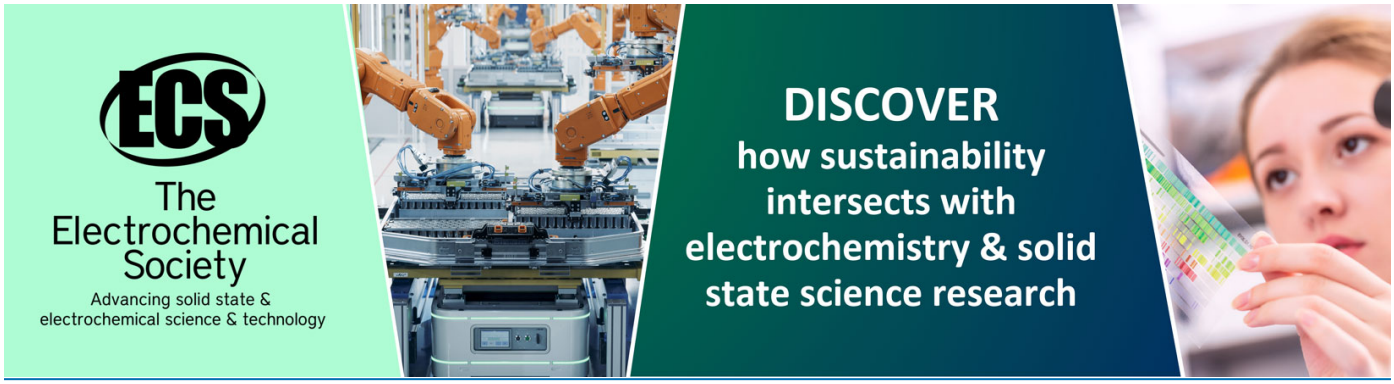
Binarization of Document Image Using Optimum Threshold Modification



IOP Conf. Series: Journal of Physics: Conf. Series **1019** (2018) 012022 doi:10.1088/1742-6596/1019/1/012022

Binarization of Document Image Using Optimum Threshold Modification

Wan Azani Mustafa[1*], Mohamed Mydin M. Abdul Kader[1]
[1]Faculty of Engineering Technology, Universiti Malaysia Perlis, UniCITI Alam Campus, Sungai Chuchuh, 02100 Padang Besar, Perlis, Malaysia

wanazani@unimap.edu.my

Abstract. Document image binarization is one important processing step, especially for data analysis. In this paper, a new binarization based local thresholding technique 'WAN' was presented. The proposed algorithm is known as 'WAN' after the first name of the author in this paper. WAN has been inspired from the Sauvola's binarization method and exhibits its robustness and effectiveness when evaluated on low quality document images. The objective of the WAN method is to improve the Sauvola method and achieve a better binarization results, specifically, for non-uniform document images. The results of the numerical simulation indicate that the WAN method is the most effective and efficient (f-measure 72.274 and NRM = 0.093) compared to the Sauvola method, Local Adaptive method, Niblack method, Feng Method, and Bernsen method.

1. Introduction

There are many challenges addressed in handwritten document image binarization, such as faint characters, bleed-through and large background ink stains [1]–[3]. Document image binarization is the process that segments the document image into the text and background by removing any existing degradations [1], [4]. Recently, many document image binarization methods have been proposed in the literature [5]–[8]. However, selecting the most optimum threshold for binarization is a difficult task due to the presence of a variety of degradations in document images [9]–[10].

Previous studies concentrated on proposing a new method or algorithm to solve the degradation of document images. In 2008, Nikolaos and Dimitrios reviewed a few enhancement and binarization techniques to find the best approach for the future research [11]. They summarized that combination of pre-processing and binarization algorithm able to improve and finally provide the new method. Many researchers agree that it's very difficult to propose a perfect algorithm since the document image in badly condition dealing with many information such as text and structure [11]–[13]. Gatos *et al.* [12] discussed the challenges and strategies to improve the document image binarization based on a combination of Multiple Binarization Techniques and Adapted Edge Information. This approach has a number of advantages: firstly, (i) combining the binarization results of several state-of-the-art methodologies; (ii) incorporating the edge map of the grey scale image; and (iii) applying efficient image post-processing based on mathematical morphology for the enhancement of the final result. Research finding by Shijian *et al.* [14] also points towards the edge information to propose a new method. However, they concentrated on the surface and stroke edge. The result is more effective compared to the Gatos method [12], Sauvola

method [15] and Otsu method [16]. In 2009, Reza and Mohamed published a paper in which they described the new model of a low quality document image using virtual diffusion processes [17]. This technique focuses on the shadow- through and bleed-through problem.
Besides, Laurence *et al.* [18] mentioned the significant relationship between Total Variation regularization and Non-local Means filtering in order to binarize a degrade document image. This approach consists two steps; (1) in order to reduce the effect of background noise, they proposed to apply the Total Variation framework and the result considered as a mask image. (2) The Non-local means was performed to clean the image from noise and bleed-through. However, Shi *et al.* [19] found differences suggesting that the technique based on shape feature is more effective compared to the technique based on background normalization [18], [20]. The threshold selection is based on the stroke width consistency and the result performance is superior compared to the existing binarization technique such as Otsu method [16], Niblack method [21], and Multistage Adaptive Thresholding (MAT) method [22].
Several studies investigating document binarization based on Otsu modification [23]–[25]. Starting 2010, Nina *et al.* [23] mentioned the significant combination between the recursive extension of Otsu thresholding and selective bilateral filtering of scanned handwritten images. This approach considered background estimation before applying the post- processing stage [23], [25]. The above findings contradict the study by Zhang and Wu [24]. They examined the modification algorithm based on the Adaptive Otsu method. The proposed method is based on three main steps; (1) applied the Wiener filter in order to eliminate noise, (2) improved adaptive Otsu's method, and (3) dilation and erosion operators were performed to preserve stroke connectivity and fill possible breaks, gaps, and holes. The advantage of this approach is that faster processing time compared to Recursive Otsu Thresholding Method [23] and AdOtsu method [25]. Similarly, Reza and Mohamed also proposed a new novel method based on the Otsu modification known as AdOtsu method [25]. The main idea of this technique was considered parameterless behavior such as average stroke width and the average line height. A positive result was achieved compared to Sauvola method [15], Otsu Method [16], and Lu and Tan method [14].
In this paper, a new binarization method based on the maximum threshold was discussed. The proposed method inspired from Sauvola method and is known as 'WAN' method. The proposed method experiments on 14 non-uniform document images. A few image quality assessment such as f-measure, sensitivity, NRM (Negative Rate Metric), and Peak Signal Noise Ratio (PSNR) was performed in order to compare the effectiveness the method. Summary, this paper is organized in the following sections: Section 2.0 describes the proposed binarization methods. Section 3.0 presents the analysis of results and Section 4.0 explains the conclusion of this work.

2. Proposed Approach

The proposed algorithm is inspired by the Sauvola method. In this paper, we put forward our proposition of calculating the binarization threshold which is likely to work better for many (if not all) types of degraded and noisy documents. The Sauvola method able to solve the problem of black noise depending on the impact on the standard deviation value by using a range of grey level values in the images [15], [26]–[27]. However, the Sauvola method failed to segment if the contrast between the foreground and background is small or if the text is in thin pen stroke text. So, the WAN method was proposed to overcome this problem by obtaining the maximum threshold value. The main advantage of the proposed method over Sauvola is that it considerably improves binarization for "lost" detail images by shifting up the binarization threshold. The Sauvola algorithm is denoted as follows;

$$T = m\left(1 - k\left(1 - \frac{\sigma}{R}\right)\right) \tag{1}$$

Where, R is the grey level (128), m is the mean value, σ is the standard deviation, and k was set is 0.2 (default value). This method outperforms Niblack algorithm in images where the text pixels have near 0

grey value and the background pixels have near 255 grey values. However, in images where the grey values of text and non-text pixels are close to each other, the results degrade significantly. Based on research, the mean value m will give a high effect on the threshold value as shown in figure 1. According to figure 1, example, if the mean value less than average (m -20) the result was blurred and more information details were lost. Otherwise, if the mean value more than the average (m +20) the result is better and improved. In this paper, the specific value was found to replace the normal mean value. The aim is to find the specific value more than the normal average.

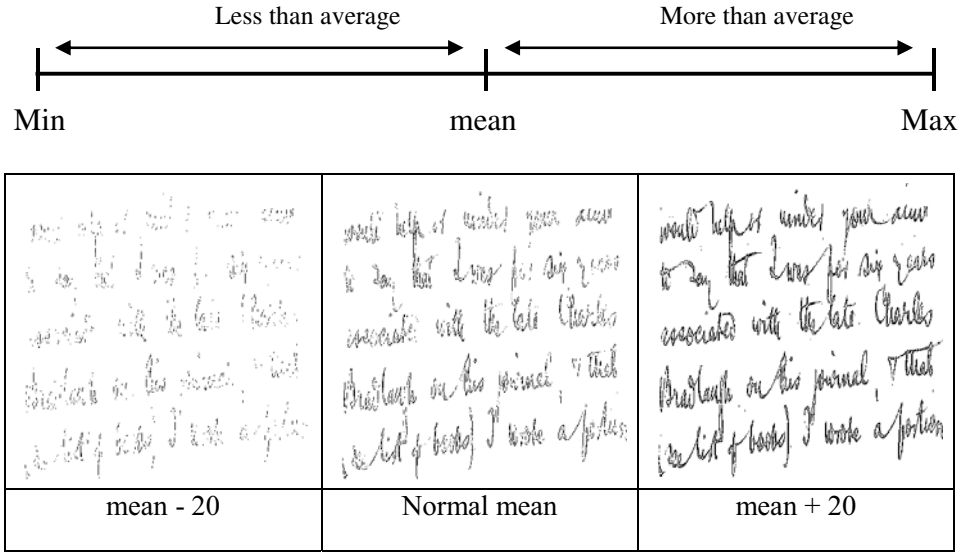


Figure 1. Binarization effect after applying the different mean value.

Actually, the Sauvola method failed to binarize the low contrast region because the threshold value is low. Therefore, the proposed method tends to increase the threshold value to segment the information in the low contrast region. However, if the threshold value is higher, it's will introduce noise and artefact on the resulting image. So the specific and the maximum threshold value needs to be proposed. In this paper, the maximum mean is calculated in order to replace the original mean. The maximum mean equation is depicted as follows;

$$m_{\max} = \frac{\max(x, y) + mean}{2} \tag{2}$$

Where, $\max(x, y)$ is the maximum intensity of input image and $mean$ is the original mean for the whole image. The average between the highest intensity and mean image was calculated. The main target is to improve the lost details on binarization result and at the same time to reduce the noise and any artefact. The final proposed algorithm is;

$$T = \frac{\max(x, y) + mean}{2}\left[1 - k\left(1 - \frac{\sigma}{R}\right)\right] \tag{3}$$

where, k and R value used a default value form Sauvola method. From this algorithm, the low contrast

problem can be solved and automatically increased the binarization result. In order to evaluate the proposed method and compare the results with a few local methods, 14 document images were tested and the results are given in the following section.

3. Experimental Result

In this paper, 14 document images from Handwritten Document Image Binarization Contest (H-DIBCO) [28] dataset experimented. The images contain various degradations such as shadows, non-uniform illumination, stains, smudges, bleed-through and faint characters [1]. All the processed images are in greyscale images and the size of each image is 400 × 400 pixels, 72 dpi, and 8-bit depth. All the programs were written in MATLAB from an Asus laptop with AMD Athlon™ II P320 Dual-Core Processor 2.10GHz and 3.00GB RAM. Three examples of document images were illustrated in figure 2. The first row shows the original image with degradation and shadowing problem and follows by binarization methods from Sauvola method and lastly by the proposed method (WAN). Based on visual criteria, the proposed algorithm seems to outperform the other methods with respect to image quality and preservation of meaningful textual information. Besides, the Sauvola result shows poor binarization compared to the proposed method. The problem of Sauvola method already overcome using the proposed algorithm.

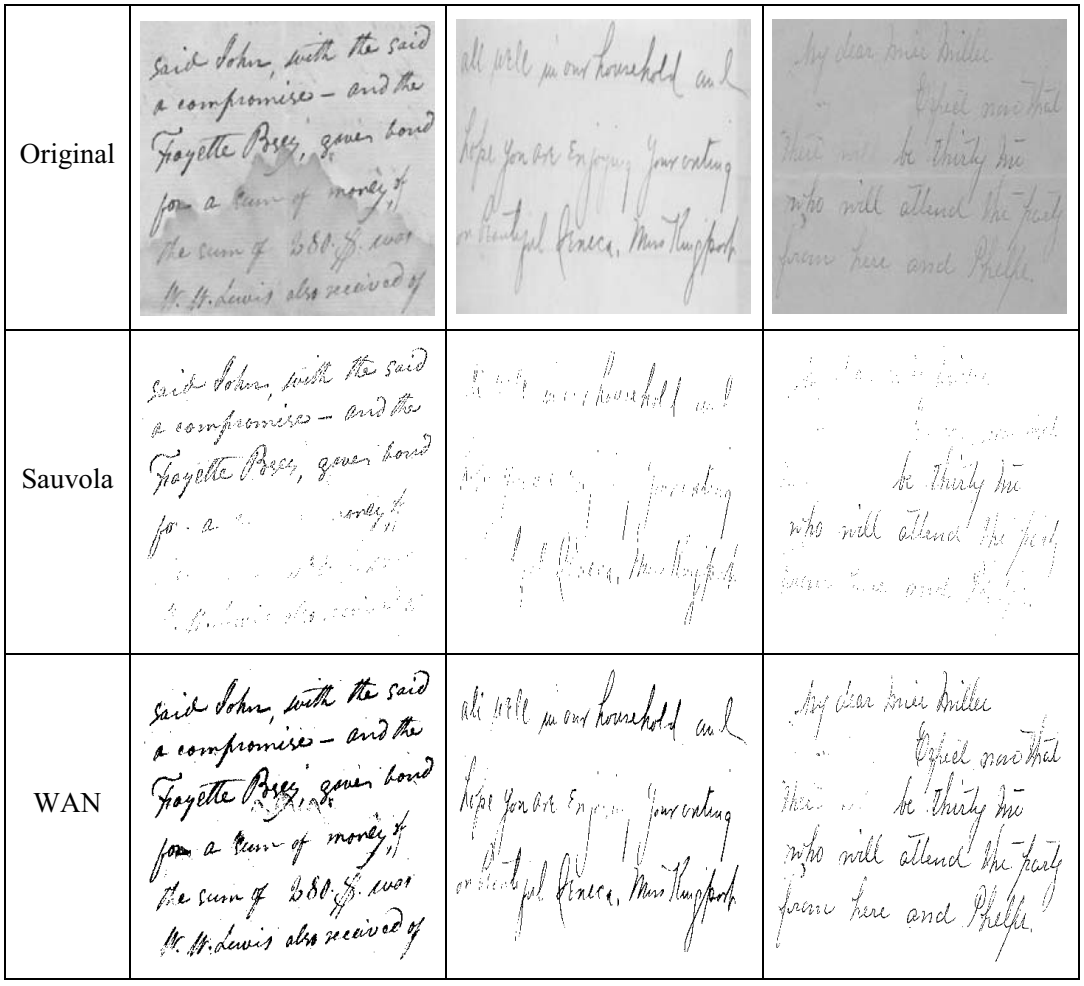


Figure 2. Comparison resulting image between Sauvola method and WAN method.

Next, a few image quality assessment (IQA) was calculated to compare with the Sauvola method and proposed method. In this paper, the evaluation based on the F-measure, Sensitivity, NRM (Negative Rate Metric), and PSNR (Peak Signal to Noise Ratio) was obtained. All the assessment equation can be referred in H-DIBCO [29]–[30]. The highest F-measure, Sensitivity, and PSNR while lowest NRM value represents a good binarization result. Table 1 provides the comparison result between Sauvola method and the proposed method (WAN). Average, in term of f-measure and sensitivity, the proposed method achieved the highest result (f-measure = 72.274 and sensitivity = 87.676) compared to the Sauvola method (f-measure = 56.321 and sensitivity = 45.343). Besides, a comparison based on NRM again shows that the performance of the proposed method obtained 0.093, which is lower than the Sauvola methods. However, based on PSNR the proposed method slightly lower (13.614) compared to the Sauvola method (14.612).

Table 1: The comparison based on a few objective measurements on Sauvola method and WAN method.

	Sauvola				**WAN**			
	F-measure	Sensitivity	NRM	PSNR	F-measure	Sensitivity	NRM	PSNR
H1	83.153	74.797	0.128	17.631	76.402	94.907	0.042	14.765
H2	47.349	31.301	0.344	12.281	83.990	86.119	0.078	15.545
H3	8.148	4.251	0.479	12.145	88.065	86.392	0.071	18.266
H4	84.884	75.039	0.125	18.871	86.658	94.114	0.035	18.519
H5	82.120	75.949	0.123	17.065	34.259	99.138	0.130	6.262
H6	77.328	66.711	0.168	15.906	52.439	98.400	0.070	9.314
H7	52.074	36.028	0.321	13.581	76.918	71.639	0.147	15.463
H8	85.501	78.754	0.108	17.225	69.182	99.462	0.036	12.017
H9	52.726	36.014	0.320	11.310	90.638	90.816	0.052	16.678
H10	51.646	35.338	0.324	11.799	52.334	94.531	0.124	7.471
H11	20.827	11.676	0.442	11.171	56.766	92.480	0.100	9.166
H12	31.078	18.442	0.408	13.285	74.132	61.438	0.194	16.091
H13	29.484	17.560	0.413	13.918	70.233	60.740	0.199	16.043
H14	82.176	72.942	0.136	18.375	73.823	97.293	0.029	14.989
Average	**56.321**	**45.343**	**0.274**	**14.612**	**72.274**	**87.676**	**0.093**	**13.614**

Besides that, in order to prove the improvement of binarization, the increment percentages was calculated. The result of increment based on f-measure, sensitivity, NRM, and PSNR was illustrated in figure 3.

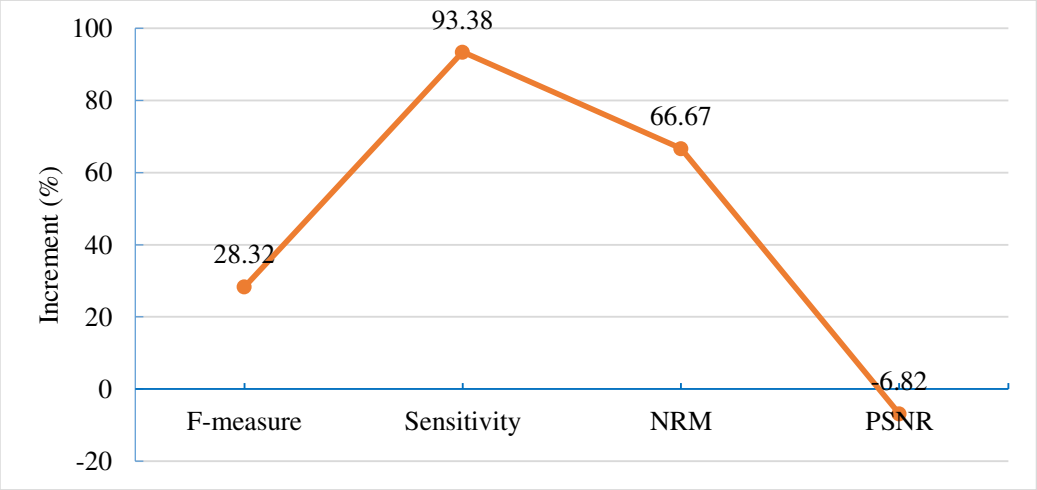


Figure 3: The increment after applying the WAN method.

The histogram in figure 3 indicates the increment (%) of binarization after employing the proposed method. Based on figure 3, the evaluation result based on sensitivity shows the highest improvement which is 93.38% and f-measure obtained lower increment which is 28.32% compared to the Sauvola method. However, in term of PSNR, the proposed method slightly decreased (6.82%) compared to the Sauvola method.

4. Conclusion

Document binarization is an important application in vision processing. In this paper, a new binarization method was proposed for low quality document image known as 'WAN' method. The aim of this paper is to improve the Sauvola method and achieved a better binarization compared to a few existing binarization methods. The proposed method performs better than the contemporary methods, especially when the input image has very few or no text (white image) and also when the intensity variations between the text and background are extremely low. Based on the result performance, the proposed method achieved a good result in term of F-measure, Sensitivity and NRM compared to the original Sauvola method. In the near future, we will propose a new algorithm which will use the more reliable methodology to enhance the work.

5. Reference

[1] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A combined approach for the binarization of handwritten document images," *Pattern Recognit. Lett.*, vol. 35, pp. 3–15, Jan. 2014.

[2] B. M. Singh, R. Sharma, D. Ghosh, and A. Mittal, "Adaptive binarization of severely degraded and non-uniformly illuminated documents," *Int. J. Doc. Anal. Recognit.*, pp. 393–412, 2014.

[3] D. Rivest-Hénault, R. Farrahi Moghaddam, and M. Cheriet, "A local linear level set method for the binarization of degraded historical document images," *Int. J. Doc. Anal. Recognit.*, vol. 15, no. 2, pp. 101–124, 2012.

[4] W. A. Mustafa and H. Yazid, "Illumination and Contrast Correction Strategy using Bilateral Filtering and Binarization Comparison," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 1, pp. 67–73, 2016.

[5] P. K. More and D. D. Dighe, "A Review on Document Image Binarization Technique for Degraded Document Images," *Int. Res. J. Eng. Technol.*, pp. 1132–1138, 2016.

[6] X. Chen, L. Lin, and Y. Gao, "Parallel nonparametric binarization for degraded document images," *Neurocomputing*, vol. 189, pp. 43–52, 2016.
[7] G. Vara Lakshmi & P. Kamala, "Improving Degraded Document Images Using Binarization Techniques," *Int. J. Electron. Commun. Eng.* , vol. 5, no. 2, pp. 1–8, 2016.
[8] M. Soua, R. Kachouri, and M. Akil, "Improved Hybrid Binarization based on Kmeans for Heterogeneous document processing," in *9th International Symposium on Image and Signal Processing and Analysis, ISPA 2015*, 2015, pp. 210–215.
[9] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Third. Upper Saddle River, NJ, USA: Prentice Hall, 2008.
[10] H. Tanaka, "Threshold correction of document image binarization for ruled-line extraction," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009, pp. 541–545.
[11] N. Ntogas and V. Dimitrios, "A Binarization Algorithm for Historical Manuscripts," in *International Conference on Comunications*, 2008, pp. 41–51.
[12] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," *19th Int. Conf. Pattern Recognit.*, pp. 1–4, Dec. 2008.
[13] P. Vinicius, K. Borges, J. Mayer, and E. Izquierdo, "Document Image Processing for Paper Side Communications," in *IEEE Transactions On Multimedia*, 2008, vol. 10, no. 7, pp. 1277–1287.
[14] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *Int. J. Doc. Anal. Recognit.*, vol. 13, no. 4, pp. 303–314, Oct. 2010.
[15] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33. pp. 225–236, 2000.
[16] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," in *IEEE Transactions On Systrems, Man, And Cybernetics*, 1979, vol. 20, no. 1, pp. 62–66.
[17] R. F. Moghaddam and M. Cheriet, "Low quality document image modeling and enhancement," *Int. J. Doc. Anal. Recognit.*, vol. 11, no. 4, pp. 183–201, Feb. 2009.
[18] L. Likforman-Sulem, J. Darbon, and E. H. B. Smith, "Enhancement of historical printed document images by combining Total Variation regularization and Non-local Means filtering," *Image Vis. Comput.*, vol. 29, no. 5, pp. 351–363, Apr. 2011.
[19] J. Shi, N. Ray, and H. Zhang, "Shape based local thresholding for binarization of document images," *Pattern Recognit. Lett.*, vol. 33, no. 1, pp. 24–32, Jan. 2012.
[20] W. A. Mustafa, H. Yazid, and S. Yaacob, "Illumination Normalization of Non-Uniform Images Based on Double Mean Filtering," in *IEEE International Conference on Control Systems, Computing and Engineering*, 2014, pp. 366–371.
[21] W.Niblack, *An Introduction to Digital Image Processing.* Englewood Cliffs: Prentice-Hall, 1986.
[22] F. Yan, H. Zhang, and C. R. Kube, "A multistage adaptive thresholding method," *Pattern Recognit. Lett.*, vol. 26, pp. 1183–1191, 2005.
[23] O. Nina, B. Morse, and W. Barrett, "A Recursive Otsu Thresholding Method for Scanned Document Binarization," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2010, pp. 307–314.
[24] Y. Zhang and L. Wu, "Fast Document Image Binarization Based on an Improved Adaptive Otsu ' s Method and Destination Word Accumulation," *J. Comput. Inf. Syst.*, vol. 6, no. 7, pp. 1886–1892, 2011.
[25] Reza Farrahi Moghaddamn Mohamed Cheriet, "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization," *Pattern Recognit.*, vol. 45, pp. 2419–2431, Jun. 2012.
[26] B. Bataineh, S. N. H. S. Abdullah, and K. Omar, "An adaptive local binarization method for document images based on a novel thresholding method and dynamic windows," *Pattern Recognit.*

Lett., vol. 32, pp. 1805–1813, 2011.

[27] J. Sauvola, T. Seppanen, S. Haapakoski, M. Pietikiiinen, M. Vision, M. P. Group, and I. Oulu, "Adaptive Document Binarization," in *International Conference on Document Analysis and Recognition*, 1997, pp. 147–152.

[28] K. Zagoris, "H-DIBCO '16," 2016. [Online]. Available: http://vc.ee.duth.gr/h-dibco2016/. [Accessed: 01-Jan-2016].

[29] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," *2009 10th Int. Conf. Doc. Anal. Recognit.*, no. Dibco, pp. 1375–1382, 2009.

[30] W. A. Mustafa and H. Yazid, "Background Correction using Average Filtering and Gradient Based Thresholding," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 5, pp. 81–88, 2016.