International Journal of Electrical, Electronics and Computer Engineering **9**(1&2): 06-11(2020)

ISSN No. (Online): 2277-2626

Image Refinement of Degrade Documents using Image Binarization

Huma Khan and Anas Iqbal

All Saints College of Technology Bhopal (Madhya Pradesh), India.

(Corresponding author: Huma Khan)

(Received 13 April 2020, Accepted 27 July 2020)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: **Binarization is the process converting a multi-tone image into a bi-tonal image. In the case of document images, it is typical to map foreground text pixels to black and the rest of the image (background) to white. In many applications, binarization is a critical preprocessing step and helps facilitate other document processing tasks such as layout analysis and character recognition. In such pipelines, the quality of the binarization can greatly affect system performance, as errors made in the binarization step can propagate to downstream tasks. As a standalone application, binarization can serve as a noise removal process to increase document readability. The file size of binary images is often orders of magnitudes smaller than the original gray or color images, which makes them cheaper to store on disk. Additionally, with the rise of digital archives, file size can become a concern as large numbers of images are viewed over the Internet. If a person can still recognize the text in the binary images, then this compression can be obtained with virtually no loss in semantic image content. The process of extracting text from the background of the document image is known as document image binarization. Edge detection techniques play a crucial role in this process. In this research work, a frame work has been proposed for digitizing historical documents. It suggests to use Markov random field that can evaluate contrast of pixels. Pixels are segmented and background pixels are filtered out using wiener filtering. Hence, foreground regions are extracted from the background region. Extensive experiments were conducted on various DIBCO datasets and the results shows significant increase in the PSNR value by 40-50%.**

Keywords: Document Binarization, PSNR Ratio, DIBCO.

I. INTRODUCTION

Historical document images are, in general, more difficult to binarize than modern scanned documents. This is in part due to their degraded state and partly because many historical documents are digitized with cameras, which do not have controlled illumination conditions like scanners. Some camera produced images have uneven illumination due to bad lighting or because the page is not flat (e.g., curved edges near book bindings). these documents suffer serious degradations. They may be affected from environmental conditions such as moisture, paper fold outlines, ink stains, document aging, etc. Therefore it is important to preserve them as a whole to avoid further degradations. In preserving the historical documents, the first step is to digitalize the physical document. The document may be inaccurately recognized due to scanning or capturing errors, illumination conditions and quality of documents. In the case of historical documents, it may have low ink /print quality, faded strokes, embedded neighboring figures competing with characters for recognition or document printed with tiny letters that contribute to the severity of the problem posed for high quality precision document recognition system.

II. IMAGE BINARIZATION

Binarization is the process of converting a multi-level input image to a new bi- level image i.e, each pixel intensity of the new image is represented by a value of either 0 or1. The pixels contain relevant information for the specific application. They are then classified into foreground and background. Foreground pixels are those that carry textual information or contain ink strokes. Binarization finds out the region of interest from a given image directed for a particular application. It is used in applications like optical character recognition, document layout analysis, text segmentation, writer identification, historical document preservation etc.

Fig. 1. Original image.

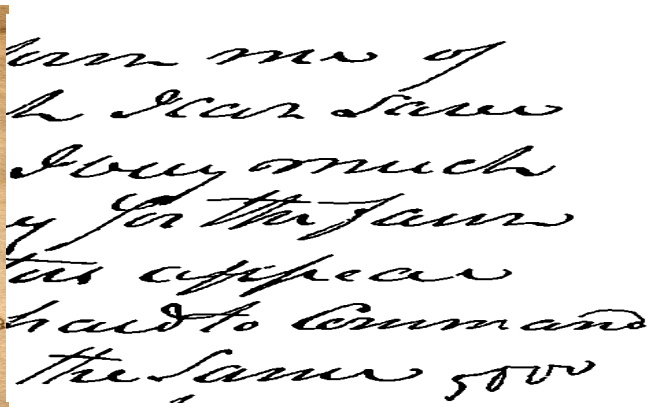


Fig. 2. After binarization.

III. PROPOSED WORK

In this work a framework for digitations of historical physical document has been proposed. Contrast of the pixel is evaluated using Markov random function. By using this function, problem of uneven illumination of the digital image is sorted. Following this, we use this energy to differentiate foreground and background ink. It incorporates advanced discontinuities in terms of regularity of the overall function of the power, distorting ink limits to align the edges and allow harder smoothing incentive. The following paragraphs describe all these points in more detail below with taking example of handwritten document i.e. HW3 from the dataset DIBCO-13.

Proposed algorithms for handwritten historical document binarization is show in Fig. 3. It consist color to gray conversion of input image, contrast measurement of each pixels, pixel segmentation, wiener filter and refined foreground regions.

Grayscale Conversion: -The input image is converted to grayscale image. This is done to smoothen background texture and eliminate noisy areas.

Discrete Wavelet Transform: - The proposed techniques use the DWT transformation scheme. The input image is decomposed into four components, namely, LL, HL, LH and HH, where the former letter corresponds to frequency offset of the row either low or high and the latter refers to filter applied to the columns.

The approximate details are given in lowest resolution level LL whereas rest three refer to detail parts and gives the vertical high (LH) containing vertical details, horizontal high (HL) containing horizontal details and high (HH) frequencies referring to diagonal details of the image.

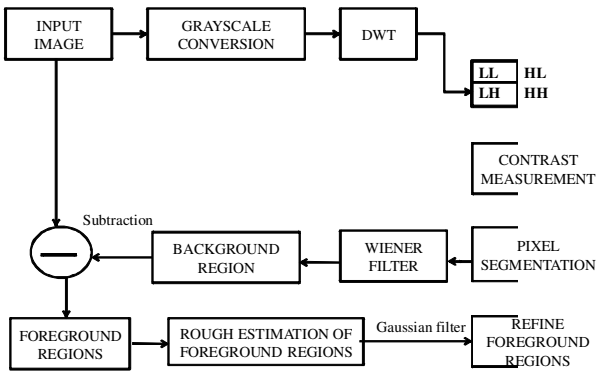


Fig. 3. Proposed Frameworks for Document Binarization

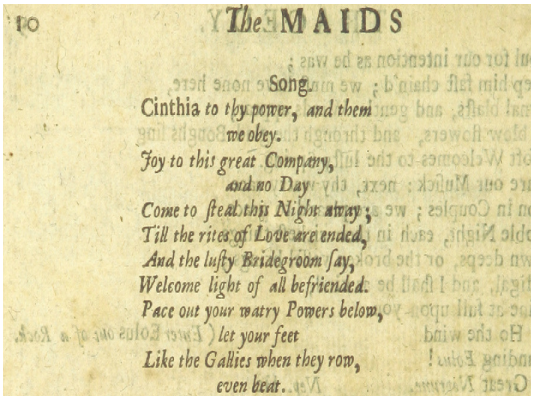


Fig. 4. Input Image from DIBCO-17.

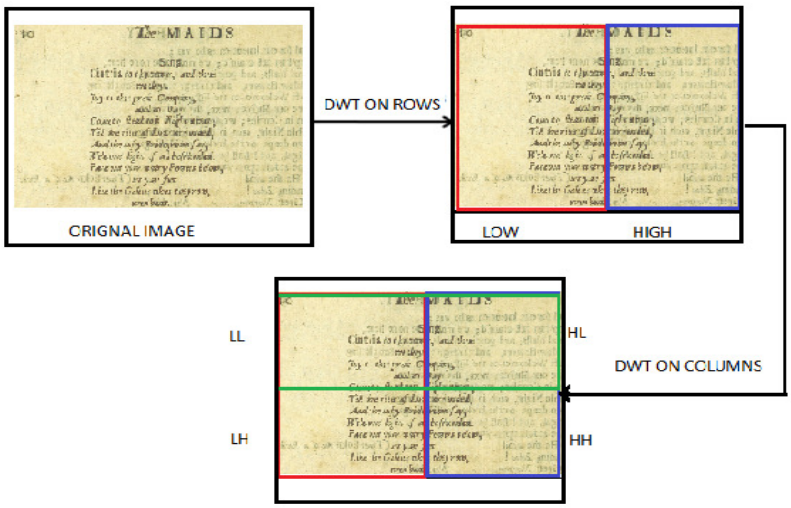
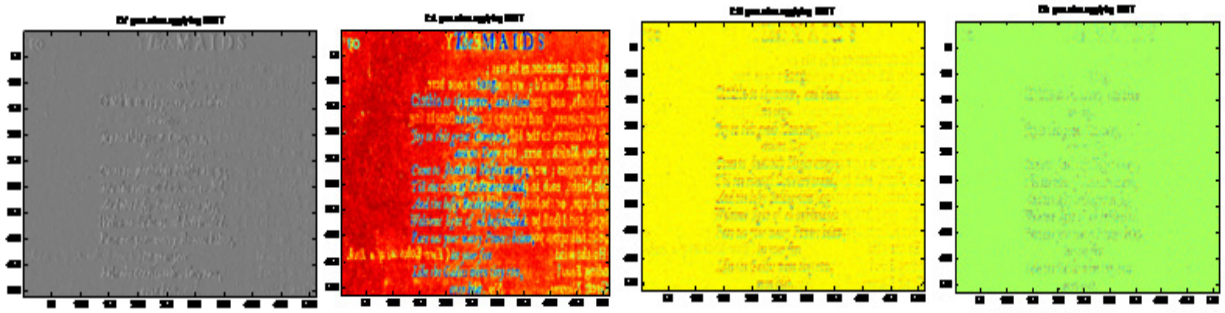


Fig. 5. DWT Transform of Image.

Fig. 5 illustrates the basic, one-level, two-dimensional DWT procedure. Firstly, we apply a one-level, one-dimensional DWT along the rows of the image. Secondly, we apply a one-level, one-dimensional DWT along the columns of the transformed image from the first step. As depicted in Figure 4 (left), the result of these two sets of operations is a transformed image with four distinct bands: (1) LL, (2) LH, (3) HL and (4) HH. Here, L stands for low-pass filtering, and H stands for high-pass filtering. The LL band corresponds roughly to a down-sampled (by a factor of two) version of the original image. The LH band tends to preserve localized horizontal features, while the HL band tends to preserve localized vertical features in the original image. Finally, the HH band tends to isolate localized high-frequency point features in the image.

Proposed document digitization scheme apply contrast Measurement and pixel segmentation over each band separately to obtain better horizontal, vertical, diagonal and approximate detail.



(a) (b) (c) (d)

Fig. 6. (a)Vertical, (b) Approximation (c) Horizontal (d) Diagonal detail coefficients of input image.

Contrast Measurement: Our framework uses Markov random field for measuring contrast of a pixel against its background. Markov random field is so chosen due to the fact that the regions in images are often homogenous i.e., neighboring pixels may have similar properties such as intensity, color or texture, etc. MRF is a probabilistic model that captures such constraints. A threshold is also calculated that segregates pixels into three categories namely foreground pixels, background pixels, and uncertain pixels. Thicker line patterns are better detected with large window size. The image contrast is evaluated by the following Eqn. (1)

$$p_c(i,j) = \frac{p_{c,max}(i,j) - p_{c,min}(i,j)}{p_{c,max}(i,j) + p_{c,min}(i,j) + \epsilon} \quad (1)$$

Wiener filter: -Wiener filter can be used to improve both the resolution and signal to noise ratio. It is the MSE-optimal stationary linear filter, which is useful for the images that have been degraded by additive noise and blurring. Calculation of the Wiener filter requires the assumption that the signal and noise processed are second-order stationary. Proposed scheme uses statically threshold that was calculated by using Markov model that filters out foreground image as noise. As show in Eqn. (2).

$$W_f(i,j) = \frac{t_{image} * c_{foreground}}{|t_{image}|^{2} * c_{foreground} + c_{background}} \quad (2)$$

However, if the image contains non-uniform background or too much noise, the contrast of the image may contain several peaks. Using a single threshold value to binaries the entire image would not produce a good binary image. Wiener filter is applied to the image foreground pixels are filtered out as noise. Thus we obtain background pixels. After subtracting background region from input image, we obtain foreground regions.

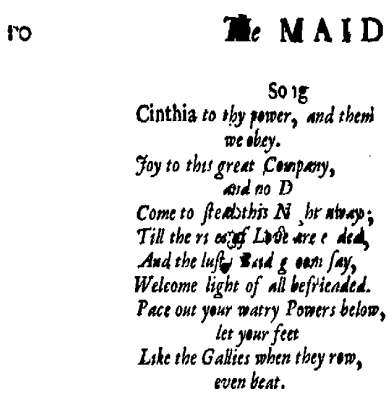


Fig. 7. Filtered Background Image.

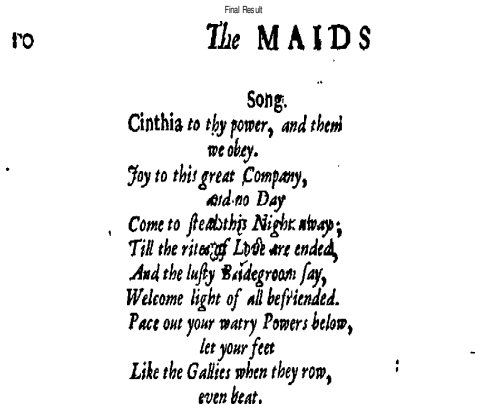


Fig. 8. Final Image.

Rough estimation of foreground regions: The work uses Gaussian filter to estimate foreground regions. It helps to reduce the noise and smooth out the image.

In Gaussian smoothing, weights give higher significance to pixels near the edges. This in turns reduces edge blurring. Therefore, Gaussian noise can be reduced using Gaussian smoothing. The degree of smoothing can be controlled by σ (larger σ for more extensive smoothing). The weights are calculated according to a Gaussian function:

$$G(i,j) = c.e^{\frac{i^2+j^2}{2\sigma^2}} \quad (3)$$

IV. RESULTS

Table 1 presents the results of binarization using the three implemented methods over the first dataset (DIBCO 2013) as well as the final ranking. The best method results are highlighted in bold. Overall, our method achieved the best results for all four measures, which confirms clearly its high accuracy in dealing with different documents types under various problems.

PSNR is calculated by using mean squire error of MSE. Both parameters are calculated by the following formulas.

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M * N}$$

The experimental results shows that the proposed algorithm gives the better performance compared to previous approach.

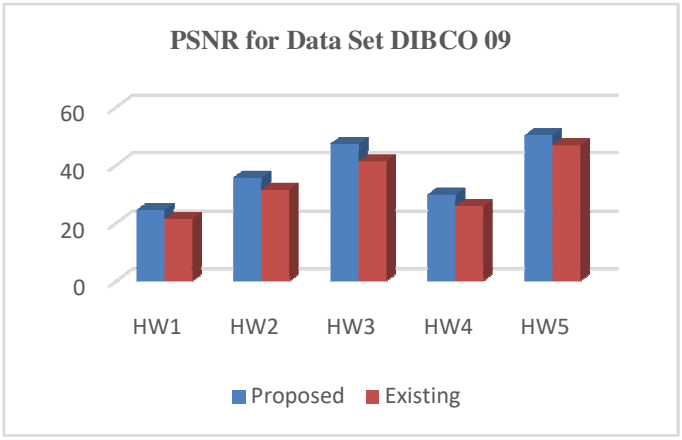


Fig. 9. PSNR Comparison Graph.

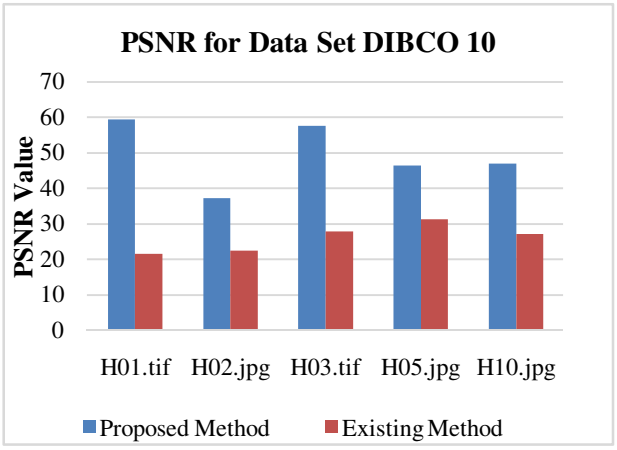


Fig. 10. PSNR Comparison Graph.

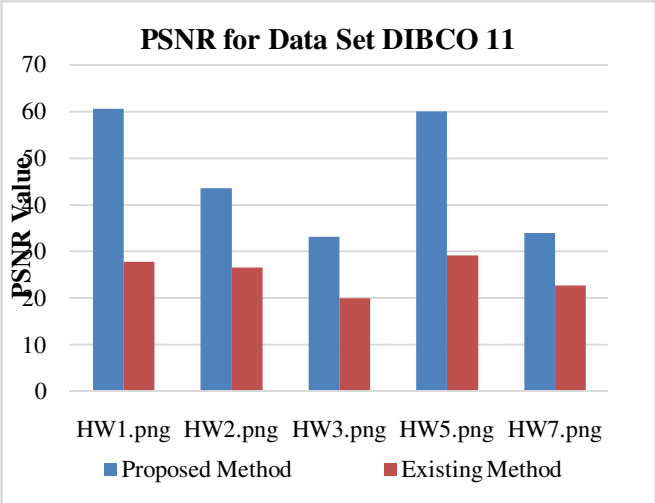


Fig. 11. PSNR Comparison Graph.

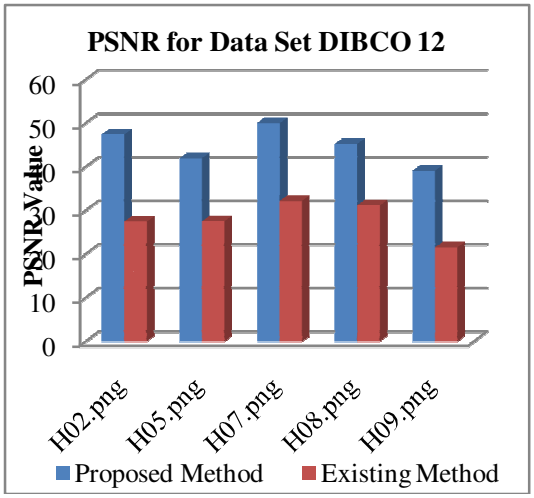


Fig. 12. PSNR Comparison.

This dataset contained handwritten images with added blur, noise and back to front interference problems. To a certain level, all methods were affected by the existence of these complications in these documents, as shown by the registered values in this dataset

V. CONCLUSION

The proposed novel robust approach for image binarization of degraded historical documents is on algorithm based of Markov random field model. Firstly, the algorithm includes grayscale conversion. Discrete wavelength transform is then applied to the image. Contrast of the pixels is measured using Markov Random field. Pixels are segmented into three categories namely background, foreground and uncertain pixels. Rough foreground regions are estimated and a Gaussian filter is applied. It smooths out the image and reduces noise. The number of single edges is reduced by almost half in this case because of its ability to recover weak and low intensity parts of the strokes on the edges. Nevertheless, our method has had a major inconvenience. The proposed method was evaluated on DIBCO datasets with promising results as compared to the existing binarization methods. Significant increase in PSNR about 40-50% can be observed by using the proposed method. The work done in this thesis has overcome the drawbacks of detecting the distorted edges by using edge detection. This framework used MRF and tried to overcome the problems occurred in the degraded historical documents.

REFERENCES

[1]. Rahman, N.A., Zuki, S.A.M. and Yassin, I.M., (2012). "A review of image processing technique in particle mixing analysis. *IEEE 2012*, 466 – 469.

[2]. Jing Zhang, and Nath, B. (2004). "Image processing techniques of landmines: a review", IEEE 2004, pp 143 – 148.

[3]. J. Kittler, and J. Illingworth, (1986). "Minimum error thresholding. *Pattern Recognition*, **19**, 41-47.

[4]. J. Sauvola, and M. Pietikinen, (2012). "Adaptive document image binarization," Pattern Recognition 2000, **33**, pp. 225-236.

[5]. Zemouri, E.T.T Chibani, and Y. Brik, (2014). "Restoration based Contourlet Transform for historical document image binarization. *IEEE*, 309-313.

[6]. D. Hebert, Nicolas, and S. Paquet, (2013). "Discrete CRF Based Combination Framework for Document Image Binarization, 1165-1169.

[7]. H.Z Nafchi R.F Moghaddam, and M. Cheriet, (2013). "Application of Phase-Based Features and Denoising in Postprocessing and Binarization of Historical Document Images, 220–224.

[8]. S. Milyaev, O. Barinova, and T. Novikova, (2013). "Image Binarization for End-to-End Text Understanding in Natural Images", *IEEE,* 228-232.

[9]. Bolan Su, Shijian Lu, and Chew Lim Tan, (2012). "Robust Document Image Binarization Technique for Degraded Document Images ", *IEEE,* 1408-1417.

[10]. Bolan Su, Shijian Lu; and Chew Lim Tan, (2012). "A learning framework for degraded document image binarization using Markov Random Field", *IEEE,* 3200 – 3203.

[11]. Yinghui Zhang, Tianlei Gao, DeGuang Li, and Huaqi Lin, (2012). "An improved binarization Algorithm of QR code image", *IEEE,* 2376 – 2379.

[12]. Bolan Su, Shijian Lu, and C.L Tan, (2011). "Combination of Document Image Binarization Techniques", *IEEE,* 22-26.

[13]. Yuanping Zhu, (2008). "Augment document image binarization by learning", *IEEE,* 1-4.

[14]. Stathis, P. Kavallieratou, and E. Papamarkos, (2008). "An evaluation survey of binarization algorithms on historical documents.
[15]. Gatos, B. Pratikakis, I. and Perantonis, S.J., (2008). "Efficient Binarization of Historical and Degraded Document Images", *IEEE,* 447-454.
[16]. J.He, Q.D. MDo, A.C. Downton, and J.H. Kim, (2005). "A comparison of binarization methods for historical archive documents", *IEEE*, 538-542.