Thresholding Techniques for Ancient Document Images

Jyoti[1] and Dr. Ajit Kumar[2]

[1]Department of Computer Science, Punjabi University, Patiala, India
[2]M.M. Modi College, Patiala, India

Abstract: Binarization of an image is a method of separating pixel values from dual collections, black as foreground and white as background. Thresholding found to be a well-known technique used for binarization of document images. It is the starting step of all the document image analysis and refers to the conversion of gray scale image to binary image. There are different thresholding techniques, but there is no way to select single or best method which is used for all documents. The main purpose of this paper is to present the study on various existing thresholding techniques and compared their results.

Keywords: Thresholding, Ancient Document, OCR.

I. INTRODUCTION

Thresholding is performed in pre-processing stage of an image and refer to distinguish foreground image from background. Thresholding is usually performing either using local, global, hybrid techniques. It converts an image of up to 256 grey levels to black and white image. Thresholding techniques are used as a text locating. The simplest way to get an image binarized is to choose a threshold value, and classify all pixels with the value above the threshold. The threshold should be selected in such a way that it retains most of the text information & suppresses the background.

Although thresholding is studied for many years but still the degraded document image problem is unsolved due to the high inter/intra-variation between the text stroke and the background across different images. Binarization for camera based images were analysed Images are degraded due to aging, smearing and smudging of text, seeping of ink due to other side of page, variation in contrast and illumination of image. Text extraction [1] is basically a system that receives input in the form of a still image or a sequence of images and the output is only text part of the image, which is recognised by Optical character recognition (OCR). A survey of binarization techniques for non-destructive testing images and document images was given by different researchers.

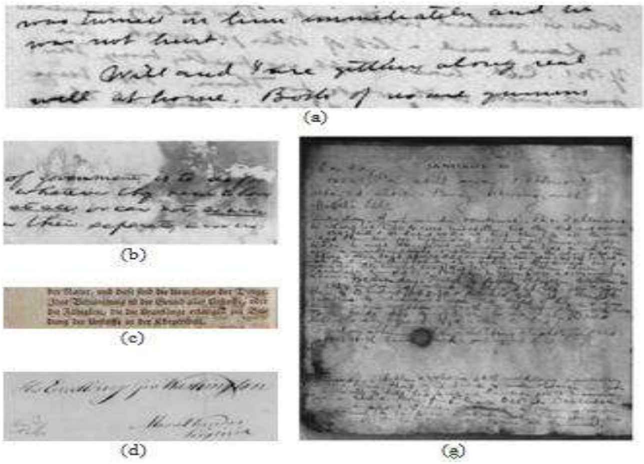


Fig. 1. Historical and degraded document image

Figure (1) shows the degraded images. Figure (a) is a handwritten degraded image shows variation in terms of stroke connection and stroke brightness and image background. Figure (b) shows the damaged background due to passage of time. Figure (c) is handwritten document. Figure (d) shows the effect the low contrast. Where the ink seeps to the other side. In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in Fig. 1(e).

II. ANCIENT DOCUMENT

Ancient document collections are valuable resources for human history. There is a huge collection of ancient documents that have invaluable knowledge about the history, culture and religion of a particular region. Typically, only small groups of people are allowed access to such collections, because the preservation of the material is of great concern. These documents have deteriorated due to age and lack of preservation facilities. This leads to the problems like varying contrast, uneven background, and strain on document, shadow paging, front and back merging. Thus these document need to be enhanced for improving their quality for readability and future references. Computer technology can aid in preserving the knowledge contained in these documents by storing these documents in multimedia format for future reference and through internet, these rare documents will be available to large number of interested individuals.

III. CHALLENGES OF RESEARCH

Ancient documents are difficult to binarize due to the challenges like:

- Shadow Effect
- Damaged Background
- Back And Front Merging
- Strains On Document
- Low Contrast
- Handwritten Document

IV. PARAMETERS

Parameters used to compare results: Precision rate, recall rate and F-measure.

Precision rate is the probability that retrieved document is relevant. High precision means that everything returned was a relevant result. It can be evaluated as:

$$\text{Precision rate} == \frac{TP}{TP+FP} \qquad (1)$$

TP is a number of ink pixels correctly classified as ink and FP is number of pixels that are the part of paper, but are misclassified as ink)

Recall rate is the probability that a relevant document is retrieved in a search. It is calculated as the fraction of correct instances from all instances that actually belongs to the relevant subset, high recall rate means you have not missed anything but you may have a lot of useless result to sift through. It can be evaluated as:

$$\text{Recall rate} = \frac{TP}{TP+FN} \qquad (2)$$

(TP is a number of ink pixels correctly classified as ink and FN number of ink elements classified as paper)

F-measure is the combines of precision rate and recall rate and is the harmonic mean of precision and recall, or balanced F-score:

$$F = 2\,\frac{Precision\ *Recall}{Precision\ +Recall} \qquad (3)$$

V. TECHNIQUES

A number of techniques are proposed by different researcher to solve the problem of image degradation. Thresholding methods are categorized into different groups according to the information they are exploiting.
A survey of binarization techniques for non-destructive testing images and document images was given by Sezgin and Sankur[2] in which various techniques were divided into six categories: Measurement Space Clustering, Histogram shape, Entropy, Object Attributes, and local gray-level surface. A method proposed by Otsu used clustering analysis based method. Method proposed by Johansen et al. and Kapur et al. are entropy based. Sauvola et al. and Ni black are based on image variance. Kittler et al. consider error measure in calculating the optimal threshold.
Different methods for thresholding are discussed below:

A. *Fixed Thresholding Method*

In Fixed Thresholding binarization method [3] fixed threshold value is used to assign 0's and 1's for all pixel positions in a given image. The basic formula for this method is described as under.

$$g(x,y) = 1 \ if \ f(x,y) >= T \qquad (4)$$
$$0 \text{ otherwise}$$

Where is global threshold value. For different threshold value results are illustrated below.

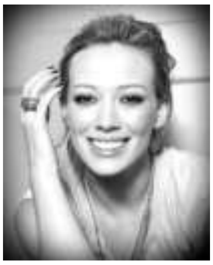


Fig. 2. ORIGINAL IMAGE

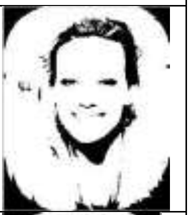
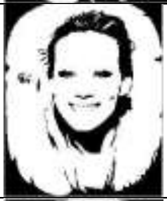
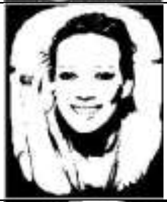
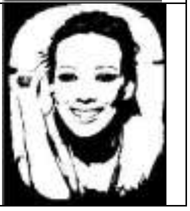
THRESHOLD VALUE	OUTPUT
T=110	
T=129	
T=138	
T=158	

Table 1

As shown in Table 1, outputs vary as given an input threshold value. From the table it can be noticed that output at threshold T=158 is better for this image. But T=158 is not optimal threshold value for all the images. So it is very difficult to choose an optimal threshold value for a input image. To overcome this difficulty we will discuss many more thresholding techniques from where we will calculate optimal threshold value of input image.

B. *Otsu Method*

Otsu [4] Method, a non parametric and unsupervised method of automatic threshold selection for picture segmentation. Histogram of an image represents object and background with a graph describing deep and sharp valley between two peaks respectively as to select the threshold at bottom of this valley. Where as in real picture, it is difficult to detect the valley bottom precisely, when the valley is flat and broad, imbued with noise, or when peaks have extremely unequal heights. To solve this type of problem, some techniques are proposed such as:

(i) Valley sharpening
(ii) Difference histogram method
(iii) Apply directly to histogram.

However, such methods need unstable calculation and there are no criteria of calculating "goodness" of the threshold. In this select an optimal threshold from the 3 discriminate criterion : mainly by maximizing the discriminate measure, η (total variance of level) as it is independent of k (level) as other two are dependent and $\boldsymbol{\sigma_B^2}$ based on first order variance (class mean) and $\boldsymbol{\sigma_W^2}$ second order variance(class variance).

This maximization is selected in sequential search by using the simple cumulative quantities i.e. only zero and first order cumulative moments of gray level histogram are used and the range of 'k' is fixed. Proposed method is used for analyzing the further important aspects as it can be used as a measure to evaluate the separability of classes for the original extension to multi thresholding problem is feasible by virtue of the criteria on which this method is based. This method leads to stable and automatic selection of threshold based on integration of the histogram. As a result this method is recommended as the most simple and standard method.

$$\sigma^2_{within} (T) = \omega_B (T) \sigma^2_B + \omega_o (T) \sigma^2_o(T) \quad (5)$$

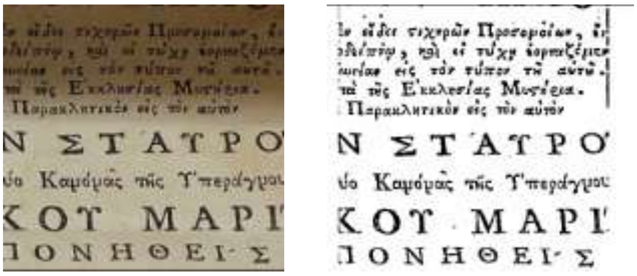
$$\omega_B (T) = \sum_{i=0}^{T-1} p(i) \quad (6)$$

$$\omega_o (T) = \sum_{i=T}^{L-1} p(i) \quad (7)$$

[0, L-1] range of intensity level

$\sigma^2_{B=}$ the variance of pixels in the background (below threshold)

σ^2_o (T) = the variance of pixels in the foreground (above threshold)



(a) (b)

Fig. 3. (a) original image, (b) Image with Otsu method.

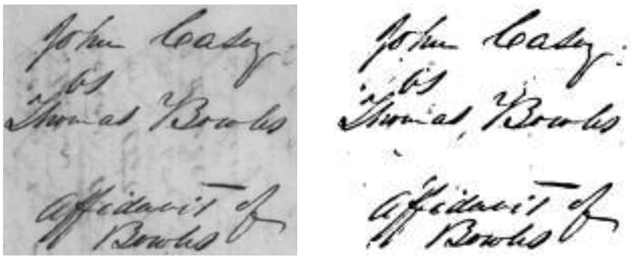
C. Minimum Error Thresholding

A computationally efficient solution [5] to the problem of minimum error thresholding is derived under the assumption of object and pixel grey level values being normally distributed. The method is applicable in multithreshold selection. We explain the minimum error thresholding originated by Kittler and Illingworth using relative entropy. Let us consider an image whose pixels assume grey level values, g, from the interval [0, n].It is convenient to summarise the distribution of the grey levels in the image in the form of a histogram h(g) which gives the frequency of occurrence of each grey level in the image.

The histogram can be viewed as an estimate of the probability density function p(g) of the mixture population is composed of grey levels of object and background pixels. In the following we shall assume that each of the two components p (g|i) of the mixture is normally distributed with mean μ_i standard deviation σ_i and a priori probability P_i, i.e.

$$p (g) = \sum_{i=0}^{n} Pi\, p(g|i) \quad (8)$$

$$p (g|i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\frac{(g-\mu)}{2\sigma 2i} \quad (9)$$



(a) (b)

Fig. 4. (a) original image, (b) Binary image

D. J.Kitter And J.Illingworth

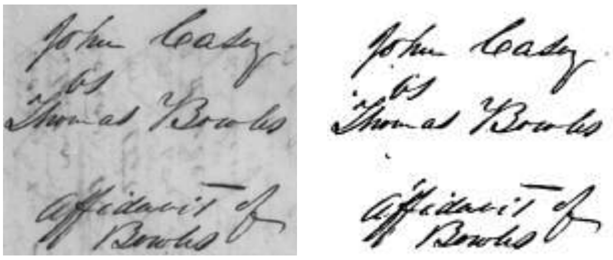
It is an efficient solution to the problem of minimum error thresholding under the assumption of object and pixel gray level values being normally distributed. Basically, this method is alternative solution to the nagwa and rosenfeld as they distribute the population in mean and variance normally and then population parameter is inferred from grey level histogram by fitting.

J.Kitter And J.Illingworth [6] approach is computationally involved as it require optimization of "goodness" of fit" criteria function by hill climbing, so the method is derivate of simpler technique for finding optimal threshold τ (Bayes minimum error). It minimizes the criteria instead of minimizing error threshold selection as it gets the amount of overlapping between object and the background population. The smaller the overlap between density function, smaller the classification error. Therefore, the value of threshold yielding the lowest value of criteria will get the best-fit model and therefore minimize the error. This minimization of criteria gives a threshold value for segmenting the square from the background. In case of non-uniform illumination where, Otsu method fails and results in salt and pepper noise, this method uses variable thresholding techniques.

By using adaptive window, an optimal threshold is selected for each window. It use bilinear interpolation of defining each pixel threshold and show the binary image, so this is the local method of binarizarion. The method is applicable in multi thresholding selection. Minimization of criteria can also be done by iterative approach, by using dynamic clustering algorithm, which is to calculate bayes minimum error rule value, and then using those values in computing the threshold value. If this threshold value is equal to old threshold value, then terminate otherwise go to the iterative process. At last, it proves that a general strategy is to run the threshold selection algorithm for several initial threshold and compare the results.$p_B(t)$ and $p_F(t)$, the p_{mix} (t) mixture of these two Gaussian distribution.

$$p_{mix} (t) = \alpha p_B (t) + (1-\alpha)p_F(t) \quad (10)$$

Where α is determined by the portions of background and foreground in the image.



(a) (b)

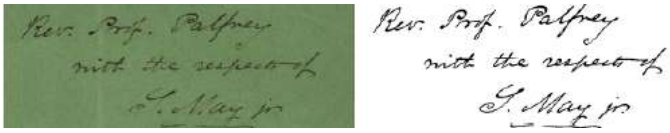
Fig. 4. (a) Original image, (b) Image with J.Kitter And J.Illingworth.

E. Niblack Method

Niblack [7] is a method which calculates the result based on local mean and local standard deviation. The threshold of the pixel (x, y) is calculated by the following formula:

$$T(x, y)=m(x, y) + k*s(x, y) \qquad (11)$$

Where T is the threshold, m(x, y) is the average mean, s(x, y) is the local standard deviation and k is constant. The size of the local window is large enough to suppress the degraded noise and small enough to preserve the local image detail. A window of size 15-by-15 is perfect. To adjust the percentage of total pixel belong to foreground object in the boundaries is depend on the value of k.



(a) (b)

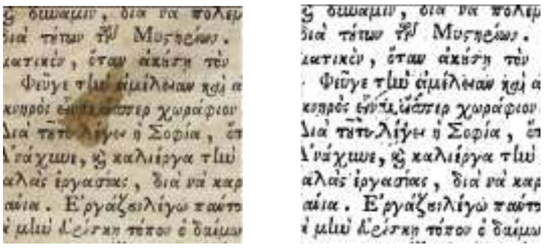
Fig. 5. (a)Original image, (b) Ground truth.

F. Sauvola Method

Sauvola method [8] is a adaptive thresholding method. The calculation of thresholding for each pixel is calculated using local Mean and Standard Deviation of sub-image with dynamic range of standard deviation.

$$T(x, y) = m(x, y) [1+k \{1-\frac{s(x,y)}{R}\}] \qquad (12)$$

Where k and R are constants.



(a) (b)

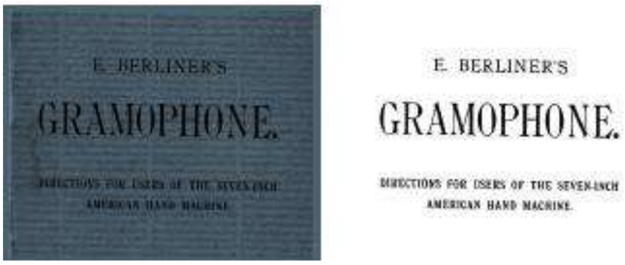
Fig. 6. (a) original image, (b) Binary image with Sauvola.

G. Bernsen method

This method is a local binarization method which computes the threshold value from the pixel of image [9]. Below equation used to calculate the threshold:

$$TBernsen= (N_{low}+ N_{high})/2 \qquad (13)$$

Where N_{high} and N_{low} are the gray level values of the window.



(a) (b)

Fig. 7. (a) original image, (b) Binary image.

H. B Gatos Method

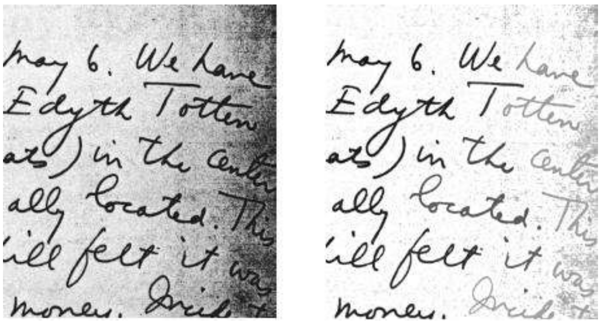
Thresholding of the document by new approach which is a combination of Thresholding methodologies and a various factors:

1.) Efficient pre-processing is been performed by using Weiner filter to remove noise and make the image smooth.

2.) Thresholding that is performed by applying different Thresholding techniques on the image such as local global and many more. These Thresholding result are combined to produce binary result.

3.) Edge information of grey level image is combined with the binary result of previous step. From these they only select that probably belong to text area according to criteria and smoothing algorithm in order to fill text area.

Enhancement [10] of image by mathematical morphology. Comparison is made with the other algorithm and found that it has 91.9% f-measure while other have less than 90%, so it proves the new adaptive technique for Thresholding and degraded documents.



(a) (b)

Fig. 8. (a) original image, (b) Binary image

I. T Kasar Method

T Kasar [11] had given a degraded document image. A contrast map from gray scale is constructed and degraded image is taken and converted into the grey scale image. And create contrast map. Wiener filter is also used. A pre-processing stage of source image is essential for historical and degraded documents for the elimination of noisy areas, smoothing of background texture as well as contrast enhancement between background and text areas. Then edge width value is evaluated. Some post-processing is further applied to improve the document binarization quality.

J. *Extreme Value Theory*

This method [12] presents a thresholding method that can deal with degradations such as shadows, low-contrast, large signal-dependent noise, non-uniform illumination, smear and strain. A pre-processing method based on morphological operations is applied to suppress light/dark structures connected to image border. It is based on difference of gamma functions .Next Generalized Extreme Value Distribution (GEVD) is a method used to find actual threshold with a significance level. This paper emphasizes on region of interest (with the help of morphological operations) and generates less noisy artifacts (due to GEVD). It is simple than other methods and works on degraded documents and natural scene images.

K. *Binarization for Camera Based Images*

This paper [13] presents a binarization method for camera based natural scene (NS) images based on edge analysis and morphological dilation. Image is converted to grey scale image and edge detection is carried out using canny edge detection. The edge image is dilated using morphological dilation and analyzed to remove edges corresponding to non-text regions. The image is binarized using mean and standard

Deviation of edge pixels. Post processing of resulting images is done to fill gaps and to smooth text strokes. The algorithm is tested on a variety of NS images captured using a digital camera under variable resolutions, lightening conditions having text of different fonts, styles and backgrounds. The results are compared with other standard techniques.

L. *Image Extraction Using Steerable Directional Filters*

Automatic text line extraction is one of the processes to analyze these documents in which influence the accuracy of text recognition. The best approach for text line extraction is proposed by The Centre for Unified Biometrics and Sensors (CUBS) [14]. It uses the concepts of CUBS approach, it basically extract text lines from the historical document images. It is based on the three local connectivity maps. One has the orientation angles of the text lines, and it is generated by dynamic steerable directional filter and map is modified by using a mode filter to determine the paragraph map in the documents. Based on these values, the adaptive local connectivity map (ALCM) is generated by the static steerable directional filter to estimate the location of the line. This approach solves the problem of the ALCM binarization that the CUBS approach has used, and gives the advantage of the extracting the text in the document besides the text lines segmentation.

M. *Binarization Using Bit-plane Slicing*:

This method [15] performs document image thresholding using bit-plane slicing. According to this approach, from gray scale image we extract the 8 bit planes and processed separately, at the end, the results are combined to give the final output. This method efficiently work for all kinds of historical documents except those with both, less noise and low contrast.

N. *A recursive Otsu thresholding method*

This method [16] presents two types of variations. One combines a recursive extension of Otsu thresholding and selective bilateral filtering to allow automatic binarization and segmentation of handwritten images. Another is based on recursive Otsu method and also adds improved background normalization and a post-processing algorithm to become robust and to perform adequately even for the images that present bleed-through artifact. Resultant of these technique segment the text in historical documents comparable to and in few cases better than many state-of-the-art approaches based on their performance as evaluate using the dataset from the recent ICDAR 2009 Document Image Binarization Contest.

VI. *Acknowledgement*

We would like to take this opportunity to express our profound gratitude and deep regard to my Project Guide Prof- Ajit Kumar, for his exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project. His valuable suggestions were of immense help throughout my project work. His perceptive criticism kept me working to make this project in a much better way. Working under him was an extremely knowledgeable experience for me.

VII. *Conclusion*

This paper has focused on the different Thresholding technique. The main objective of this paper is to evaluating the thresho lding techniques for degraded image binarization. It has been found that each technique has its own benefits and limitations; no technique is best for every case. In near future we will propose algorithm which will use more reliable methodology to enhance the work.

REFERENCE

[1] Bawa, Rajesh K., Sethi, Ganesh K. *A review on binarization algorithms for camera based natural scene images, ICACCI'12* Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 2012, pp. 873-878.

[2] Sezgin M., Sankur B. *Survey over image thresholding techniques and quantitative performance evaluation,* Journal of Electronic Imaging Vol. 13 (1), 2004, pp. 146-165.

[3] M.I. Sezan, *A peak detection algorithm and its application to histogram-based image data reduction*, Computer Vision, Graphics, and Image Processing 49 (1) (1990) 36–51.

[4] Otsu N. *A threshold selection method from gray level histograms,* IEEE Transactions on Systems, Man, and Cybernetics (SMC-9), 1979, pp. 62-66.

[5] Gonzalez R. C. and Woods R.E. *Digital Image Processing,* Third Edition, 2009, pp. 763-764

[6] Kittler J. and Illingworth J. *Minimum error thresholding,* Pattern Recognition, Vol 19(1), 1986, pp. 41-47.

[7] Niblack W. *An Introduction to Image Processing,* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[8] Sauvola J. and Pietikainen M. *Adaptive document image binarization*, Pattern Recognition, Vol 33(2), 2000, pp. 225-236.

[9] Madhuri Latha, Chakravarthy, *An Improved Bernsen Algorithm Approaches For License Plate Recognition*, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) ISSN: 2278-2834, ISBN: 2278-8735. Volume 3, Issue 4 (Sep-Oct. 2012).

[10] Gatos B., Pratikakis I. And Perantonis S.J.*Text Detection in Indoor/Outdoor Scene Images*, First International Workshop on Camera-based Document Analysis and Recognition (CBDAR'05), 2005, pp. 127-132.

[11] Kasar T., Kumar J. and Ramakrishnan A.G. *Font and Background Color Independent Text Binarisation*, Camera Based Document Analysis and Recognition (workshop of ICDAR), 2007, pp. 3-8.

[12] Fernando B. and Karaoglu S. *Extreme Value Theory Based Text Binarization In Documents and Natural Scenes* The 3rd International Conference on Machine Vision (ICMV 2010), 2010, pp. 144-151.

[13] Bawa, Rajesh K., Sethi, Ganesh K. *A Binarization technique for extraction of devanagri text from camera based images*, Signal & Image Processing: An International Journal Vol. 5 Issue 2, 2014, pp.29.

[14] Z. Shi, S. Setlur, and V. Govindaraju, *A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines*, pp. 176-180, 2009.

[15] Y. Zhu, C. Wang, and R. Dai, *Document image binarization based on stroke enhancement*, in International Conference on Pattern Recognition, 2006, pp. 955-958.

[16] Otsu N. *A threshold selection method from gray level histogram. In: IEEE Transactions on Systems, Man, and Cybernetics 9 (1978); no. 1. p. 62-66.*