Document Images Binarization Using Hybrid Combination of Fuzzy C-Means and Deghost Method

W A Mustafa1*, Hairy Aziz1, Wan Khairunizam2, I Zunaidi3, Z M Razlan2 and Shahriman A B2

1Department of Electrical Technology Engineering, Faculty of Engineering Technology, Universiti Malaysia Perlis, UniCITI Alam Campus, Sungai Chuchuh, 02100 Padang Besar, Perlis, Malaysia
2School of Mechatronics Engineering, Universiti Malaysia Perlis, 02600 Arau, Malaysia
3Faculty of Technology, University of Sunderland, St Peter's Campus, Sunderland, SR6 0DD, United Kingdom

*E-mail:wanazani@unimap.edu.my

Abstract. This paper presents a document binarization approach to the document image. Thousands of historical documents usually hold important information within them. It is usually stored in the national archives and library around the globe waiting to be scanned to retrieve the content it holds. However, many environmental factors, improper handling, and the poor quality of the materials used in the document creation cause it to suffer a high degree of degradation of various types. In order to retrieve the content of the degraded historical document, a binarization approach on document images must be applied. The processes of the binarization separate pixel value of an input image into two values which is white as the background and black as foreground text. The proposed system consists of three parts. In the first part, the image pre-processing operation is done before the binarization process to enhance image quality. In this part, Contrast Stretching and Mean Filter is applied onto the image to remove noise on the image. The second part will be to apply the binarization algorithm on the document image that has undergone an image pre-processing operation. By applying the Fuzzy C-Means algorithm to the document images, the images will be converted to a binary image and divided into two components, which is text and background. The last step of the proposed method will be performing the Deghost operation to remove “ghost” entities that may have appeared on the document image. The method will undergo imaging quality analysis, such as PSNR, Accuracy, and F-measure to determine the effectiveness of the proposed method. The experimental results on H-DIBCO 2013 dataset show the robustness, reliability, and efficiency in the proposed approach.

1. Introduction

Thousands of historical document images usually hold important information within them. Document images are usually badly degraded and suffer from degradation as it began to age over time. Some type of degradation that the document image suffers is bleed through, smear and uneven illumination image contrast variation [1–6]. It is important to retrieve text from a document image as it may hold important information. The major problem faced is degradation, when it occurs, the document image is in a very poor quality which in most cases is hard to decipher [7–11].

In these past years, an increasing amount of literature on discussing the automated image processing method that has been performed solely purpose document images has shown an increasing pattern [12,13]. The proposed image processing approaches are basically based on a new algorithm and methodology [14,15]. Mostly, researchers have agreed that the difficulties of document image binarization are badly degradation of the old and historical manuscript of historical document images [16–21]. In 2014, an improved degraded document image binarization using a guided image filter was proposed by Kaur et al [22]. The proposed method provides in-depth analysis of the work of a new
technique which has the ability to binarized documents in a more efficient manner. It consists of three parts, the first part will be responsible for applying the guided image filter to smooth and restore a degraded document which will improve the visibility of the degraded image. Now second part comes into action, in this part adaptive image contrast enhancement are going to be applied. At last, the third part is under action which can do the true binarization. The contrast map is going to be binarized using adaptive thresholding thus it will be integrated with a Hybrid edge map to recognize the text stroke edge pixels. The comparative analysis has shown that the proposed algorithm provides quite a significant improvement over the available algorithms. Image Quality Analysis (IQA) acquired the average of 0.9896 for Accuracy, 95.84% for F-measure, and 22.89 for PSNR.

Niortogiannis et al. [23] have done a combined approach for the binarization of handwritten document images. The research was inspired by the low detection rate of faint characters in binarization of handwritten document images. Initially, background estimation is applied along with image normalization based on background compensation. Afterward, global binarization is performed on the normalized image. In the binarized images, very small components are discarded and representative characteristics of a document image such as the stroke width and the contrast are computed. Furthermore, local adaptive binarization is performed on the normalized image taking into account the aforementioned characteristics. Finally, the two binarization outputs are combined at connecting component level. For DIBCO 2011 dataset, the average F-measure and PSNR are 94.05% and 21.65 respectively.

In the year 2017, Peng et al. [24] have reported Document Image binarization using Convolutional Encoder-Decoder. In the proposed method, mid-level document image representations are learned by a stack of convolutional layers that compose the encoder in this architecture. Then, by mapping low-resolution representations of the original size through the decoder, the binarization image is obtained, which is then composed by a series of transposed convolutional layers. The method compares the proposed method with other binarization algorithms both qualitatively and quantitatively on the public dataset. The experimental results show that the proposed method has comparable performance to the other hand-crafted binarization approaches and has more generalization capabilities with limited in-domain training data. For H-DIBCO 2016 dataset, the average F-measure and PSNR are 88.07% and 18.13 respectively.

Document image binarization using a discriminative structural classifier was written by author Ahmadi et al. [25]. The study done was to propose a statistical framework for binarizing degraded document images based on the concept of conditional random fields (CRFs). The CRFs are discriminative graphical models which display conditional distributions that are used in structural classifications. The proposed graphical framework enables the depending labeling of all the sites of the image despite the independent pixel-by-pixel binarization of other methods. The performance of the proposed system is evaluated on different document image datasets and is compared with several well-known binarization methods. Experimental results show comparable performance with respect to other state-of-the-art methods. By using DIBCO 2009 dataset, the average F-measure and PSNR are 91.12% and 18.63 respectively.

In this paper, a binarization method based on Fuzzy C-Means and Deghost technique is proposed. The proposed method came from a combination of the Fuzzy C-Means clustering algorithm and Deghost method. In order to prove the effectiveness of the proposed method, three image quality assessment (IQA) which are Accuracy, F-measure, and Peak Signal Noise Ratio (PSNR) have been calculated. This paper is organized into four sections. Section 2 and 3 explains the Fuzzy C-Means and Deghost method, Section 4 shows the results performance in terms of the IQA. Section 5 describes the summary of the work.

2. Fuzzy C-Means Clustering Algorithm

The first step is applying Fuzzy C-means algorithm into the processed image. The process is needed to separate text from the background. The standard FCM objective function for partitioning \( \{x_k\}_{k=1}^{N} \) into \( c \) clusters is given by Equation [26].
\[ J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^p \| x_k - v_i \|^2 \]  \hspace{1cm} (1)

where \( \{ v_i \}_{i=1}^c \) are the prototypes of the clusters and the array \( \{ u_{ik} \} = U \) represents a partition matrix, \( U \in U \), namely in Equation 2 [26].

\[ U \{ u_{ik} \in [0,1] \mid \sum_{i=1}^{c} u_{ik} = 1 \forall k \text{ and } 0 < \sum_{k=1}^{N} u_{ik} < N \forall i \} \]  \hspace{1cm} (2)

The parameter \( p \) is a weighting exponent on each fuzzy membership and determines the amount of fuzziness of the resulting classification. The FCM objective function is minimized when high membership values are assigned to voxels whose intensities are close to the centroid of its particular class, and low membership values are assigned when the voxel data is far from the centroid. The proposed technique, two clusters value has been used.

3. Deghost Method

The next processed is applying the Deghost method to the processed image. The Deghost method removes “ghost” entities in the processed image.

![Figure 1: Cross section through an image, illustrating how “ghost” objects might occur](image)

The main steps of the algorithm:

1. The gradient magnitude image \( G \) of the smoothed image is calculated, using Sobel’s edge operator.
2. A value for Objects having an average gradient below a threshold \( T_p \) is selected. There is no automatic method to specify \( T_p \), so it is specified by trial and error.
3. For all four-connected print components, the average gradient of the edge pixels is calculated.

Based on figure 2 after applying the Deghost method on the image, some unwanted noise after the binarization process is successfully removed or reduced. Figure 2 (top) image shows the resulting after applied segmentation process using FCM and figure 2 (bottom) shows the result of a combination between FCM and enhanced using the Deghost technique.
Figure 2: Image before (top) and after applying the Deghost method (bottom).

4. Results And Discussion
In this study, 16 document images from H-DIBCO 2013 dataset which have non-uniform image background is used as a test subject. The result of document image after the binarization process is shown in Figure 3. According to Figure 3, it is clear that the proposed method is better than the comparative method as it contains less noise and its accuracy is comparable to ground truth. Therefore, it is hard to compare all the method in Figure 3 using only human naked eyes. Thus, Image Quality Analysis (IQA) must be referred to find the most effective one.

| Method      | Image 1 | Image 2 | Image 3 |
|-------------|---------|---------|---------|
| Original    | ![image](image1) | ![image](image2) | ![image](image3) |
| Ground truth| ![image](image1) | ![image](image2) | ![image](image3) |
Figure 3: Image Binarization using H-DIBCO 2013 for selected binarization method
Based on Figure 3, the proposed method images are comparable to the ground truth image as both images are almost identical from the comparison below. The proposed method image also has a low amount of noise compared to another method in Figure 3. It can be stated from proposed method image, that it is better than the other comparison method. In order to assess the overall performance of each selected binarization method, a quantitative analysis is conducted. The performances of different binarization techniques are evaluated by using three objective indices. These indices are F-measure, Accuracy, and PSNR. The pixels similarity of the resultant segmented image and the segmented image that use the selected method determines the quality of the segmented image. All calculation is based on true positive (TP), true negative (TN), false positive (FP), precision (P), recall (R), the maximum pixel value \( (\text{Max}_I) \), mean squared error (MSE), and false negative (FN). The segmentation accuracy can be obtained by calculating the percentage of pixels that are correctly segmented and background in the image. Accuracy score reaches its best value at 100 and worst at 0. The accuracy equation can be referred in Equation 3 [27,28].

\[
\text{Accuracy} = \left( \frac{TP + TN}{TP + TN + FP + FN} \right)
\]  

(3)

The result of IQA for Accuracy is presented in Table 1. Bernsen method gives out the lowest value of 85.68% for Accuracy compare to other methods. This is due to the processed images that still have much noise after the binarization process. However, the proposed method achieves the highest result with 96.06%. From the data in Table 1, it is apparent that all selected method gave the satisfied result.

**Table 1**: Accuracy performance against H-DIBCO 2013 dataset for the selected method

| Image | Bernsen | Triangle | Multiple Threshold | Wolf | Proposed Method |
|-------|---------|----------|-------------------|------|-----------------|
| 1     | 99.17   | 98.58    | 97.91             | 96.29| 98.26           |
| 2     | 93.10   | 96.90    | 97.15             | 93.38| 98.02           |
| 3     | 96.84   | 95.85    | 95.43             | 93.48| 95.72           |
| 4     | 68.33   | 97.73    | 99.23             | 96.97| 99.50           |
| 5     | 70.24   | 77.43    | 98.93             | 97.47| 99.39           |
| 6     | 91.82   | 6.69     | 95.63             | 94.15| 98.09           |
| 7     | 94.36   | 98.47    | 97.95             | 97.73| 97.89           |
| 8     | 68.36   | 94.91    | 98.96             | 95.18| 99.13           |
| 9     | 92.79   | 97.61    | 97.55             | 95.04| 95.87           |
| 10    | 80.79   | 95.53    | 97.82             | 92.41| 98.82           |
| 11    | 96.56   | 98.74    | 98.33             | 94.76| 97.88           |
| 12    | 89.42   | 66.91    | 77.81             | 88.90| 80.09           |
| 13    | 90.13   | 87.41    | 54.51             | 84.23| 93.59           |
| 14    | 89.30   | 84.89    | 59.33             | 90.81| 96.24           |
| 15    | 87.96   | 93.50    | 95.34             | 80.31| 92.80           |
| 16    | 61.77   | 89.21    | 96.16             | 89.47| 95.60           |
| Average | 85.68   | 86.27    | 91.13             | 92.54| **96.06**       |
Besides, F-measure also was calculated based on Equation 4 [27].

\[
F - measure = 2 \times \left( \frac{P \times R}{P+R} \right) \times 100
\]  

(4)

The F-measure score is the harmonic average of the precision and recall, where the F-measure score reaches its best value at 1 and worst at 0. The result of F-Measure should be higher to show a well-segmented image. The result of IQA for F-measure is presented in Table 2. From the data in Table 2, it is apparent that all selected method gave the satisfied result. The proposed method achieves the highest result with 74.96%. This is due to the processed images that have been well binarized. However, the Wolf method gives out the lowest value of 23.32% for the F-Measure compare to other methods.

| Image | Bernsen | Triangle | Multiple Threshold | Wolf | Proposed Method |
|-------|---------|----------|-------------------|------|-----------------|
|       | 1       | 88.53    | 84.33             | 63.50| 9.60            |
|       | 2       | 60.53    | 80.30             | 73.11| 2.54            |
|       | 3       | 75.65    | 75.20             | 47.14| 2.37            |
|       | 4       | 19.44    | 80.65             | 91.28| 53.73           |
|       | 5       | 14.56    | 21.58             | 79.33| 32.32           |
|       | 6       | 56.62    | 12.53             | 51.52| 22.49           |
|       | 7       | 33.61    | 69.86             | 19.16| 2.31            |
|       | 8       | 19.76    | 69.87             | 90.64| 33.83           |
|       | 9       | 52.05    | 79.72             | 68.23| 4.75            |
|       | 10      | 42.59    | 79.95             | 86.10| 26.06           |
|       | 11      | 76.85    | 91.45             | 87.12| 44.22           |
|       | 12      | 64.73    | 42.86             | 51.18| 41.54           |
|       | 13      | 74.91    | 74.49             | 44.70| 42.70           |
|       | 14      | 62.34    | 60.04             | 35.82| 32.03           |
|       | 15      | 73.47    | 86.64             | 87.68| 13.14           |
|       | 16      | 25.82    | 60.44             | 76.03| 9.47            |
| Average| 52.59   | 66.87    | 65.78             | 23.32| **74.96**       |

PSNR is most commonly used to measure the quality of reconstruction of lost compression code. The equation of PSNR as follows in Equation 5 [27].

\[
PSNR = 10 \log_{10} \left( \frac{\text{Max}^2}{\text{MSE}} \right)
\]  

(5)

PSNR is most commonly used to measure the quality of reconstruction of lost compression code. The result of IQA for PSNR is presented in Table 3. Bernsen method gives out the lowest value of 10.13 dB for PSNR out of all the selected methods. This is due to the processed images that have not well reconstructed. However, the proposed method achieves the highest result with 16.03 dB. From the data in Table 3, it can be said that all selected method gave the satisfied result.
Table 3: PSNR performance against H-DIBCO 2013 dataset for the selected method

| Image | Bernsen | Triangle | Multiple Threshold | Wolf | Proposed Method |
|-------|---------|----------|--------------------|------|-----------------|
| 1     | 20.79   | 18.49    | 16.79              | 14.30| 17.60           |
| 2     | 11.61   | 15.09    | 15.46              | 11.79| 17.04           |
| 3     | 15.00   | 13.82    | 13.40              | 11.86| 13.68           |
| 4     | 4.99    | 16.43    | 21.15              | 15.19| 22.98           |
| 5     | 5.26    | 6.46     | 19.72              | 15.97| 22.15           |
| 6     | 10.87   | 0.30     | 13.60              | 12.33| 17.18           |
| 7     | 12.49   | 18.17    | 16.88              | 16.44| 16.75           |
| 8     | 5.00    | 12.93    | 19.83              | 13.17| 20.63           |
| 9     | 11.42   | 16.21    | 16.10              | 13.05| 13.84           |
| 10    | 7.17    | 13.50    | 16.61              | 11.20| 19.68           |
| 11    | 14.63   | 18.99    | 17.77              | 12.81| 16.73           |
| 12    | 9.75    | 4.80     | 6.54               | 9.55 | 7.01            |
| 13    | 10.06   | 9.00     | 3.42               | 8.02 | 11.93           |
| 14    | 9.71    | 8.21     | 3.91               | 10.36| 14.25           |
| 15    | 9.19    | 11.87    | 13.32              | 7.06 | 11.43           |
| 16    | 4.18    | 9.67     | 14.16              | 9.78 | 13.56           |
| Average| 10.13   | 12.12    | 14.29              | 12.05| 16.03           |

5. Conclusion
Today, there is a strong move toward digitalization to preserve the content of the historical document for future generations. The huge amount of digital data produced requires automatic processing, enhancement, and recognition. There is a need for binarization approach on the document image to retrieve information on a document. In this paper, a combination of the Fuzzy C-Means clustering algorithm and Deghost method are used to create a binarization technique. The target is to suggest a maximum threshold value in order to detect the text image. Besides, the goal of this study is to acquire result in term of reducing blur image and eliminate the noise. In order to analyze the result performance, a few IQA such as F-measure, and PSNR was calculated. The proposed method successfully to improve the image quality in term of achieved highest result.

6. References
[1] Mustafa W A and Yazid H 2017 Image Enhancement Technique on Contrast Variation : A Comprehensive Review J. Telecommun. Electron. Comput. Eng. 9 199–204
[2] Pardhi S and Khart G U 2017 An Improved Binarization Method for Degraded Document Int. J. Res. Advent Technol. 1–5
[3] Lu D, Huang X, Liu C, Lin X, Zhang H and Yan J 2016 Binarization of degraded document image based on contrast enhancement Chinese Control Conf. CCC 2016-Augus 4894–9
[4] Mysore S, Gupta M K and Belhe S 2016 Complex and degraded color document image binarization 3rd Int. Conf. Signal Process. Integr. Networks, SPIN 2016 157–62
[5] Mustafa W A and Kader M M M A 2018 Binarization of Document Images: A Comprehensive Review J. Phys. Conf. Ser. 1019 1–9
[6] Mustafa W A and Kader M M M A 2018 Document Image Database ( 2009 - 2012 ): A Systematic Review J. Phys. Conf. Ser. 1019 1–7
[7] Ayyalasomayajula K R and Brun A 2017 Document Binarization Combining with Graph Cuts and Deep Neural Networks 36th Swedish Symposium on Image Analysis pp 1–6
[8] Obafemi-Ajayi T, Agam G and Frieder O 2010 Historical document enhancement using LUT classification Int. J. Doc. Anal. Recognit. 13 1–17
[9] Mustafa W A and Yazid H 2016 Illumination and Contrast Correction Strategy using Bilateral Filtering and Binarization Comparison J. Telecommun. Electron. Comput. Eng. 8 67–73
[10] Kuk J G and Cho N I 2009 Feature Based Binarization of Document Images Degraded by Uneven Light Condition 2009 10th Int. Conf. Doc. Anal. Recognit. 748–52
[11] Kaur J and Mahajan R 2014 A Review of Degraded Document Image Binarization Techniques Int. J. Adv. Res. Comput. Commun. Eng. 3
[12] Mustafa W A, Khairunizam W, Ibrahim Z, Ab S and M.Razlan Z 2018 Improved Feng Binarization Based on Max-Mean Technique on Document Image IEEE International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA) (IEEE) pp 1–6
[13] Mustafa W A, Aziz H, Khairunizam W, Ibrahim Z, Ab S and M.Razlan Z 2018 Review of Different Binarization Approaches on Degraded Document Images IEEE International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA) (IEEE) pp 1–8
[14] Mustafa W A, Yazid H and Jaafar M 2018 An Improved Sauvola Approach on Document Images Binarization J. Telecommun. Electron. Comput. Eng. 10 43–50
[15] Mustafa W A and Kader M M M A 2018 Binarization of Document Image Using Optimum Threshold Modification J. Phys. Conf. Ser. 1019 1–8
[16] Kaviya Selvi K and Sabeenian R S 2015 Restoration of degraded documents using image binarization technique ARPN J. Eng. Appl. Sci. 10 2813–7
[17] Ranganatha D and Holi G 2015 Hybrid binarization technique for degraded document images Souvenir of the 2015 IEEE International Advance Computing Conference, IACC 2015 pp 893–8
[18] Zemouri E T, Chibani Y and Brik Y 2014 Restoration based Contourlet Transform for historical document image binarization Int. Conf. Multimed. Comput. Syst. -Proceedings 0 309–13
[19] Singh B M, Sharma R, Ghosh D and Mittal A 2014 Adaptive binarization of severely degraded and non-uniformly illuminated documents Int. J. Doc. Anal. Recognit. 393–412
[20] Jia F, Shi C, He K, Wang C and Xiao B 2018 Degraded document image binarization using structural symmetry of strokes Pattern Recognit. 74 225–40
[21] Jyoti D, Raj B, Sharma A and Kapoor K 2016 Document Image Binarization Technique for Degraded Document Images By Using Morphological Operators Int. J. Adv. Res. Ideas Innov. Technol. 2 1–7
[22] Kaur E J and Mahajan R 2014 Improved Degraded Document Image Binarization Using Guided Image Filter Int. J. Sci. Res. Educ. 4 242–9
[23] Nitrogiannis K, Gatos B and Pratikakis I 2014 A combined approach for the binarization of handwritten document images Pattern Recognit. Lett. 35 3–15
[24] Peng X, Cao H and Natarajan P 2017 Using Convolutional Encoder-Decoder for Document Image Binarization 2017 14th IAPR Int. Conf. Doc. Anal. Recognit. 708–13
[25] Ahmadi E, Azimifar Z, Shams M, Famouri M and Shafiee M J 2015 Document image binarization using a discriminative structural classifier Pattern Recognit. Lett. 63 36–42
[26] Mesquita R G, Silva R M A, Mello C A B and Miranda P B C 2015 Parameter tuning for document image binarization using a racing algorithm Expert Syst. Appl. 42 2593–603
[27] Lokhande S S and Dawande N A 2015 A survey on document image binarization techniques Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015 pp 742–6
[28] Mustafa W A and Yazid H 2016 Background Correction using Average Filtering and Gradient Based Thresholding J. Telecommun. Electron. Comput. Eng. 8 81–8