UNSUPERVISED TEXT BINARIZATION IN HANDWRITTEN HISTORICAL DOCUMENTS USING K-MEANS CLUSTERING

Huseyin Kusetogullari

Department of Computer Science and Engineering, Blekinge Institute of Technology,
Karlskrona, Sweden.
(e-mail: hku@bth.se)

ABSTRACT

In this paper, we propose a novel technique for unsupervised text binarization in handwritten historical documents using k-means clustering. In the text binarization problem, there are many challenges such as noise, faint characters and bleed-through and it is necessary to overcome these tasks to increase the correct detection rate. To overcome these problems, preprocessing strategy is first used to enhance the contrast to improve faint characters and Gaussian Mixture Model (GMM) is used to ignore the noise and other artifacts in the handwritten historical documents. After that, the enhanced image is normalized which will be used in the postprocessing part of the proposed method. The handwritten binarization image is achieved by partitioning the normalized pixel values of the handwritten image into two clusters using k-means clustering with k = 2 and then assigning each normalized pixel to the one of the two clusters by using the minimum Euclidean distance between the normalized pixels intensity and mean normalized pixel value of the clusters. Experimental results verify the effectiveness of the proposed approach.

Index Terms— Handwritten text binarization, image processing, k-means clustering, document images.

1. INTRODUCTION

Recently, handwritten document images have become a major research subject in the areas of image processing and pattern recognition to resolve different handwritten image problems. Amongst these problems, handwriting binarization is one of the most important and challenging problem in the handwritten document images and it has been used in many applications such as text observation, segmentation, character detection and recognition [1, 2, 3, 4]. However, historical document images may be affected with various factors which may reduce the handwriting legibility and cause the degradation on the handwritten images. Some of these factors are characteristics of digital camera, noise, different lighting conditions, deterioration and faint handwriting on documents. Current handwriting binarization methods have mostly ignored such important factors in their methods which are often observed on images and this may cause reducing the accuracy rate of character recognition when they are applied to the historical document images. Therefore, it is necessary to take into account these factors in the handwritten binarization method for removing the artifacts for better binarization of handwritten.

Text binarization is defined as the process of finding set of text and non-text pixels on the document images. The set of pixels which comprise the binary image are obtained by using the handwritten image. Many methods have been proposed and developed to find the text binarization through the document images. The existing methods are usually based on thresholding approaches and they are mainly classified in two different categories, namely global and local. Global thresholding methods use a single threshold value which is applied to the whole image. On the other hand, local thresholding methods find a local thresholding value based on the statistics and parameters within a moving window (e.g. mean μ, standard deviation σ). For instance, Otsu [5] proposed a general thresholding method to obtain the binary image and Moghaddam et al. [6] improved the general thresholding text binarization method. In [7], local thresholding based text binarization approach is proposed which estimates local statistics of pixel intensities within a window and adapts the local threshold according to those local statistics. However, the method fails to remove the noise in the binary text image. Sauvola et al. [8] proposed adaptive document image binarization which is based on the local thresholding approach to determine the binary image and decrease the noise but the correct text detection rate is also low under strong undesired artifacts in the document images. In [9], edge-based local thresholding method is proposed to create the handwritten binary image and it uses several steps to achieve the result. Besides this, self-training learning-based document image binarization method is proposed in [10] and the method first divides document image pixels into three different groups which are foreground pixels, background pixels and undesired pixels. After that, proposed learning-method is trained from the given document images and undesired pixels are classified using the learned pixel classifier. The method is successful to create document image binary but training the decision making

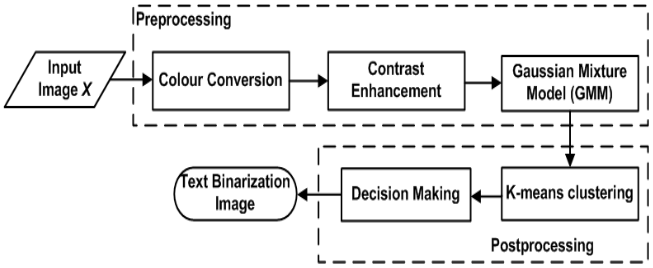
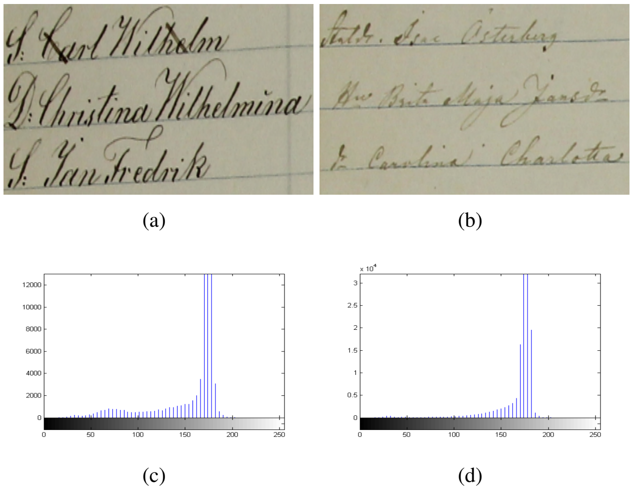


Fig. 1. Block diagram of the proposed method.

method is not an efficient approach. Furthermore, many other text binarization methods have been presented to create binary text image [11, 12, 13, 14]. Generally, thresholding methods have ability to detect the clear handwritten on the document image but they are unable to detect the faint characters or to remove the noise successfully. Therefore, using thresholding methods on handwritten document images may not provide effective results because of noise, faint characters and bleed-through.

In this paper, a novel technique is proposed for unsupervised text binarization in handwritten documents using k-means clustering. Unsupervised text binarization technique mainly uses the automatic analysis of handwritten document images and it is not necessary to train the system for the learning process of classifiers. In the proposed method, binary document image is created by using both pre-processing and postprocessing methods which are as follows: 1) Contrast enhancement and noise removal method; 2) Normalization of pixel values; and 3) k-Means clustering. In the proposed method, the contrast enhancement is first applied to the document image to improve the faint characters and then, Gaussian Mixture Model (GMM) is used to ignore the noise on the enhanced document image. In the final step of the pre-processing, each document image pixel is normalized and normalized pixel values are partitioned into two clusters using k-means algorithm. Each cluster is represented with a mean of normalized pixel values. After that, handwritten binary image is achieved by assigning each pixel of the normalized image to the one of the clusters according to the minimum Euclidean distance between its normalized pixel value and mean normalized pixel values of the clusters. Simulation results demonstrate an improvement of correct text binarization rate comparing to the state-of-the-art methods. Numerical experiments, on different handwritten document images, illustrate the effectiveness and efficiency of the proposed approach.

The rest of the paper is organized as follows. In Section 2, we describe the steps of our proposed method. Section 3 demonstrates the performance of the proposed approach. Section 4 concludes the paper.



(a) (b)

(c) (d)

Fig. 2. Different handwritten image examples, (a) Good quality, (b) Bad quality, (c) Histogram of good quality of image in (a), (d) Histogram of bad quality of image in (b).

2. PROPOSED METHOD

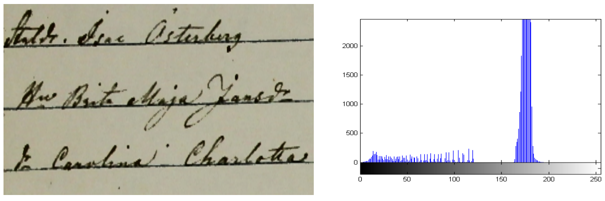
Let us consider a handwritten document image $X = \{x_1(i,j)| 1 \leq i \leq H, 1 \leq i \leq W\}$, with a size of $H \times W$. The purpose of the proposed method is to create a handwritten binary image that represents important (text) and unimportant (non-text) pixels occurred on the handwritten document image. The text binarization problem can be modeled as a binary classification problem and it is defined as:

$$X_b(i,j) = \begin{cases} 0, & \text{text pixel} \\ 1, & \text{non-text pixel} \end{cases} \quad (1)$$

where i and j are the pixel coordinates of X, X_b denotes the binary text image, '0' indicates that there is a text pixel intensity for the corresponding pixel but '1' indicates that there is no text pixel value on the document image. Obtaining the binary text image is a very complex problem. Therefore, we propose a new method to create binary text image from the handwritten document image. Let $\Delta = \{a_t, a_{nt}\}$ be the set of classes associated with important (denoted by a_t) and unimportant (denoted by a_{nt}) pixels on the image X. The proposed approach has three important steps to assign the pixel intensity values into two different clusters $\Delta = \{a_t, a_{nt}\}$, as shown in Fig. 1. In order to achieve the result, following steps are used: 1) Contrast enhancement and noise removal method; 2) Normalization of pixel values; 3) k-Means clustering technique with $k = 2$ to cluster the normalized pixel values into two clusters corresponding to a_t and a_{nt}.

2.1. Contrast Enhancement and Noise Removal Method

In order to increase the correct text detection rate, it is necessary to enhance the quality of text on the images because



(a) (b)

Fig. 3. Handwritten image using contrast enhancement, (a) Illustration of enhanced bad handwritten image, (b) Histogram of the enhanced handwritten image.

using faint characters in the unsupervised method will cause failing to detect the text. For instance, Fig. 2 shows two different handwritten document images. Fig. 2(a) has a high quality text image which is clear to observe and recognize the handwritten on the document image. On the other hand, another handwritten document image, shown in Fig. 2(b), has faint characters which are difficult to observe and recognize. Faint characters can be appearing broken or blurred on the image and this will cause reducing the correct detection rate of handwritten binarization. Besides this, histograms of two different image, shown in Figs. 2(c) and (d), indicate that it is necessary to apply contrast enhancement method to improve the faint characters on the document image shown in Fig. 2(b). Proposed method was combined with the modified histogram for contrast enhancement (MHCE) [15] to enhance the handwriting image. By enhancing the contrast of an image, shown in Fig. 2(b), will also improve the undesired artifacts on the image such as noise. In order to overcome the noise problem, background of the original image is combined with the foreground of the enhanced handwritten document image by using Gaussian Mixture Model (GMM) to ignore the noise on the background of the enhanced handwritten document image. Thus, the faint characters enhanced and improved, and various noises are ignored on the resulting image as shown in the histogram in Fig. 3(b).

2.2. Normalization

Another step of the proposed method is to normalize the input image X and normalized image X_n is defined as each pixel value of X is divided by the maximum pixel value of all pixel intensities of the input image X. Normalized image X_n can be defined as follows:

$$X_n(i,j) = X(i,j)/m \tag{2}$$

where m is defined as the maximum pixel value of all pixel intensities of X. As a result, the normalized pixel intensities will be in the region of [0,1] and it will be used in the post-processing of the proposed method.

2.3. Unsupervised based Document Image Binarization

After estimating the normalized image, unsupervised algorithm is used to make the decision whether the corresponding pixel intensity is text pixel or non-text pixel. To make the decision, we used k-means clustering method as a final step of the proposed method for creation of handwritten binary image which is an efficient and fast unsupervised learning approach [16]. The purpose of the clustering method is to partition the normalized pixel intensities into two different clusters for creation of binary image. In the proposed method, k is considered as 2 because the purpose of the method is to cluster the pixel intensities of the input image into two clusters and text pixels are represented as black pixel intensities and non-text pixels are represented as white pixel intensities in the resulting image. Thus, handwritten binary image is generated by using the unsupervised approach. In order to apply k-means algorithm, we will use two inputs which are the normalized pixel values and the number of clusters. Let μ_t and μ_{nt} be the two cluster means of normalized pixel intensities for text a_t and non-text a_{nt} classes, respectively. First, two mean values of two clusters are randomly chosen and the normalized pixels are labeled as text or non-text pixels by using the Euclidean distance technique [16]. Thus, labeled pixels are partitioned into two clusters and then update the mean values of two clusters over the normalized image. The process continues until no changes in the clusters are detected. The expectation is that the values of the normalized pixels to the μ_t are smaller than the values of normalized pixels to the μ_{nt}. The unsupervised thresholding approach is mathematically defined as follows:

$$X_b(i,j) = \begin{cases} 0, & W_t \leq W_{nt} \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

where,

$$W_t = (x_n(i,j) - a_t)^2, W_{nt} = (x_n(i,j) - a_{nt})^2 \tag{4}$$

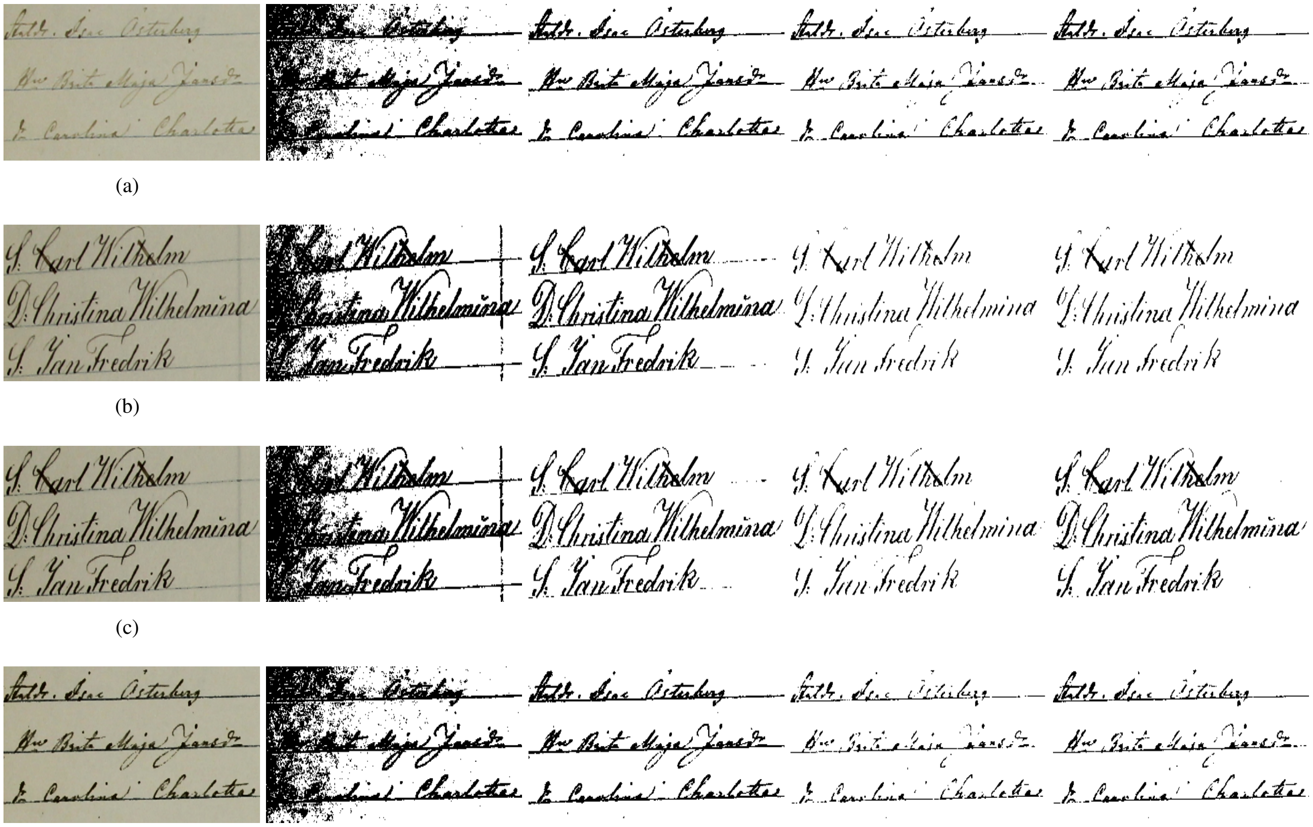
where $x_n(i,j)$ is the normalized pixel value of X_n at the pixel coordinates i and j. Using equations 3 and 4, the cluster whose pixels have lower average value in the normalized image is assigned as the a_t class, and the other cluster is assigned as a_{nt} class. Note that, a_t cluster is assigned as ”0” pixel value which indicates that the corresponding pixel location involves a text pixel value and a_{nt} cluster is assigned as ”1” pixel value which indicates that the corresponding pixel location involves a non-text pixel value.

3. EXPERIMENTAL RESULTS

To assess the qualitative and quantitative performance, proposed method was compared with the other text binarization methods which are OTSU [5] and local based thresholding approach [7]. In the first experiment, we use two different handwritten document images which are good and bad quality of images, shown in Fig. 4(a) and (b), respectively. As

Method	Good			Bad			Enhanced Good			Enhanced Bad		
	P_{FA}	P_{MD}	P_{TE}	P_{FA}	P_{MD}	P_{TE}	P_{FA}	P_{MD}	P_{TE}	P_{FA}	P_{MD}	P_{TE}
Proposed Method	12.6	17	29.6	26.2	24.5	50.7	10.3	4.2	14.5	11.2	2.1	13.3
Otsu [5]	23.4	40.3	63.6	42.2	38.2	80.4	37.2	13.2	50.4	22	21.8	43.8
Niblack [7]	35.6	23	68.6	33.3	24.6	57.9	22.6	10.3	32.9	17.2	12.3	29.5
Avg.	23.86	26.86	53.93	33.9	29.1	63	23.36	9.23	32.6	16.8	12.06	28.86

Table 1. Quantitative measures on different test images, shown in Fig. 4.



(a)

(b)

(c)

(d) (e) (f) (g) (h)

Fig. 4. Qualitative handwritten binarization results by using different handwritten binarization methods on handwritten test images, (a) Original good quality of handwritten image, (b) Original bad quality of handwritten image, (c) Contrast enhancement of good quality of handwritten image, (d) Contrast enhancement of bad quality of handwritten image, (e) Proposed method by using 10 iterations, (f) Proposed method by using 1000 iterations, (g) Global thresholding based binarization method [5], (h) Local thresholding based binarization method [7].

shown in Fig. 4(c) and (d), the input images are improved by using the contrast enhancement method. The binary images in the second and third column of Fig. 4 are the resulting images by using the proposed unsupervised method with 10 and 1000 iterations in k-means clustering, respectively. Fig. 4(g) shows the result by the OTSU text binarization method [5], where black pixels denote the text pixels and white pixels denote the non-text pixels. Results by using Niblack method [7] are given in Fig. 4(h).

Compared with the result in [5], we can see that most text pixels are detected by using the proposed method. For example, for the image in the third row and third column in Fig. 4, almost all text regions are correctly detected as the text pixels from the input image by using the proposed method, while the most text pixels are missed by the method in [5]. The other results of the Otsu thresholding method [5] are shown in Fig. 4, which provide the worst performance comparing to the results of the proposed method and local thresholding method. Generally, the proposed method provides the best performance comparing to the other text binarization methods because it improves the handwritten text by using the contrast enhancement method and removes artifacts on the document image while other methods suffer from these artifacts, and global and local-based thresholding methods are unable to provide accurate results in terms of finding text pixels on the document images.

3.1. Quantitative Results

In order to understand and analyze the performance of the proposed handwritten binarization method, quantitative experiments have been used on different handwritten document images. We evaluated the results by applying quantitative experiments on the input test images together with their ground truth handwritten binary images. Once the binary handwritten binary image has been obtained by using the proposed method, the quantities which are false alarm rate (PFA), missed detection rate (PMD) and total error rate (PTE) are used to obtain the results to compare between the estimated binary handwritten binary image and the ground truth handwritten binary image. The metrics are mathematically defined as follows: $PFA = FA/Nn \times 100$ and $PMD = MD/Nt \times 100$, in which FA is the number of the non-text pixels that are incorrectly obtained as the text pixels, and MD estimates the number of text pixels that are incorrectly detected as the non-text pixels. The sum of both quantities forms TE such that $PTE = PFA + PMD$ with Nn and Nt denoting the total number of non-text and text pixels in the ground-truth binary image, respectively.

Table 1 illustrates the quantitative results for different handwritten document images, shown in Fig. 4, and they are computed using quantitative metrics. Based on the results, we can see that the average value of four PTE values for the results in [7] is 47.23%. For the results of the OTSU thresholding method in [5], the average value of four PTE values is 59.55%. The proposed method provides better results in terms of finding text and non-text pixels in different document images with the average value of four PFA and four PMA are 15.07% and 11.95%, respectively. Besides this, the average value of four PTE values for the proposed method is 27.02%. As a result, the lowest total error PTE is estimated by using the proposed method and the highest accuracy rate of text and non-text detection is achieved by using the proposed method.

In the second experiment, implemented methods were applied to ten different handwritten document images to obtain the binary text images. Table 2 illustrates the average computative results of ten different binary text images by using the quantitative experiments. Based on the results, proposed method is the best performing approach of the comparison, with the average false alarm rate PFA, missed alarm rate PMD and total error rate PTE of 14.2, 13.5, and 27.7, respectively. According to the results, it is clearly seen that the lowest quantitative results are estimated by using the OTSU thresholding method [5]. Consequently, the highest correctly detection rate of text pixels is achieved by using the proposed method.

Method	P_{FA}	P_{MD}	P_{TE}
Proposed	14.2	13.5	27.7
Otsu [5]	23.2	36.9	60.1
Niblack [7]	21.4	26.1	47.5

Table 2. Quantitative measures on ten different test images.

4. CONCLUSION

In this paper, we have presented a new unsupervised method to find text and non-text pixels over the handwritten documents. Our algorithm consists of three main steps. First, a preprocessing method is applied to the input document images to enhance the contrast of the text, to remove the noise and normalize the pixel intensities. The final step is used as an unsupervised method to detect the text pixels and non-text pixels from the normalized pixel intensities. By running the k-means clustering approach as a post-processing of the method, text pixels and non-text pixels from the normalized pixel intensities are partitioned into two different clusters. Finally, handwritten binary image is achieved by assigning the text pixels as black pixel values and non-text pixels as white pixel values. The proposed method presented in this paper finds the text pixels effectively and qualitative and quantitative tests on different data sets show that our method remarkably reduce the detection error rate comparing to the state-of-the-art text binarization methods.

5. ACKNOWLEDGEMENT

This work is part of the research project ”Scalable resource-efficient systems for big data analytics” funded by the Knowledge Foundation (grant: 20140032) in Sweden.

6. REFERENCES

[1] M.T.M. Gary, J.C.H. Poon, *”A fuzzy-attributed graph approach to handwritten character recognition”*, in Proc. IEEE Int. Conf. on Fuzzy System, pp. 570-575, 1993.

[2] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk and B. Girod, *”Robust Text Detection in Natural Images With Edge-Enhanced Maximally Stable Extremal Regions”*, in Proc. IEEE Int. Conf. on Image Processing (ICIP), pp. 2609-2612, 2011.

[3] R. F. Moghaddam, M. Cheriet, *”A variational approach to degraded document enhancement”*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 8, pp. 1347-1361, 2010.

[4] B. Z. Varghahan, M. C. Amirani and S. Mihandoost, *”Enhancement and Cleaning of Handwritten Data*

by using Neural Networks and Threshold Technical", in Proc. IEEE Int. Conf. on App. of Inf. and Comm. Tech., pp. 1-4, 2011.

[5] N. Otsu, *"A threshold selection method from grey-level histograms"*, IEEE Transaction on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

[6] R. F. Moghaddam and M. Cheriet, *"AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization"*, Pattern Recognition, vol. 45, no. 6, pp. 24192431, 2012.

[7] W. Niblack, *"An introduction to digital image processing"*, Englewood Cliffs, Prentice Hall, N. J., pp. 115-116, 1986.

[8] J. Sauvola and M. Pietikainen, *"Adaptive Document Image Binarization"*, Pattern Recognition, vol. 33, no. 2, pp. 225–236, 2000.

[9] B. Gatos, K. Ntrirogiannis, I. Pratikasis, *"Dibco 2009: document image binarization contest"*, International Journal on Document Analysiz and Recognition, pp. 1-10, 2010.

[10] B. Su, S. Lu, C.L. Tan, *"A self-training learning document binarization framework"*, in Proc. IEEE International Conference on Pattern Recognition, pp. 31873190, 2010.

[11] Y. Chen, G. Leedham, *"Decompose algorithm for thresholding degraded historical document images"*, IEE Proceedings Vision, Image and Signal Processing, vol. 152, no. 6, pp. 702-714, 2005.

[12] H. S. Don, *"A noise attribute thresholding method for document image binarization"*, Internatinoal Journal on Document Analysis and Recognition, vol. 4, no. 2, pp. 131-138, 2001.

[13] M. L. Feng, Y. P. Tan, *"Contrast adaotive binarization of low quality document images"*, IEICE Electronics Express, vol. 1, no. 16, pp. 501-506, 2004.

[14] B. Gatos, I. Pratikasis, S. J. Perantonis, *"Adaptive degraded document image binarization"*, Pattern Recognition, vol. 39, no. 3, pp. 317-327, 2006.

[15] T. Arici, S. Dikbas, Y. Altunbasak, *"A histogram modification framework and its application for image contrast enhancement"*, IEEE Trans. Image Process., vol. 18, no. 9, pp. 19211935, 2009.

[16] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, *"An Efficient k-Means Clustering Algorithm: Analysis and Implementation"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, 2002.