Evaluation of Binarization Methods for Aged Printed Myanmar Documents

Aye Su Phyo, Khin Nwe Ni Tun
University of Computer Studies, Yangon
ayesuphyo@ucsy.edu.mm, knntun@ucsy.edu.mm

Abstract

Binarization is one of sub phases of preprocessing step of optical character recognition (OCR). Binarization is separation of foreground text from background of document image. The accuracy of OCR mainly relies on binarization's result. This paper compares several alternative binarization algorithms for aged printed Myanmar documents. The algorithms evaluated are global thresholding (Otsu), Local thresholding (Niblack, Sauvola, Wolf, Feng and Nick). It is found that the binarized images more stable if filters (Wiener and Gaussian) are prior used before applying binarization algorithms. Another one is that local thresholding is suit for aged Myanmar documents. Among local thresholding, Niblack, Sauvola and Wolf are the more suitable algorithms based on the experimental results. The quality of binarized images is verified by using different assessment parameters like mean square error (MSE), signal to noise ratio (SNR) and peak signal to noise ratio (PSNR). This work aims to get the high accuracy of recognition steps with the main objective of developing OCR of aged printed Myanmar documents.

1. Introduction

The importance of digital libraries for information retrieval cannot be denied. The previous cultural, historical and religion books contain invaluable knowledge but it is very time consuming to search the required information in these paper books. Different methods have been devised to facilitate this information search. These include word spotting, optical character recognition, etc. Although a lot of work has already been done in this field, it still remains an inviting and challenging field of research, mainly because the results achieved so far are not satisfactory for huge volumes of data; especially if the document base consists of a set of ancient printed documents of relatively degraded quality [1, 6, 8, 20].

Optical Character Recognition (OCR) is renovation of scanned images of printed text, handwritten text into computer readable text. OCR is widespread techniques of computerizing printed data such that can be automatically searched, trimly saved, shown online, and exercised in computer. There are number of phases while doing recognition such as image acquisition, preprocessing, segmentation, feature extraction, classification, etc. Preprocessing consists of noise removal, binarization, skew correction, size normalization, boundary detection, and thinning, etc. Aim of binarization is to extract text and eliminate the background. Binarization plays an important role in OCR since its performance is quite critically the degree of success in subsequent character segmentation and recognition [17].

In general, the document binarization deals with the two categories: (i) global, (ii) local. In global approach, a single threshold value is selected for the entire image and is processed with value it mainly result good in separation of foreground and background intensity but in poor contrast, variable intensity of foreground-background these method fails to binarize the image. Local approach, local area information for calculating the threshold value uses for each pixel or sub image. This approach is used in the case of historical documents, there exists degradation such as shadow and non-uniform illumination, ink seeping and strains as it deals with them in adaptive manner. Documents are degraded as due to aging, humidity, temperature, poor quality and much more. There is always as different approach will get better result for a different type of degraded image. Generally, there are three basic steps to process a degraded image: (i) preprocessing of a document image, (ii) applying suitable binarization technique, and (iii) post processing [9, 16, 22].

Preprocessing is mainly denoising the image by using filters in Spatial (Mean, Median, and Wiener) and Fourier (Butterworth, and Gaussian) domain. Binarization is done many methods. The basic method for performing binarization is Otsu

method, which is based on maximizing the separation between two pre assumed classes. As well as there are some local methods – Niblack and Sauvola methods. After performing this, there is binary image that's containing text, black spots, broken edges, etc. Removal, join of these problem and enhancement of the image is performed in post processing [11].

In a literature survey, we found that there are many techniques which are being used by these researchers for the process of binarization but the prominent techniques [24] are Otsu, Niblack, and Sauvola. In this paper, we investigated three well-known binarization methods (Otsu, Niblack, and Sauvola) and also developed other adaptive Niblack methods which are Feng, Wolf, and Nick. Finally, we compared the quality of six methods and found that there is no single method which is successful for each images. We visually analyzed which type of method is suitable for which image. And then, we compared the assessment parameter values of these binarized images.

2. Review of the Binarization Methods

The purpose of document image binarization is to produce an image with black (or white) text on white (or black) background. Pixels in the gray image will belong to only two distinct gray levels – one gray level corresponding to the background and the other corresponding to the text pixels [5, 7, 21].

As discussed above, binarization techniques can be global and local. Global binarization is not useful in the case where a document image contains degradations such as variable illumination and contrast. Some referred studies on global binarization methods are due to Abutaleb [2], Kapur et al [13], Kittler and Illingworth [12], Otsu [18]. Otsu approach to global binarization is the most successful approach for OCR systems because of its computational efficiency and effective [19]. Otsu proposed a global thresholding method based on statistical methods. Any of the existing global method cannot solve the problem of non-uniform illumination as the threshold value is same for whole of the document but illumination is not.

Local binarization techniques perform better in the presence of non-uniform illumination and degradation in document images. A number of approaches to local thresholding are available in the literature [5, 6, 8, 10, 14, 15]. Bernsen et al [10] proposed a local thresholding method based on mean and contrast information for the calculation of threshold over a local region. According to Trier and Taxt [19] evaluation, Bernsen approach almost completely recovers text but produces large background noise especially where the document regions do not contain text.

Niblack [24] proposed a method that calculates a pixel-wise threshold by shifting a rectangular window across the image. According to Gatos et al [4], Niblack method fails when the background contains light texture. Sauvola et al [15] proposed a similar method by making some assumptions based on the distribution of gray values associated with foreground and background pixels. Sauvola approach shows robust results in the processing of severely degraded document images. It can handle variable illumination, noise, and resolution variation in the document image but fails in very light and very dark background [4]. Gatos et al developed a method, using a combination of existing techniques for degraded documents with uneven background. Gatos et al method is an adaptive local binarization approach, which removes smear and strains, variable illumination, low contrasts, large signal-dependent noise and shadow types of degradations from the document image. On the basis of comparison of results with existing binarization approaches discussed in [8, 15, 18], Gatos method is promising but fails in case of low illumination.

On the basis of the literature review of local and global thresholding algorithms, we conclude that none of the approaches produce satisfactory results on severely degraded images that contain distinct type of degradations such as variable background and shadows, non-uniform illumination and contrast, and ink bleed-through.

3. Binarization Methods

3.1. Global Thresholding

The global thresholding chooses a fixed intensity threshold value T (from 0 to 255). If the intensity value of any pixel of an input image is more than T, the pixel is set to white otherwise it is black.

3.1.1. Otsu Algorithm

One of the most common global thresholding, Otsu's thresholding chooses the threshold to minimize the intraclass variance of the thresholded black and white pixels.

3.2. Local Thresholding

The local thresholding is an adaptive one in which a threshold value is determined over a small region. Local thresholding performs better in case of badly illuminated images and document image analysis as threshold computation is dependent on region characteristics.

3.2.1. Niblack's Algorithm

Niblack's algorithm [24] calculates pixel-wise threshold by sliding a rectangular window over the gray level image. The threshold T is computed by using the mean m and standard deviation s, of all pixels in the window, and is denoted as:

$$T = m + k * s \quad (1)$$

where *k* is a constant, which determines how much of the total print object edge is retained, and has a value of between 0 and 1. The value of *k* and the size SW of the sliding window define the quality of binarization. The value of *k* controls the amount of text region inside the local window. To conserve local details and handle local illumination level, one requires small window size but if we choose too small window size, then it will not cover object and eliminate noise in the gray image. Window size (SW) of 15×15 and *k*=-0.2 are recommended by Trier and Jain [19].

3.2.2. Sauvola's Algorithm

Sauvola's algorithm [15] is a modification of Niblack's which is claimed to give improved performance on documents in which the background contains light texture, big variations and uneven illumination. In this algorithm, a threshold is computed with the dynamic range of the standard deviation, *R*, using the equation:

$$T = m * (1 - k\left(1 - \frac{s}{R}\right)) \quad (2)$$

where m and s are again the mean and standard deviation of the whole window and k is constant. Values of R=128 and k=0.5 are used.

3.2.3. Wolf's Algorithm

To address the issues in Sauvola's algorithm, Wolf et al [5] propose to normalize the contrast and the mean gray value of the image and compute the threshold as:

$$T = (1 - k) * m + k * M + k * s/R(m - M) \quad (3)$$

where *k* is fixed to 0.5, *M* is the minimum gray value of the image and *R* is set to the maximum gray value standard deviation obtained over all the local neighborhoods (windows).

This method in most cases outperforms its predecessors. However, degradation is observed in its performance if there is a sharp change in background gray values across the image. This is due to the fact that the values of *M* and *R* are calculated from the whole image. So even a small noisy patch could significantly influence *M* and *R* values, thus eventually calculate misleading binarization thresholds.

3.2.4. Feng's Algorithm

Instead of calculating dynamic range of gray-value standard deviation from the whole image like [14], Feng et al [2] propose calculating it locally introducing the notion of two local windows, one contained within the other. The values of local mean m, the minimum gray-level M, and standard deviation s are calculated in the primary local window while the dynamic range standard deviation R_s is calculated in the larger window termed as " secondary local window ". Binarization threshold is then computed as:

$$T = (1 - \alpha_1) * m + \alpha_2 * \left(\frac{s}{R_s}\right) * (m - M) + \alpha_3 * M \quad (4)$$

where ,

$$\alpha_2 = k_1(\frac{s}{R_s})^\gamma \quad (5)$$

$$\alpha_3 = k_2(\frac{s}{R_s})^\gamma \quad (6)$$

Based on the experimental experiences of authors, γ is set to 2 while the values of other parameters, α_1, k_1 and k_2 are proposed to be in the range 0.1-0.2, 0.15-0.25 and 0.01-0.05 respectively. This method addresses well the *R*-problem in the Wolf's algorithm. However, the introduction of three parameters leaves the robustness of this method questionable.

3.2.5. Nick's Algorithm

Nick [23] is an improvement in the Niblack algorithm with an advantage that it improves binarization for "white" and light page image by

shifting down the binarization threshold. Binarization threshold is then computed as:

$$T = m + k\sqrt{(\sum p_i{}^2 - m^2)/_{NP}} \qquad (7)$$

where k is the Niblack constant and m is the mean of gray value. pi is the pixel value of gray image and NP is the number of pixels. The value of k can vary from -0.1 to -0.2 depending upon the application requirement. k closes to -0.2 make sure that noise is all eliminated but characters can break a little bit, while with values close to -0.1, some noise pixels can be left but the text will be extracted crisply and unbroken. So, for OCR, the value of k must be set at -0.1 and in application where we don't desire any noise, k should be -0.2.

4. Assessment Parameters

For performance evaluation of binarizaiton algorithms, three assessment parameters are used, which are as follows:

4.1. Mean Square Error (MSE)

MSE is one of many ways to quantify the difference between the original image and binarized image. The MSE represents the average of the square of the errors between the original image and the binarized image.

4.2. Signal to Noise Ratio (SNR)

SNR is calculated as the ratio of average signal to average noise power.

4.3. Peak Signal to Noise Ratio (PSNR)

The measure of PSNR relies on word length of an image pixels and it is calculated as the ratio of peak signal power to average noise power.

5. Input Data Discussion and Experimental Results

5.1. Input Data Discussion

In this work, a set of 150 document images is collected from (old book named " khwe yue " by Dr. Tin Myint was published in 1968, at Gardian Press), (old book named " khit pyaung taw lan yae tha mine win meint khun myar - No. 2" was published in 1969, at Sar Pay Beint Man Press) and (old book named " Myanmar kyae taw nay toe ei ka byar myar " by Mya Than Tint was published in 1975, at Mya Mya Win Press) which are of about 50 years old.

Figure 1. Example of input image

Typical noisy document images having non-uniformly distributed noise are presented in Figure 1. The background of the input images is golden brown in color. It is converted into gray image before applying filters shown in Figure 2.

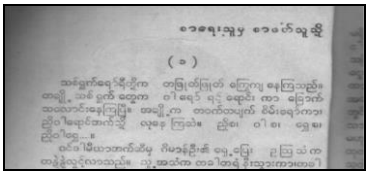

Figure 2 . Gray image of input image

In Figure 3 and 4, the gray image is applied two filters – Wiener and Gaussian Low Pass Filter both of which are the most efficient for aged printed Myanmar documents [3].

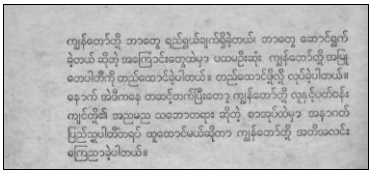

Figure 3 . Wiener filtered image

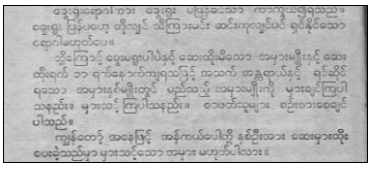

Figure 4 . Gaussian filtered image

5.2. Experimental Results

The experimental for this work is carried using MATLAB (R2013a) (8.1.0.604) with the following configuration: Intel ® Core ™ i7-3770 CPU @3.40GHz, 4.00 GB RAM, 32 bit OS, Windows 7 Ultimate.

5.2.1. Results of Gray Image

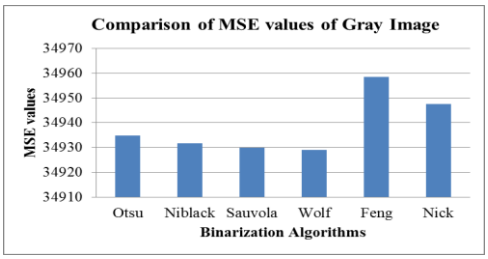

Figure 5. Comparison of MSE values of Gray Image

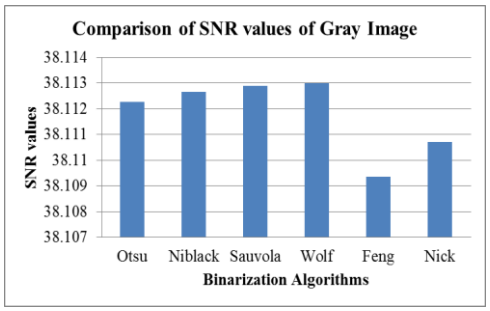

Figure 6 . Comparison of SNR values of Gray Image

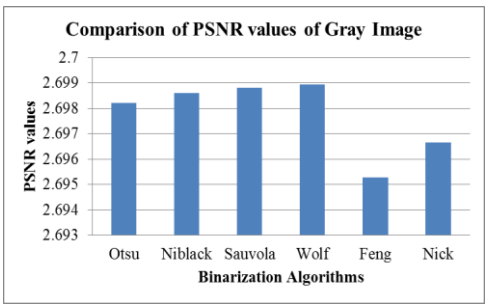

Figure 7. Comparison of PSNR values of Gray Image

The comparison of MSE values of gray image applying these six binarization algorithms is shown in Figure 5. The smaller the MSE values, the better the result of binarization. Among these six algorithms, the MSE values of Otsu, Niblack, Sauvola, and Wolf algorithms seem differ in the small amount.

The comparison of SNR and PSNR values of gray image are shown in Figure 6 and 7. The larger the values of SNR and PSNR, the smaller the amount of noise in this image. Like MSE values, the values of SNR and PSNR of Otsu, Niblack, Sauvola, and Wolf algorithms differ in the small amount.

5.2.2. Results of Wiener Image

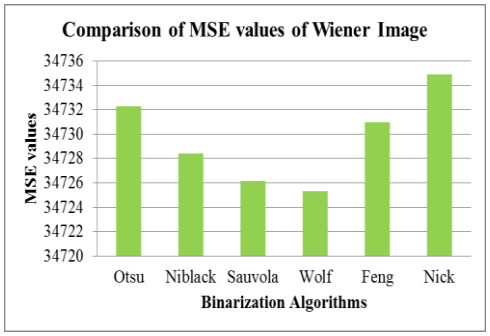

Figure 8. Comparison of MSE values of Wiener Image

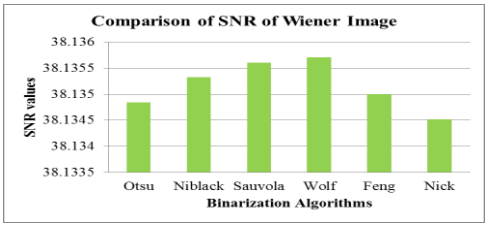

Figure 9 . Comparison of SNR values of Wiener Image

In Figure 8, 9 and 10, the comparison of MSE, SNR and PSNR values of Wiener filtered image are shown respectively. The values of these six algorithms are not different much. However, the MSE value of Wolf algorithm is the smallest one. In addition, the other two assessment parameter values of Wolf algorithm are the greatest.

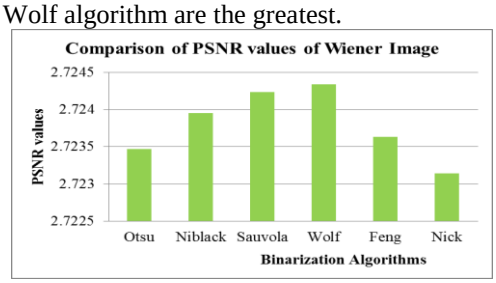

Figure 10. Comparison of PSNR values of Wiener Image

5.2.3. Results of Gaussian Image

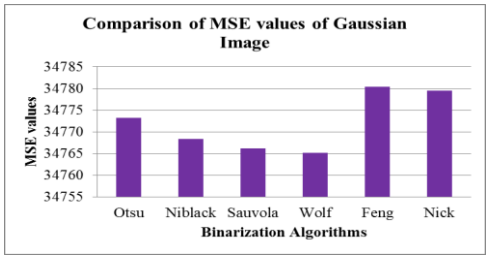

Figure 11. Comparison of MSE values of Gaussian Image

The comparison result of MSE, SNR and PSNR values of gaussian filtered image is shown in Figure 11, 12 and 13. Like the result of Wiener filtered image, Wolf algorithm gives the better result.

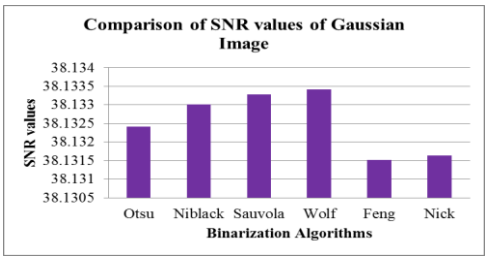

Figure 12. Comparison of SNR values of Gaussian Image

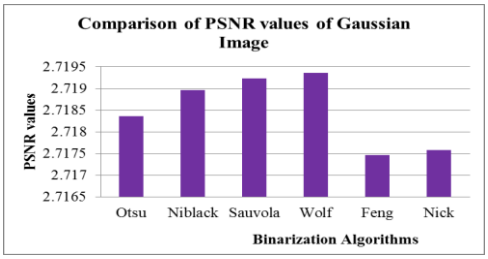

Figure 13 . Comparison of PSNR values of Gaussian Image

According to the results of these three types of images (gray, Wiener filtered and Gaussian filtered), although Wolf algorithm gives the better result, the other two (Niblack and Sauvola) algorithms also give the result that is not very different. By this experiment, when the input image is gray, the value of MSE is greater than other two input types. Hence, the prior process of binarization, filtering process, gives the better result.

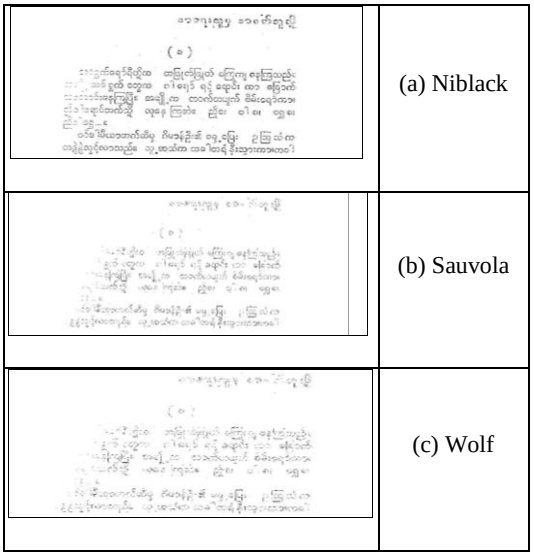

Figure 14. Some output images of input Gray image

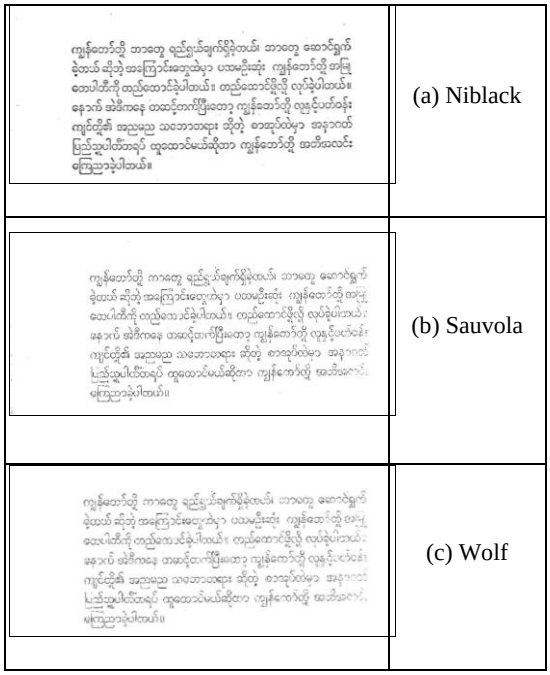

Figure 15 . Some output images of input Wiener image

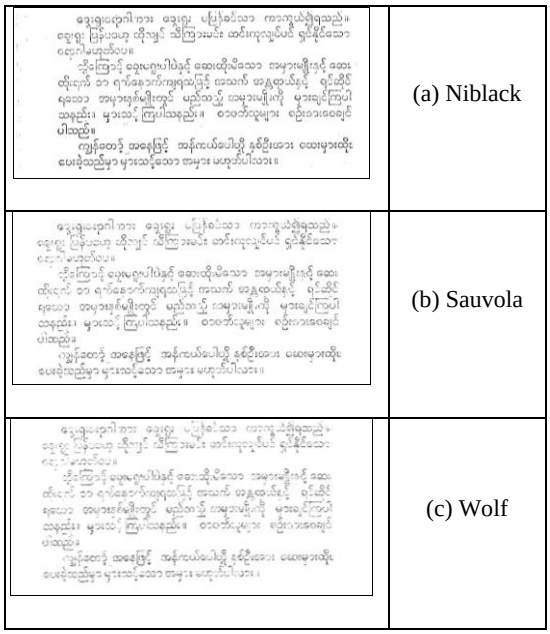

Figure 16. Some output images of input Gaussian image

According to the assessment parameter results of Wolf algorithm, it is the better one. But, according to evaluating by visual human, there is no best algorithm. In Figure (14), (15) and (16), the resultant image of Niblack algorithm is the better one. By the result of these three figures, the filtered images give the better binarization results.

Comparing the results of binarized gray image and binarized filtered image, the second one is better. And then, comparing the results of global thresholding and local thresholding, it is found that results of local one have adjusted the output among local areas so that characters will appear to have more stable appearance than global one. Among local thresholdings, the result of Niblack is better than others.

6. Conclusion and Future Work

In this work, we make an analysis of binarization algorithms. Calculating of performance results have been obtained by using 150 aged printed Myanmar documents. According to the values of assessment parameters, we have observed that if filtering step is applied before binarization step, the result is more stable. And then, the next observation is that local thresholding is suitable for aged printed Myanmar documents. Note, however, that these results apply only to the document images degraded with aging, some non-uniform illumination and some ink-bleed noise. On the other hand, documents degraded with other noises are outside of the scope of this comparison.

As a prospect for the future, the result of this analysis will be used in the digitization process of preservation of aged Myanmar documents and in the optical character recognition of aged printed Myanmar documents.

References

[1] A.Antonacopoulos, Karatzas D., "The Lifecycle of a Digital Historical Document: Structure and Content", *DocEng*, 2004.

[2] A.S.Abutaleb, "Automatic thresholding of gray-level pictures using two-dimensional entropy", *Comput Vis. Graph. Image Process*, 1989.

[3] A.S.Phyo, Dr.K.N.N.Tun, "Analysis of filters for the improvement of previous printed Myanmar documents binarization", *Sixth ICSE* 2015.

[4] B.Gatos, I.Pratikakis, "Adaptive degraded document image binarization", *Pat.Recog*, 2006.

[5] B.M.Singh, R.Sharma,"Adaptive Binarization of severely degraded and non-uniformly illuminated documents", *Springer, IJDAR*, 2014.Doermann, K.Tombre (eds.), Handbook of Document Image Processing and Recognition, Springer-Verlag London 2014.

[6] E.Badekas, N.Papamarkos, "Estimation of appropriate parameter values for document binarization techniques", *IJRA*, 24(1), 2009.

[7] H.S.Baird, "Difficult and urgent open problems in document image analysis for libraries", *1st IDIAL, (*2004).

[8] I.K.Kim, D.W.Jung, P.H.Park, "Document image binarization based on topographic analysis using a water flow model", *Pattern Recognition*, 2002.

[9] J.Bersen, "Dynamic thresholding of grey-level images", *ICPR's86*, 1251-1255 (1986).

[10] J.He, Q.K.M.Do, A.C.Downton, J.H.Kim, "A comparison of binarization methods for Historical archive documents", *IEEE, ICDAR*, 8, 2005.

[11] J.Kittler, J.Illingworth, "Minimun error thresholding", *Pattern Recognition*, 1986.

[12] J.N.Kapur, P.K.Sahoo,"A new method for gray-level picture thresholding using the entropy of the histogram", *Comput Vis. Graph. Image Process*, 1985.

[13] J.Wen, S.Li, J.Sun, "A new binarization method for non-uniform illuminated document images", *Pattern Recognition*, 46, 1670-1690(2012).

[14] J.Sauvola, M.Pietikainen, "Adaptive document image binarization", *Pattern Recognition*, 2000.

[15] M.L.Feng and Y.P.Tan, "Contrast adaptive binarization of low quality document images", *IEICE Electron*, vol. 1, No. 16, 501-506(2004).

[16] M.S.Sonawane, Dr.C.A.Dhawale, "Evaluation of Thresholding Algorithms for Document Images", *IJSR*, 2013.

[17] N.Otsu, "A threshold selection method from gray level histograms", *IEEE Trans. Syst. Man Cybern*, 9, 62-66 (1979).

[18] O.D.Trier, T.Taxt, "Evaluation of binarization methods for document images", *IEEE Trans, Pattern Anal. Mach. Intell. 17(3)*, 312-315(1995).

[19] P.J.Burt,"Recovery of distorted document images from bound volumes", *Sixth ICDAR*, 2001.

[20] R.F.Mogahaddam,"Low quality document image modeling and enhancement", *IJDAR*, 2009.

[21] R.Goyal, A.Kaur, "A review of optimal binarization techniques on documents with damaged background", *IJCST*, 2011.

[22] V.Nicole,K.Khurram, "Comparison of Niblack inspired binarization methods for ancient documents".

[23] W.Niblack, *An Introduction to Digital Image Processing*. Prentice Hall, Englewood Cliffs, 1986