with bleeding noise has been proposed. This algorithm consists of three primary processes. In the first process, a given gray-scale image has been classified into three classes: black-foreground pixels class, white-background pixels class and confused pixels class. In the second process, the confused pixels class will be classified into either of the two black and white classes. The classified image was cut into rectangles using the confused-pixels vertical and horizontal histograms. Each rectangle is a sub-image containing a region of the image with pixels having similar properties. The third is a voting process where a threshold value is selected to binarize each sub-image separately. Seven thresholding values driven from six different global binarization techniques contribute to the voting process. The binarized image is the collection of the sub-images binarization results. Four different measuring metrics have been used to evaluate the results of the proposed algorithm. The performance of the algorithm has been compared with two widely used binarization algorithms which yield a significant improvement in the binarization process of ancient manuscripts and historical documents with bleeding noise.

Keywords: binarization, thresholding, hybrid algorithm, global binarization algorithms, image measuring metrics

I. INTRODUCTION

Binarization is an essential pre-processing technique in most image processing and pattern recognition systems. The importance of this technique realise on the fact that it changes colour and gray images to simple for processing binary images. The binarization process is regarded as extracting the foreground text and shapes from any background and save it as black pixels on a white sheet. Manuscripts suffer from many types of background noise that makes the binarization process a challenge. Background noise takes many forms such as aging, deformations, ink spots, complex backgrounds, bleed through, poor quality ink and paper, and intensity degradations. Every type of this noise adds a difficulty to the binarization process. Many ancient manuscripts suffer from at least one type of background noise.

Binarization algorithms can be classified into three kinds: global; local; and hybrid [1], [2]. Global binarization algorithms work on the whole image to find a threshold value. Pixel values less than this value are considered as black foreground pixels and values greater than this threshold value are considered as white background. Otsu thresholding algorithm [3] is one of the well-known global binarization algorithms used widely in many applications. Local binarization algorithms use the information of a local window with a predefined size.

Each pixel is classified according to the local information contained in this fixed sized-window. This type of binarization works well with images that have complex background and intensity degradations. Sauvola [4], Niblack [5], Bernsen [6], and Gatos [7] are frequently used local binarization algorithms. Hybrid binarization techniques use a combination of global and local binarization algorithms. This kind of technique benefits from the fast global algorithms and the good results of the local algorithms to produce a binarized image. Sokratis et al. [8] is one of the newly developed hybrid binarization algorithms.

In this paper a new binarization algorithm for ancient manuscripts and historical documents with bleeding noise is proposed. This algorithm is composed of three processes named: classification, segmentation, and voting. The image is divided into different sized rectangles. Each rectangle in the given image chooses a binarization threshold among seven different values.
The reminder of this paper is organized as follows. Next section, we present six global binarization algorithms. Then, we introduce the new proposed binarization algorithm. After that we present the new algorithm results, and finally we draw our conclusions.

II. GLOBAL BINARIZATION ALGORITHMS REVIEW

Global thresholding algorithms are simple and fast. They use only one value to classify the gray-scale image into two classes: black pixels and white background. Otsu [3] is one of the most useful global thresholding algorithms. It

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

uses the mean and the standard deviation values of the image histogram to find a threshold. This threshold value classifies the whole image into two classes by maximizing the between-class variance of the two classes 0 and 1. The class variance is given by:

$$\sigma_B^2 = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2$$

Where,
ω_0 and ω_1 are the class 0 and class 1 occurrence respectively,
μ_0, μ_1, and μ_T are the class 0 mean, class 1 mean, and the total mean level of the original image respectively.

$$\sigma_B^2(k^*) = \max_{1\le k\le L} \sigma_B^2(k)$$

Where,
k^*, the optimal threshold,
L, the maximum gray-level of the image,
k, threshold level.
Moment-preserving thresholding algorithm [9], [10] uses the gray-level moments of the image to select a threshold. The selected threshold of the newly bi-level image must preserve the first three moments of the original image. The first three moments of the tested bi-level image is compared to those of the original image. If the three equations are equal then the selected threshold value used is a moment preserving threshold value. The moment of a gray-scale image is given by:

$$m_i = (1/n)\sum_j n_j (z_j)^i, \qquad i = 1,2,3,\ldots$$

m_i, the i_{th} moment,
n, total number of pixels in the image,
n_j, the total number of pixels in the image with gray-value z_j.
And,

$$m_i' = \sum_{j=0}^{1} p_j (z_j)^i, \qquad i = 1,2,3.$$

m_i', the i_{th} moment of the resulting bi-level image,
j, class number,
$p_j = \frac{n_j}{n}$, fraction of pixels.
The equality equation is:

$$m_i' = m_i, \qquad i = 1,2,3.$$

Entropy thresholding algorithm [10], [11] divides the image into two probability distribution functions. One of these distribution functions is used for the black pixels and the other for the background white pixels. The sum of entropies of these probability distributions should be maximized to find the threshold value. The summation A is defined as:

$$A_j = \sum_{i=0}^{j} y_i$$

y_i, number of pixels in the image with gray-level i.
And,

$$E_j = \sum_{i=0}^{j} y_i \, log y_i, \qquad j = 0,1,\ldots,n$$

E_j, entropy partial sum,
n, maximum gray-value (n = 255).

$$\varphi(j) = \frac{E_j}{A_j} - \log A_j + \frac{E_n - E_j}{A_n - A_j} - \log(A_n - A_j)$$

$\varphi(j)$, sum of entropies for the two probability distribution functions.
The fuzzy c-means clustering (FCM) technique objective is to cluster data into 3 classes. The fuzzy c-mean [12], [13] function J_m is defined as:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^m \left\|x_i - C_j\right\|^2$$

m, weighting exponent greater than 1,
u_{ij}, degree of membership of x_i in the cluster j,
x_i, the i^{th} of d-dimension measured data,
C_j, the d-dimention center of the cluster j,
$\|*\|$, norm expressing the similarity between any measured data and the center,
And,

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m},$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - C_j\|}{\|x_i - C_k\|}\right)^{\frac{2}{m-1}}}$$

Iteration stop condition is:

$$max\{|u_{ij}^{k+1} - u_{ij}^k|\} < \varepsilon$$

Where,
k, is the iteration step, and,
ε, between 0 and 1.
Two binarized thresholding values have been computed using FCM, one using the small and middle classes, and the other using the middle and large classes.
Kittler & Illingworth [14] presented the minimum error thresholding (MET) algorithm. The MET uses the amount of overlap between Gaussian models of the object and the background populations. The method objective is to find a threshold value T that minimizes this overlap area. The criterion function $J(t)$ is defines as follows:

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

$$J(t) = 1 + 2\{P_1(t)log_e\sigma_1(t) + P_2(t)log_e\sigma_2(t)\} - 2\{P_1(t)log_eP_1(t) + P_2(t)log_eP_2(t)\}$$

Where,

$$P_i(T) = \sum_{g=a}^{b} h(g),$$

$$\mu_i(T) = \frac{\sum_{g=a}^{b} h(g)g}{P_i(T)},$$

$$\sigma_i^2(T) = \frac{\sum_{g=a}^{b}\{g-\mu_i(T)\}^2 h(g)}{P_i(T)},$$

$$a = \begin{cases} 0, & i = 1 \\ T+1, & i = 2 \end{cases}$$

$$b = \begin{cases} T, & i = 1 \\ n, & i = 2 \end{cases}$$

And,
h(g), is the gray level image histogram,
g, the gray level value and is between [0, n]
And, T is found by:

$$J(T) = min_t J(t).$$

Xu et al. [15] presented a novel range-constrained Otsu method for binarization. The new method is a three step process described in the following algorithm:

Range-Constrained Otsu Method Algorithm

1. Find the threshold value T_o running the original Otsu algorithm.
2. Run Otsu algorithm again but only for pixel values between 0 and T_o. The computed threshold value is T_{new}.

Use the new threshold value T_{new} to binarize the image.

III. New Binarization Algorithm

The new algorithm consists of three processes: classification; segmentation; and voting. During the classification process given image pixels are classified into three categories: black; white; and red-confused pixels. The image is then segmented according to the red-confused pixels histogram distribution. The segmented small-images are then binarized using voting between seven different global thresholding values. The binarized image is the collection of those small-images results.

A. Classification

The gray-scale image is to be classified into three categories using Bataineh et al. [16].

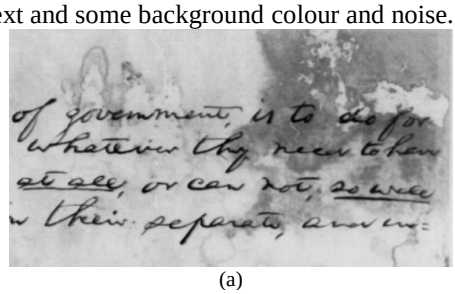
$$I(x,y) = \begin{cases} black, & i(x,y) \le T_{con} - \left(\frac{\sigma_g}{2}\right), \\ red, T_{con} - \left(\frac{\sigma_g}{2}\right) < i(x,y) < T_{con} + \left(\frac{\sigma_g}{2}\right), \\ white, & i(x,y) \ge T_{con} + \left(\frac{\sigma_g}{2}\right). \end{cases}$$

where,
$I(x,y)$, the result image,
$i(x,y)$, the pixel value,
σ_g, the global standard deviation of all the pixel-values in the image,
T_{con}, the confused threshold of the image, and can be calculated by:

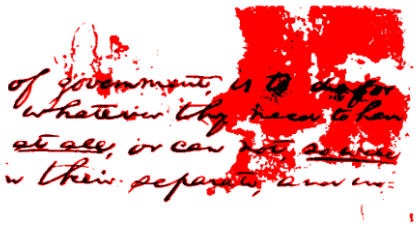
$$T_{con} = m_g - \frac{m_g^2 - \sigma_g}{(m_g + \sigma_g)\left((0.5max_{level}) + \sigma_g\right)},$$

where,
m_g, the mean value of the image pixel-values,
max_{level}, the maximum level of the image pixel-values.

Fig. 1 shows the result of the classification process. The results of a handwritten image from DEBCO 2009 [17] contest showed that the red confused class represents parts of the text and some background colour and noise.



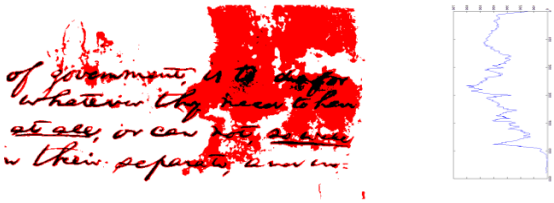
(a)



(b)

Fig. 1. Example of image classification, black pixels represent foreground pixels, red pixels represent confused class pixels, and white pixels represent background pixels (a) Original image, (b) Classified image

The segmentation process starts with computing the histogram of the red-confused pixels of the classified image in both directions – the vertical and horizontal (see Fig. 2). This histogram results are then used to segment the red pixels into rectangles. Each local minima of the smoothed histogram is considered as a segmentation line. Two adjacent vertical or horizontal segmentation lines represent two edges of the rectangle sub-image. The more we have minima points the more we are able to divide an image into sub-images.



ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

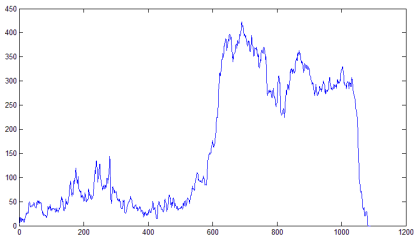


Fig. 2. Vertical and horizontal histogram of the red-configured pixels

B. Segmentation

The segmentation process starts with computing the histogram of the red-confused pixels of the classified image in both directions – the vertical and horizontal (see Fig. 2). This histogram results are then used to segment the red pixels into rectangles. Each local minima of the smoothed histogram is considered as a segmentation line. Two adjacent vertical or horizontal segmentation lines represent two edges of the rectangle sub-image. The more we have minima points the more we are able to divide an image into sub-images.

After dividing the image according to the red-confused pixels histogram into small rectangles, we start to join adjacent rectangles having the same red-pixels concentration. The algorithm for joining rectangles is described in the following algorithm:

Join Adjacent Rectangles Algorithm

1. From the list of rectangles, find the rectangle having the maximum number of red-pixels. Mark this rectangle as the main rectangle.
2. Find the main rectangle join rate:

$$join\ rate_{main\ rectangle} = \frac{red\ pixels\ count\ inside\ the\ rectangle}{rectangle\ size}$$

3. Find the $join\ condition$:
4.

$$join\ condition = join\ rate_{main\ rectangle} * join\ constant$$

Where the $join\ constant$ is a predefined number less than 1. We have set this constant to 0.5 during our experiments.

5. From the list of rectangles, and for each one adjacent to the main rectangle:

a. Find the rectangle's join rate $join\ rate_{rectangle}$.

b. If the rectangle join rate is greater than or equals to the main rectangle's $join\ condition$, then join the current and main rectangles to form a bigger rectangle.

$$join\ rate_{rectangle} \geq join\ condition$$

c. Remove the rectangle form the list of rectangles.

6. Remove the main rectangle form the list of rectangles.
7. Loop: go to step 1, until the list of rectangles is empty.

The final result of this process is an image divided into small sub-images (rectangles). Each sub-image has part of the red-confused pixels, and may contain white and/or black pixels; some of those sub-images are shown in Fig. 3.

C. Voting

The original gray-scale image was divided into smaller sub-images during the segmentation process. Each sub-image is a rectangle with two coordinate points. Every sub-image is then imposed to the six global binarization techniques mentioned before in the global binarization algorithms review section. At the end of this process we get seven global binarization thresholding values.

$$threshoulds = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$$

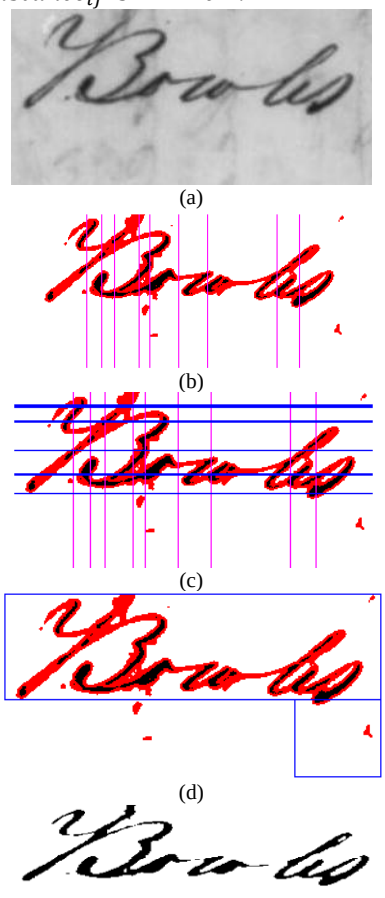
The distances between those seven thresholds are found as follows:

$$distance_{ij} = |a_i - a_j|, i \neq j, i, j = 1,2 \ldots,7$$

The two thresholding values having the smallest distance are chosen as the best two values for the thresholding process. The average of those two values is the sub-image threshold value (Sub_image_{th}) to be used during the binarization process,

$$Sub\_image_{th} = \frac{a_i + a_j}{2},$$

Where, $distance_{ij}$ is minimum.



(a)

(b)

(c)

(d)

(e)

DOI 10.17148/IJARCCE.2015.44134

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

The newly binarized sub-image is then inserted into its proper place in the white and black pixels image previously found during the classification process. Fig. 4 shows a sample of the sub-images binarization results.
The final binarized image is the collection of all the binarized sub-image results.

IV. RESULTS

The proposed algorithm was tested on degraded documents from two contests DIBCO 2009 [17], [18] and ICFHR 2010 [19]. The results are evaluated using four measuring metrics widely used in evaluating binarization techniques. The four measures consist of F-measure [20], geometric accuracy [19], peak signal-to-noise ratio (PSNR), and negative rate metric (NRM) [17], [18].

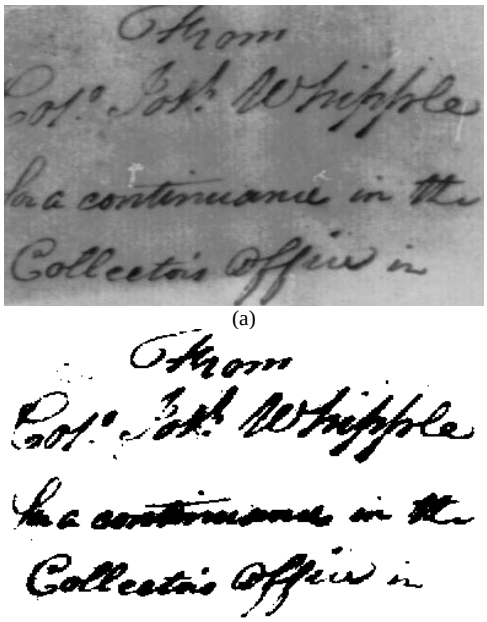


Fig. 4. Examples of sub-images and their binarized results, (a) Original sub-image, (b) Binarized sub-image.

Fig. 5 shows some of the results of the proposed algorithm for ancient printed and handwritten text. The results illustrate the capability of the algorithm to work on different backgrounds having different types of noise. The results presented in this section are raw image results that did not undergo any post processing such as smoothing or edge enhancement.

The metric results for a sample of three different types of degraded images are shown in Table 1, 2 and 3. Table 1 presents a sample of six images from three different sets of document images. Four of them are printed and handwritten text images from DIBCO2009 contest, and the other two are text documents with bleeding noise from ICFHR 2010 contest. The four measuring metrics results showed the good performance of the proposed algorithm.

TABLE 1
PROPOSED ALGORITHM RESULTS ON DEGRADED HANDWRITTEN AND PRINTED DOCUMENTS

Image file name	F-measure (%)	Geometric Accuracy	PSNR	NRM
H03*	85.72512	0.95434	15.19420	0.04547
H04*	82.48361	0.91330	15.78012	0.08419
P01*	89.46990	0.95374	15.78987	0.04591
P02*	93.87143	0.95659	16.00151	0.04293
2_10**	96.34177	0.99706	23.09820	0.00294
1_5m**	97.43154	0.99839	25.27116	0.00161

*H03, H04, P01, and P02 are part of the DIBCO2009 contest test images.
** 2_10 and 1_5m are part of the ICFHR 2010 contest test images.

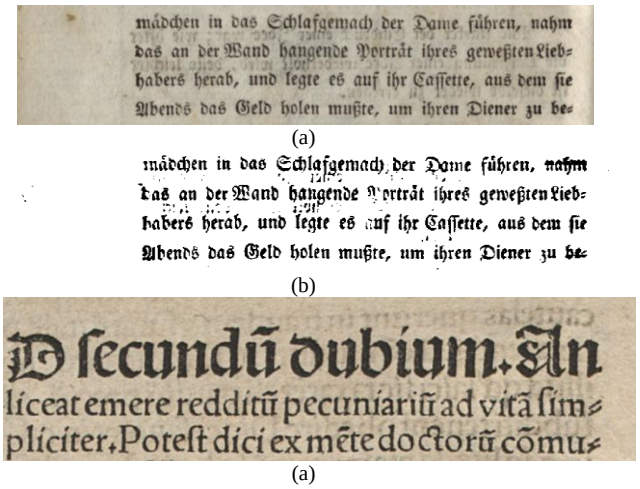


Fig. 5. Proposed algorithm results, (a) Original images, (b) Binarized images using the proposed algorithm.

DOI 10.17148/IJARCCE.2015.44134

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Table 2 and 3 show a comparison between the proposed algorithm and two well-known binarization algorithms, Otsu [3] and Sauvola & Pietikainen [4]. Two images, handwritten and printed text documents are used in this comparison. The four measuring metrics results demonstrate the improvement of the proposed binarization algorithm over the two algorithms specially the F-measure metric.

TABLE 2
HANDWRITTEN DEGRADED DOCUMENTS COMPARISON RESULTS

Measuring metric[***]	**Proposed algorithm**	**Otsu**	**Sauvola et al.**
F-measure	82.48361	40.55702	30.91580
Geometric Accuracy	0.91330	0.87294	0.80365
PSNR	15.78012	6.73124	4.85058
NRM	0.08419	0.12046	0.17737

[***] Using H04 handwritten test image from DIBCO2009 contest.

TABLE 3
PRINTED DEGRADED DOCUMENTS COMPARISON RESULTS

Measuring metric[****]	**Proposed algorithm**	**Otsu**	**Sauvola et al.**
F-measure	89.46990	88.92597	51.16599
Geometric Accuracy	0.95374	0.96918	0.85893
PSNR	15.78987	15.36796	6.38121
NRM	0.04591	0.03081	0.13124

[****] Using P01 printed test image from DIBCO2009 contest

Table 4 shows another comparison between the proposed algorithm and the two binarization algorithms, Otsu [3] and Sauvola & Pietikainen [4]. The four measuring metrics results of a historical document with bleeding noise also demonstrate the improvement of the proposed algorithm over the two well-known algorithms.

TABLE 4
HISTORICAL DOCUMENTS WITH BLEEDING NOISE COMPARISON RESULTS

Measuring metric	**Images of historical documents with bleeding noise**[+]		
	Proposed algorithm	**Otsu**	**Sauvola et al.**
F-measure	48.79574	27.03121	23.60881
Geometric Accuracy	0.93454	0.82316	0.78328
PSNR	9.27953	5.16784	4.38103
NRM	0.06344	0.16120	0.19324

[+] Using 1_10m document image from ICFHR 2010 contest.

V. CONCLUSION

The proposed binarization algorithm managed to divide a given image into sub-images each containing a region of pixels with similar properties. Each sub-image is treated separately. A suitable threshold value was selected among seven different threshold values driven from six different global binarization algorithms. The results of the individual sub-images are then combined to form a binarized image that is proved to perform better than two of the widely used binarization algorithms Otsu and Sauvola et al. The resulting binarzied images showed the capabilities of the proposed algorithm to sub-divide the image into similar-in-properties areas. In addition, the algorithm was able to select a suitable binarization threshold value for each sub-image. The results were also evaluated and compared using four measuring metrics widely used in similar assignments. The proposed algorithm showed a significant increase in f-measure, geometric accuracy, and PSNR metric values, and a significant decrease in negative rate metric values compared to Otsu and Sauvola et al.

In the future, the selection method of the threshold value is to be modified. More thresholding techniques will be added, and the selection process will benefit from all the available threshold values.

As another further work, the effect of various filters will be tested on each of the three processes of the proposed algorithm and on the final binarized image result.

ACKNOWLEDGMENT

The authors would like to thank **Dr. Mahmoud Khalil**, from the Faculty of Engineering, Ain Shams University, Cairo, Egypt for his early contribution to this work.

REFERENCES

[1] P. Stathis, E. Kavallieratou, and N. Papamarkos, "An Evaluation Technique for Binarization Algorithms," Journal of Universal Computer Science, vol. 14, No 18, pp. 3011-3030, 2008.

[2] J. Kaur, and R. Mahajan, "A Review of Degraded Document Image Binarization Techniques," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, issue 5, 2014.

[3] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, SMC-9(1), pp. 62-66, 1979.

[4] J. Sauvola, and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition, vol. 33, pp. 225-236, 2000.

[5] W. Niblack, "Segmentation," An Introduction to Digital Image Processing Englewood Cliffs, NJ, USA. PA: Prentice Hall, pp. 113-117, 1986.

[6] J. Bernsen, "Dynamic thresholding of gray-level images," Proc. of the 8th Int. Conf. on Pattern Recognition, Paris, pp. 1251–1255, 1986.

[7] B. Gatos, I. Pratikakis, and S. J. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents," Document Analysis Systems VI, Lecture Notes in Computer Science, vol. 3163, pp. 102-113, 2004.

[8] V. Sokratis, E. Kavallieratou, R. Paredes, and K. Sotiropoulos, "A Hybrid Binarization Technique for Document Images," Learning Structure and Schemas from Documents Studies in Computational Intelligence, vol. 375, pp. 165-179, 2011.

[9] W. Tsai, "Moment-preserving thresholding: a new approach," Computer Vision, Graphics, and Image Processing, vol. 29, pp. 377-393, 1985.

[10] C. Glasbey, "An Analysis of Histogram-Based Thresholding Algorithms," CVGIP: Graphical Models and Image Processing, vol. 55, pp. 532-537, 1993.

[11] J. Kapur, P. Sahoo, and A. Wong, "A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram," Computer Vision, Graphics, and Image Processing, vol. 29, pp. 273-285, 1985.

[12] R. Babuska, "*Fuzzy clustering",* Fuzzy and Neural Control, Disc Course Lecture Notes. PA: Delft University of Technology, Delft, the Netherlands, pp. 55-71, October 2001.

[13] J. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, N. 2-3, pp. 191-203, 1984.

[14] J. Kittler, and J. Illingworth, "Minimum Error Thresholding," Pattern Recognition, vol. 19, issue 1, pp. 41-47, 1986.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

[15] X. Xu, S. Xu, L. Jin, and E. Song, "Characteristic Analysis of Otsu Threshold and Its Applications," Pattern Recognition Letters, vol. 32, pp. 956-961, 2011.
[16] B. Bataineh, S. Abdullah, and K. Omar, "An Adaptive Local Binarization Method for Document Images Based on a Novel Thresholding Method and Dynamic Window," Pattern Recognition Letters, vol. 32, issue 14, pp. 1805-1813, 2011.
[17] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," 10th International Conference on Document Analysis and Recognition. pp. 1375-1382, 2009.
[18] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: document image binarization contest," International Journal on Document Analysis and Recognition (IJDAR), vol. 14, I. 1, pp. 35-44, 2011.
[19] R. Paredes, E. Kavallieratou, and R. Lins, "ICFHR 2010 Contest: Quantitative Evaluation of Binarization Algorithms," International Conference on Frontiers in Handwriting Recognition ICFHR. IEEE Computer Society, pp. 733-736, 2010.
[20] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Management, vol. 45, pp. 427-437, 2009.