International Journal of Electrical, Electronics and Computer Engineering **8**(1): 01-10(2019)

ISSN No. (Online): 2277-2626

Image Binarization of Deteriorated Historical Documents for Document Image Analysis

***Isra Naqvi*[1] *and Abdul Samee Khan*[2]**

[1]M.Tech Student, Department of Electronic and Communication Engineering,
All Saints College of Technology Bhopal (Madhya Pradesh), India
[2]Assistant Professor, Department of Electronic and Communication Engineering,
All Saints College of Technology Bhopal (Madhya Pradesh), India

(Corresponding author: Isra Naqvi)

(Received 07 September, 2018 Accepted 20 December, 2018)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Historical documents are original documents that contain important information about the person, place, event, etc. and can thus serve as primary source of the historical methodology. Despite of its importance, digitization of old documents is done for internet dissemination. The analysis of historical documents are often difficult than other types of handwritten or printed texts due to their complex structure. They may suffer from faint characters, stroke width variability, bleed through, large background ink stains and periodic shadings that leads to degrade the quality of image. Document binarization plays an important role to preserve the historical documents. It focuses on extracting the text from the background of the image. In doing this, the edge detection approach plays a crucial role. In this research work, a framework for digitization of historical physical document has been proposed. This framework suggests using Markov random function that can evaluate contrast of pixels. It overcomes the problem of appearance of a single document which varies on the factors such as uneven illumination, viewing angle and noisy background of different portions of image document. Finalbinarized document image has significant enhancement in PSNR (db.) value by 40-50%. Proposed scheme uses DIBCO database for evaluation and validation of the proposed method.

Keywords: Document Binarization, PSNR Ratio, DIBCO

I. INTRODUCTION

Historical documents are of prime importance to the nation. It contains a number of information that can reveal details of the roots of any nation, tribe, culture, religion etc. Over the time, 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

Image binarization contributes to maintain a safe and effective preservation of severely degraded documents. It is an important research theme over the years. Our work on the basic techniques is used to improve, restore the image, eliminate noise and focus on binarization. In olden days, binarization was important for sending faxes. Nowadays it is important for digitalizing text or segmentation. Binarization is the basis of segmentation. It is used as a pre-processor before OCR.

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 or 1. 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.

III. DOCUMENT IMAGE BINARIZATION CLASSIFICATION

Document Image Binarization takes place in digitized document analysis system as one of the first processes.

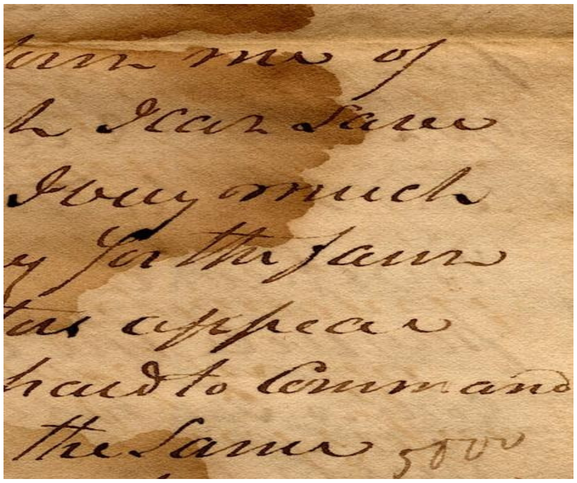


Fig. 1. Original image.

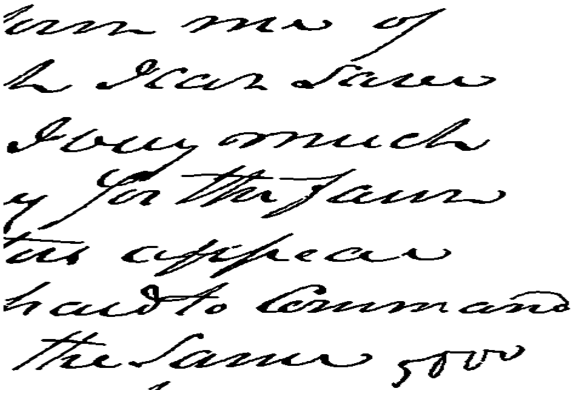


Fig. 2. After binarization.

It can be used in various image processing applications such as in optical character recognition (OCR) and self-image recovery. Many document binarization methods have been reported in the literature that can be roughly categorized into two groups: i) global thresholding methods, which assign a single threshold for the whole document image and ii) local thresholding methods which assign a threshold for each pixel or a small region of the document images. The basic thresholding techniques are illustrated below

A. The global thresholding technique

It calculates the optimal threshold for the entire image. These techniques require some calculations and can work well in simple cases. Sometimes, global thresholding is insufficient, i.e, the constant threshold value over the image is insufficient. For example, it will happen that a part of the background region is brighter than some target region. In this case, global thresholding cannot extract all the targets without any false extraction from the background region. The shaded background is partially brighter than a (darker) target and no global threshold can extract those two targets correctly.

B. The local binarization techniques

It sets different thresholds for different target pixels based on the district / local information. Generally, these techniques are sensitive to background noise due to the great variation in the case of a poor document or luminous deterioration bleed through.

C. Hybrid binarization approach

It combines both global and local threshold. The first step is to conduct a global threshold for the classification of the bottom of the document image and keep only the part that contains the foreground. The second step aims to improve the image obtained by the previous step to get the result more clearly through the application of adaptive threshold technique.

D. Dynamic Threshold Binarization

Such as redundancy and knows the threshold of pixels with gray level values of their own and their neighboring pixels and the pixel format. This method is used usually for binarizing images of poor quality. However, due to the expense of the dynamic threshold, this method has a high computational complexity and slow speed.

IV. CURRENT SCENARIO

Document image binarization is an important technique for image analysis and document pre-treatment to the text segments. Many of the techniques proposed are successfully applied in various applications such as document retrieval image. Zemouri, E.T.T Chibani et.al [5] propose the conversion contourlet to assess the quality of a deteriorated historical document. To facilitate binarization, they first improved the quality of document by applying a conversion contourlet. After reconstruction, they used the local threshold method to extract the plain text. D.Hebert, Nicolas [6] proposed a CRF based framework to explore the capabilities of the combined model by combining many of the outstanding output binarization algorithms. The Framework uses two 1D CRF models on the horizontal and vertical directions that are associated to each pixel of the by-product of the marginal probabilities calculated from both models used. The experiments were conducted on two sets of image data of document Binarization Contest (DIBCO) for 2009 and 2011, and it shows the better performance than most of the methods presented in DIBCO 2011.

H.Z Nafchi, [7] processed and analyzed the results of steps to significantly improve the performance of binarization, especially in the case of severely degraded historical documents. They presented a post processing method based on the phase-preserved denoised image and also phase congruency features extracted from the input image. Essence of the method consists of two powerful mask can be used to delete the false positive output pixel in the output of binarization. First, we get a mask value of high recovery of the image without noise using morphological operations. In parallel, the second mask is obtained based on the characteristics of phase congruency. Then, the average filter is used to remove the noise in these two masks, which are then used to correct the output of any method of binarization. This approach has been tested along with many of the previous methods in the DIBCO’09, H-DIBCO'10, DIBCO'11 H-DIBCO'12 datasets with promising and improved results. Moreover, high-performance proposals masks appear likely to use as a uncensored truth generator for binarization based learning. Milyaev [8] demonstrated that the OCR engines still work well as long as appropriate application of binarization is used on the input images. This binarization and the performance of 12 binarization methods are evaluated. This includes our assessment of the various standards and criteria prescribed uses text recognition from natural images (ICDAR 2003 and 2011). So their main conclusion is that the image binarization methods along with additional filtering of connected components generated and the OCR engine off the shelf can achieve the improved performance for end to end text understanding in natural images. Bolan Su [9] proposes a novel image binarization technique that addresses the problems of different types of document degradation such as uneven illumination and document smear. This technique uses an adaptive image contrast. Adaptive contrast of the image is a combination of local image contrast and the local gradient of the image, which is tolerant to the text and the background variation caused by different types of document degradation. In the proposed technique, they first constructed the adaptive contrast map. This contrast map is then binarized and combined along with Canny’s edge map to determine the text stroke edge pixels. The text is fragmented by local threshold that is estimated on the basis of the intensities of detected text stroke edge pixels within the local framework. The proposed method is simple, powerful, and involves a minimum set of standards. It has been tested in three public data sets used in the recent tender documents image binarization (DIBCO) for 2009 and 2011 and HDIBCO 2010 and achieved accuracy of 93.5%, 87.8% and 92.03%, respectively, which are much higher than or close to reported better performance in all three competitions. Experiments on daily data Beckley Group, which consists of several difficult images of documents of poor quality and superior performance show the proposed method compared with other techniques. Bolan Su, Shijian Lu [10] proposed a learning which makes use of method known as Markov random field to improve the performance of existing document image binarization methods. Large-scale experiments on modern document image data Binarization contest shows that significant improvements of the current methods of binarization in the application of the framework of the proposed work. Yinghui Zhang [11] proposes an improved algorithm based on the background of gray level image binarization for the non-uniform illumination QR code image. First they did the sub-block processing according to the size of the QR Code. On this basis, they use the gray level estimation formula for calculating the value of the gray level of each block. Second, they used full interpolation algorithm to build a background image in gray. Then they used this gray background image adapt to the original image to get the corrected image. Finally, the Otsu algorithm for image binarization is applied. Experiments show that the algorithm can make effective image correction varying lighting QR code, and get a good binary image. Bolan Su.[12] proposes a classification framework combining different threshold methods and produce better performance of document image binarization. Given the results of binarization of some media reported, the proposed framework divides the document image pixels into three groups, a pixel in the foreground, the background pixels and pixel of uncertainty. The uncertain pixels are iteratively classified by a classifier into background and foreground.it is based on preselected background and foreground sets. Large-scale experiments on different data sets, including document image Binarization Contest (DIBCO) 2009, and a handwritten document binarization of competition image (H-DIBCO) 2010 shows that their proposed framework outperforms most of the state-of-art methods dramatically. Yuanping Zhu [13] proposes a method for binarization on the basis of learning that can be the used in same type of document binarization improvement, especially in the quality stability. It contains two stages i.e., learning and a performing binarization stage. Binarization evaluation and optimization is done by obtaining knowledge from learning stage.

During the performing step, the result obtained from the binarization step is fed back to the binarization in order to adjust the parameters of binarization. This will then improve the overall performance. The experiments validate the improvement. Stathis [14] tries to answer the question that how well existing binarization techniques can binarize the degraded document image. They proposed a new technique to verify algorithms and document binarization. Their method is simple. It can also be implemented on any binarization algorithm since it requires only a binarization stage. Then they applied this proposed method to 30 technical binarization existing algorithms. The experimental results are displayed. Gatos [15] presents a new adaptive binarization and the improvement of historical and degraded. The proposed method is based on (i) pre-processing efficiency; (ii) the combination of the outcomes of several state-of-art binarization methodologies. (iii) the edge information incorporation and (iv) the application of effective image post-processing on the basis of mathematical morphology to improve the final result. The superior performance was demonstrated against six well-known techniques in many handwritten and machine printed historical documents mainly from the Library of Congress of the United States. The performance is based on a fixed and concrete assessment methodology. J. He, Q.D.M. Do [16] compares several alternative binarization algorithms for historical archive documents, by evaluating their impact on the performance of the recognition of the end-to-end word to identify and archive documents a complete system using commercial OCR engine. Algorithms evaluated are: global threshold, Niblack algorithms and Sauvola's algorithm. Adaptive versions of the algorithms of Niblack and Sauvola; and Niblack and Sauvola algorithms are applied to background removed images. We found that we have the archive documents, Niblack algorithm can achieve better performance than Sauvola (allegedly evolution of Niblack) algorithm, and also achieved better than the internal binarization provided as part of the commercial OCR engine performance.

V. PROPOSED WORK

In this research work a framework for digitization 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. 19 from the dataset of DIBCO-17 (Fig. 4).

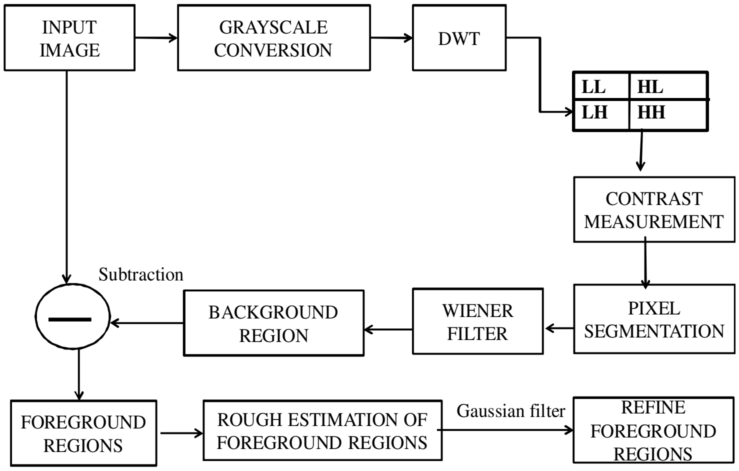
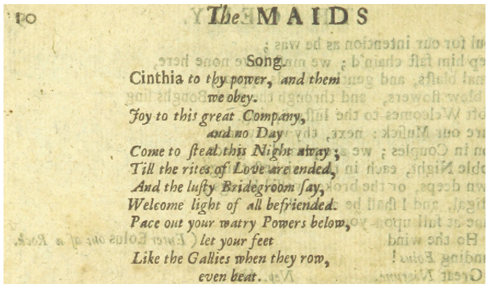


Fig. 3. Proposed Frameworks for Document Binarization.

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



10 The MAIDS

Song.

Cinthia to thy power, and them
we obey.
Joy to this great Company,
and no Day
Come to ſteal this Night away;
Till the rites of Love are ended,
And the luſty Bridegroom ſay,
Welcome light of all befriended.
Pace out your watry Powers below,
let your feet
Like the Gallies when they row,
even beat.

Fig. 4. Input Image 19 from DIBCO-17.

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

Discrete Wavelet Transform: It is an implementation of the wavelet transform using a discrete set of the wavelet scales and translations obeying some distinct rules. Wavelets permit both frequency and time analysis of signals simultaneously. It is due to the fact that energy of wavelets is concentrated in time and still posses. In other words, this transform decomposes the signal into mutually orthogonal set of wavelets, which is the main difference from the continuous wavelet transform (CWT), or its implementation for the discrete time series sometimes called discrete-time continuous wavelet transform (DT-CWT).

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.

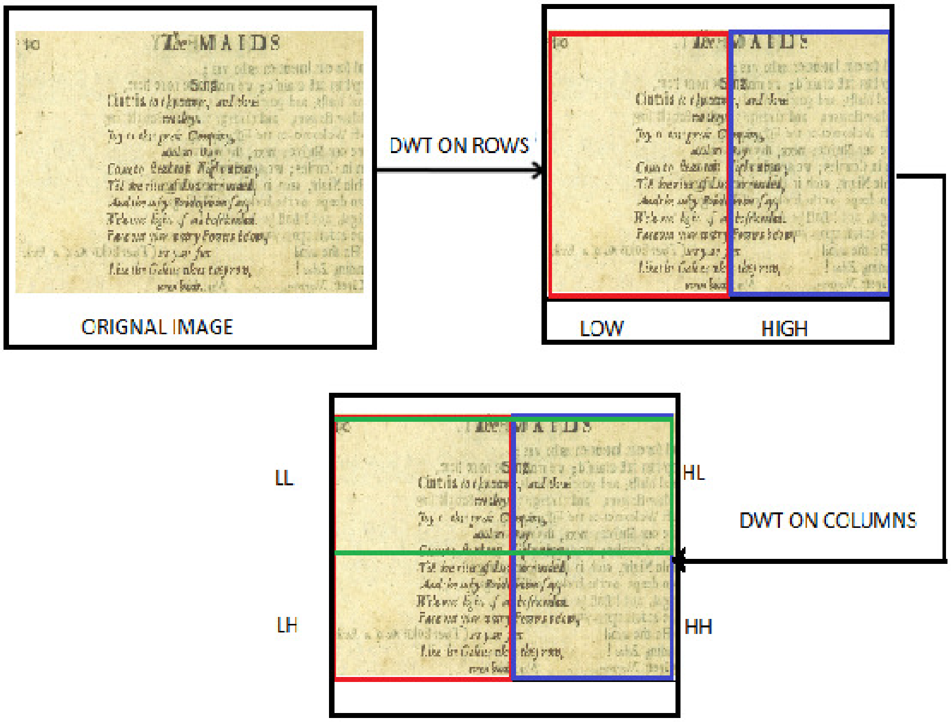
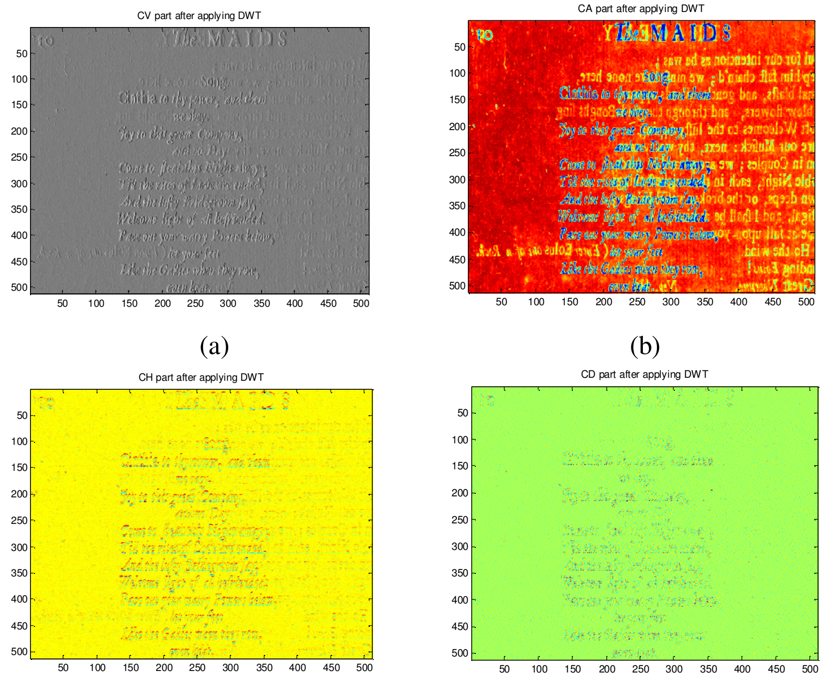


Fig. 5. DWT Transform of Image.

Figure 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 5 (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. As in the one-dimensional case, we do not necessarily want to stop there, since the one-level, two-dimensional DWT extracts only the highest frequencies in the image. Additional levels of decomposition can extract lower frequency features in the image; these additional levels are applied only to the LL band of the transformed image at the previous level.

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)

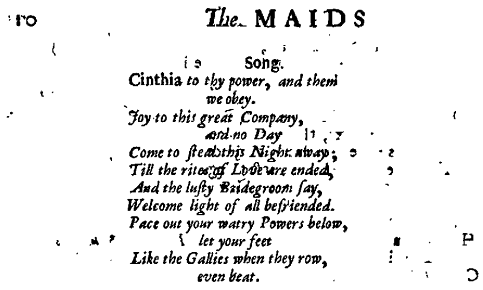
(d)

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. The value of foreground pixel that is darker, is usually a positive value and the background pixels that is lighter have negative, zero or small positive values. Thicker line patterns are better detected with large window size. The image contrast is evaluated by the following equation 1

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

Where Pc, max (i, j) and Pc, min (x; y) refer to the maximum and the minimum image intensities within a local neighborhood window. The term ϵ is a positive but infinitely small number to avoid dividing zero problems.



ro The MAIDS

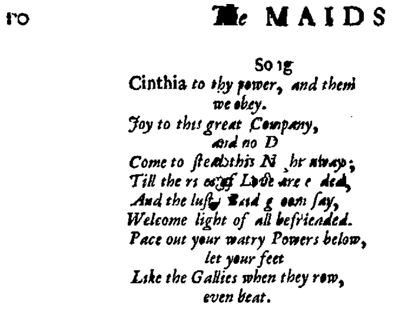
Song.
Cinthia to thy power, and them
we obey.
Joy to this great Company,
and no Day
Come to ſteal this Night away;
Till the rites of Love are ended,
And the luſty Bridegroom ſay,
Welcome light of all befriended.
Pace out your watry Powers below,
let your feet
Like the Gallies when they row,
even beat.

Fig. 7. Contrast Image.

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 equation 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.



ro The MAIDS

So ıg

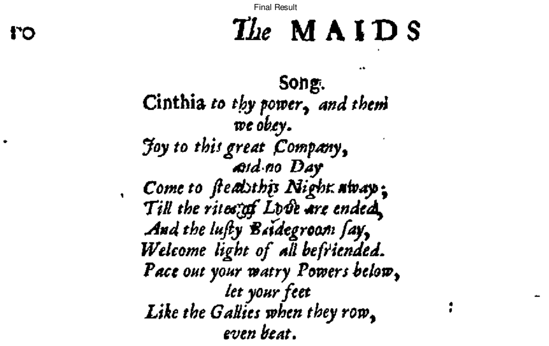
Cinthia to thy power, and them
we obey.
Joy to this great Company,
and no D
Come to ſteadthis N ht away;
Till the ri of Love are e ded,
And the luſty Bad g oom ſay,
Welcome light of all befriended.
Pace out your watry Powers below,
let your feet
Like the Gallies when they row,
even beat.

Fig. 8. Filtered Background 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:

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



Final Result

ro The MAIDS

Song.

Cinthia to thy power, and them
we obey.
Joy to this great Company,
and no Day
Come to ſteal this Night away;
Till the rites of Love are ended,
And the luſty Bridegroom ſay,
Welcome light of all befriended.
Pace out your watry Powers below,
let your feet
Like the Gallies when they row,
even beat.

Fig. 9. Final Image.

VI. RESULT

The proposed concept has been implemented on MATLAB. In this simulation, various DIBCO datasets are used.

PSNR is a ratio in an image which is often used as a quality measurement between original and reconstructed image. It 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.

Table 1: PSNR Comparison Results.

Image Name	Proposed					Existing				
Year	2009	2010	2011	2012	2013	2009	2010	2011	2012	2013
HW1	24.49	59.40	60.55	47.42	32.43	21.45	21.67	27.84	27.49	20.84
HW2	35.67	37.16	43.55	41.97	56.63	31.45	22.46	26.56	27.54	26.49
HW3	47.34	57.52	33.20	50.05	45.90	41.32	27.87	20.01	32.18	26.78
HW4	29.75	46.41	60.05	45.25	46.77	25.75	31.24	29.12	31.20	25.87
HW5	50.23	46.97	33.94	39.09	32.34	46.87	27.17	22.67	21.56	23.07

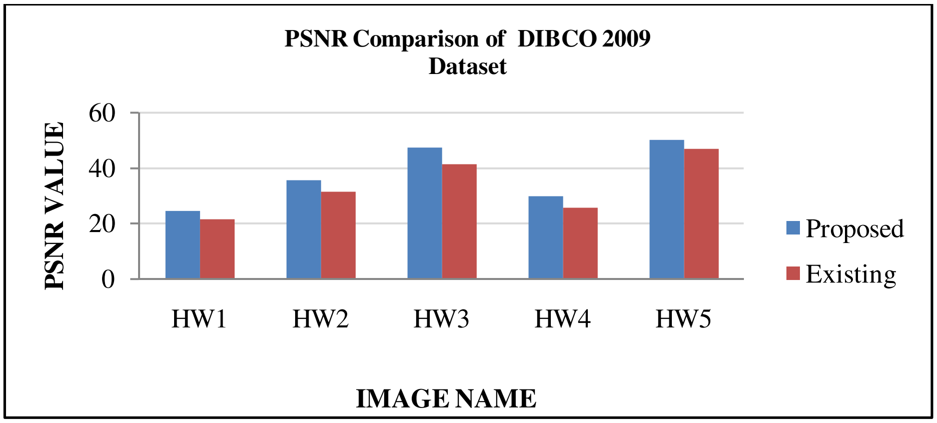


Fig. 10. PSNR Comparison Graph.

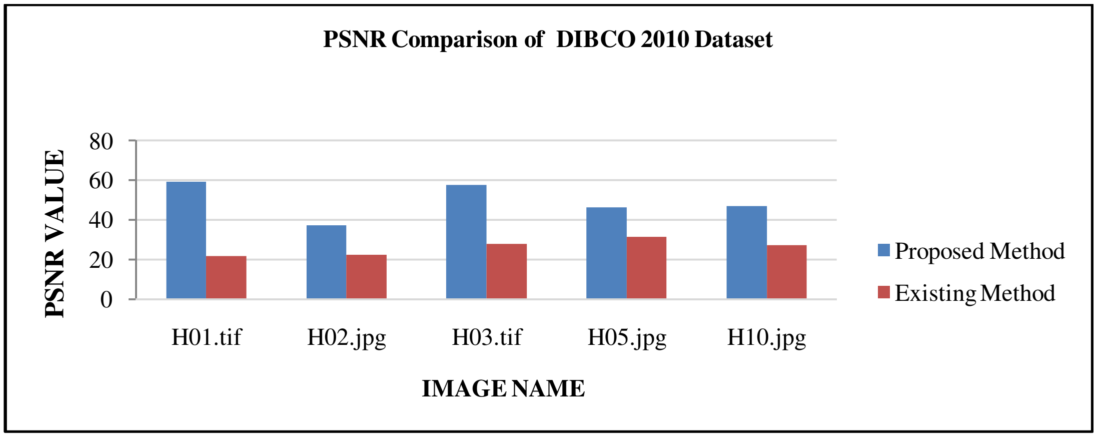


Fig. 11. PSNR Comparison Graph.

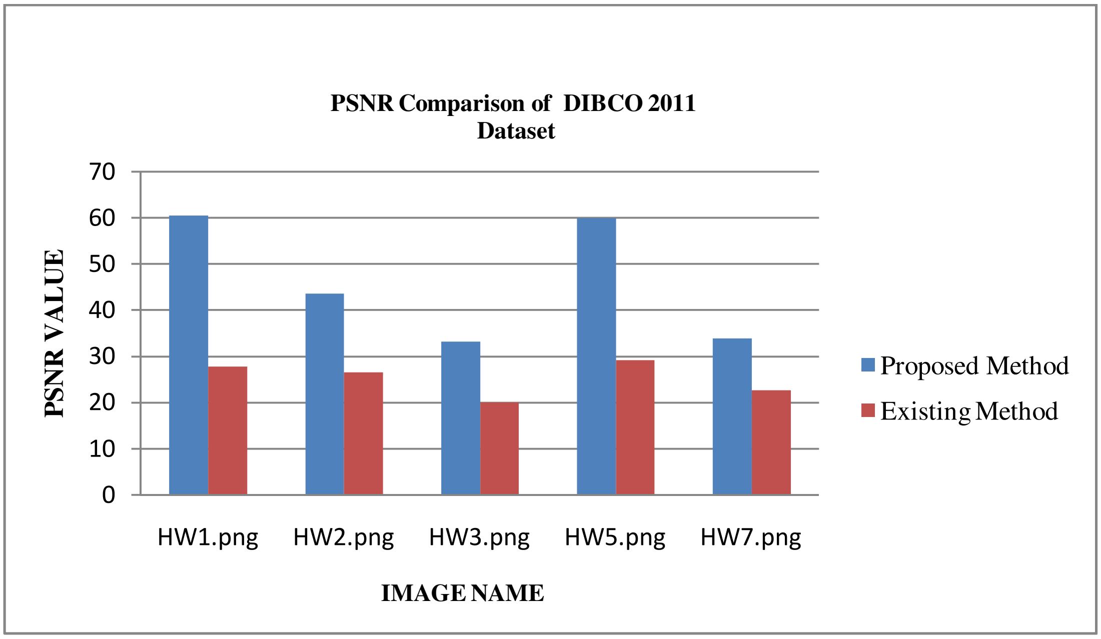


Fig. 12. PSNR Comparison Graph.

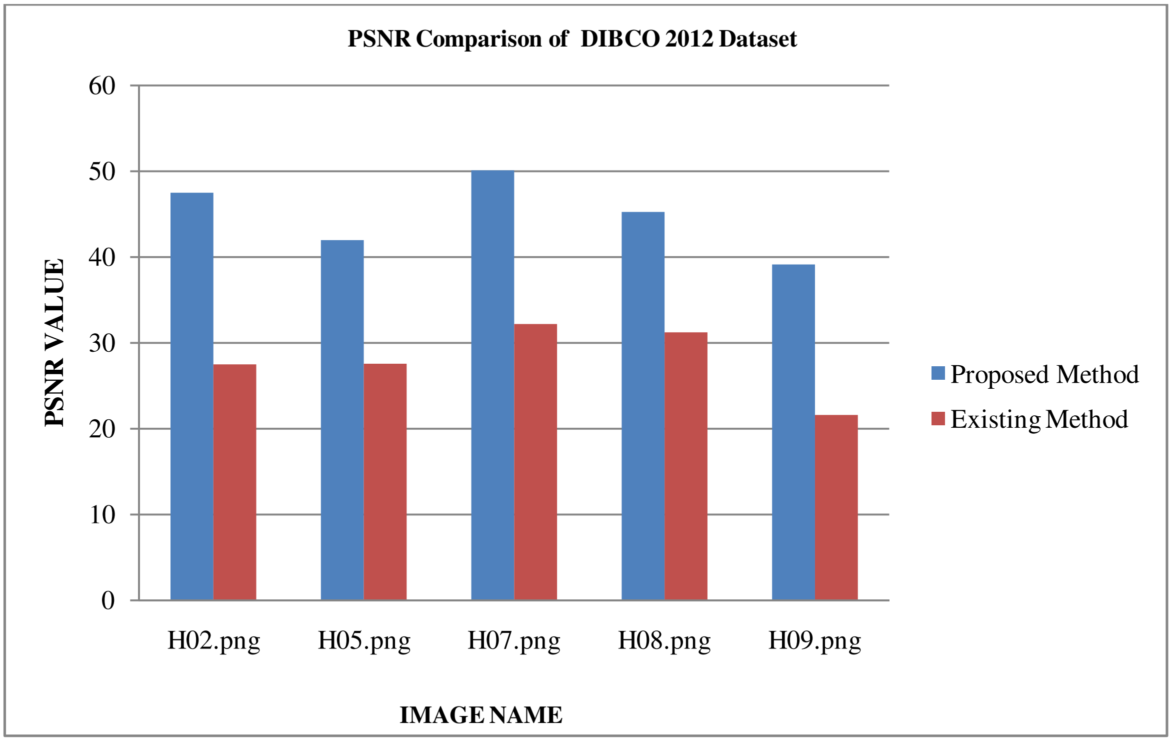


Fig. 13. PSNR Comparison Graph.

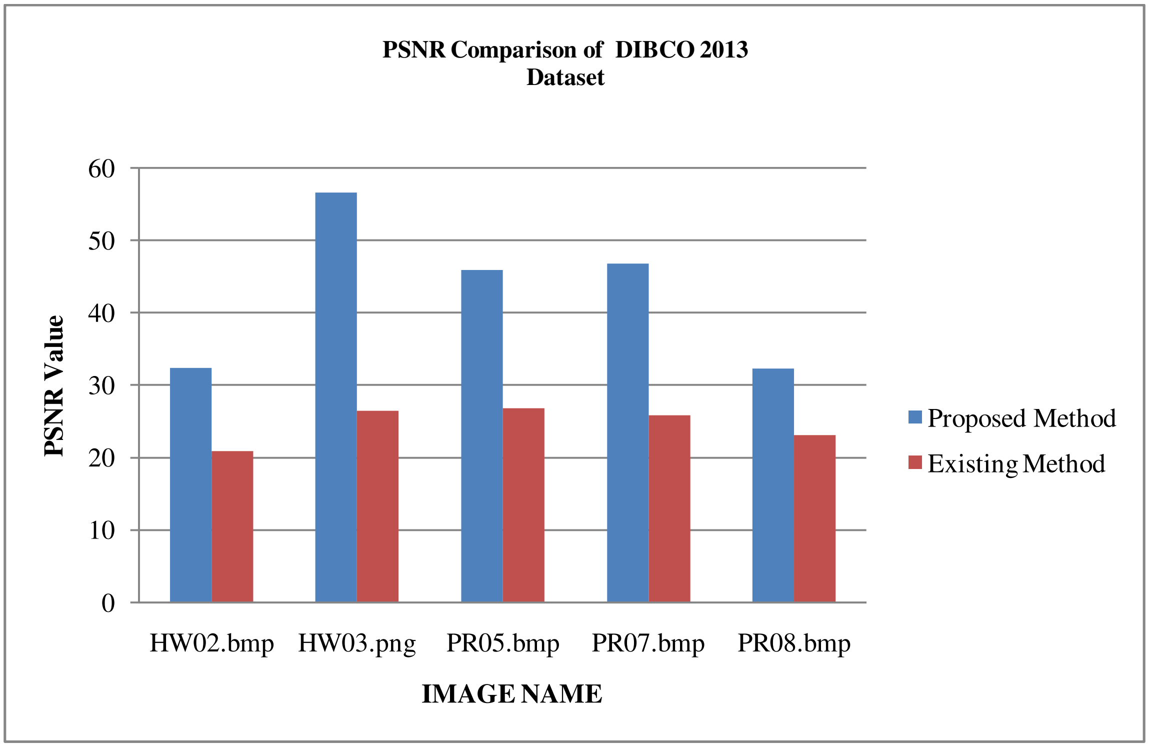


Fig. 14. PSNR Comparison Graph.

The proposal is that the higher the PSNR, the better degraded image has been reconstructed to match the original image. Hence, the better reconstructive algorithm as shown in figure 10 to 14 for DIBCO 2009 , 2010, 2011, 2012 and 2013 respectively. For any ideal binerzation it is desired to have better PSNR ratio and the proposed scheme has higher PSNR ratio as compare to the existing one.

VII. CONCLUSION

This work improves the document image binarization framework that makes use 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. 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.; Yassin, I.M., (2012). "A review of image processing technique in particle mixing analysis", IEEE 2012, pp 466–469.

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

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

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

[5]. Zemouri, E.T.T Chibani, Y. Brik, (2014). "Restoration based Contourlet Transform for historical document image binarization", IEEE, pp309-313, 2014.

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

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

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

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

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

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

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

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

[14]. Stathis, P. Kavallieratou, E. Papamarkos, (2008). "An evaluation survey of binarization algorithms on historical documents" IEEE 2008

[15]. Gatos, B.Pratikakis, I.Perantonis, S.J., (2008). "Efficient Binarization of Historical and Degraded Document Images", IEEE, pp 447-454, 2008.

[16]. J.He, Q.D. MDo, A.C. Downton, J.H. Kim, (2005). "A comparison of binarization methods for historical archive documents.' ", IEEE, pp 538-542.