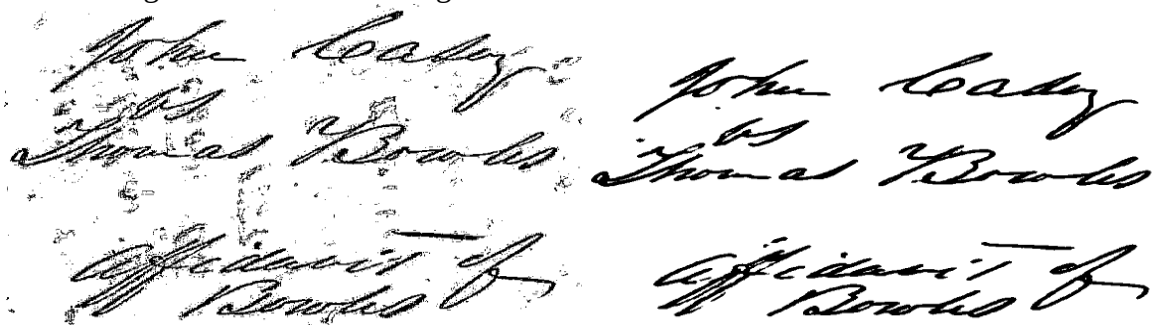
A Proposed Binarization Technique on Hand written document

Navdeep Kaur . Research Scholar, Guru Kashi University, Talwandi Sabo(Bathinda)
Monica Goyal . Assistant Professor, Guru Kashi University, Talwandi Sabo(Bathinda)

Abstract: Binarization is performed in the preprocessing stage for document inspection. Binarization of degraded document images improve the result from poor quality of the paper, the printing process, ink blot and fading document and remove noise from examine. In recent years, libraries have begun to digitize historical document that are of interest to a wide range of people, with the goal of preserving the content and making the documents available via electronic media. But for historical documents suffering from degradation due to damaged background, stained paper, holes and other factors, the recognition results drop appreciably. A degraded documented image contrast based on binarization technique that is tolerant for dissimilar types of document degradation such as uneven enlightenment and document smear.

1. Introduction

Binarization is performed in the preprocessing stage for document inspection. Binarization of degraded document images improve the result from poor quality of the paper, the printing process, ink blot and fading document and remove noise from examine. Every historical document is often degraded by the drain through, where the ink of the other side seeps through to the front. But, document binarization is removing the noise and improves the quality of historical documents. In document with uniform contrast delivery of background and foreground. Binarization is pre-processing technique which consist of separate foreground and background of document image. It convert a gray-scale document image into a binary document image. Document image binarization is an important step in the document image inspection. Image binarization technique is important for the ensuing document image operating task. Document image binarization technique is perform the important role to improve the quality of the degraded document. Document image binarization technique is important for the ensuing document image processing task such as optical character recognition(OCR).Image binarization is the process of separation of pixel values into dual collections, black as foreground and white as background. Thresholding has created to be a well known technique used for binarization of document image. The thresholding of the degraded document image to still found to be a challenging task because of the high inter/intra variation between the text stroke and the document background across various document images. Thresholding is divided into the global thresholding and local thresholding.



Input image Binarized image

ISSN: 2278 – 1323

Generally, we can use three major types: global thresholding, local thresholding and hybrid thresholding.

2. Problem Formulation

The problem undertaken for the paper is *"A Proposed binarization technique on hand written document"***.** Libraries and Museums contain extensive collections of ancient historical documents printed or handwritten in native languages. Typically, only a small group of people are allowed access to such collections, because the preservation of the material is of great concern. In recent years, libraries have begun to digitize historical document that are of interest to a wide range of people, with the goal of preserving the content and making the documents available via electronic media. But for historical documents suffering from degradation due to damaged background, stained paper, holes and other factors, the recognition results drop appreciably.
These recognition results can be improved using binarization techniques. Binarization techniques can distinguish text from background. The simplest way to get an image binarized is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The problem arises, how to select the correct threshold. The selection of threshold is performed by two methods: Global, Local.

3. Techniques

1. **Global thresholding:**

Global thresholding technique need few computations and can work well in simple cases. but global thresholding technique is fails in complex backgrounds, like as non-uniform color and poor illuminated backgrounds. These techniques are usually not suitable for degraded document image, because they do not have a clear pattern that separates foreground text and background. Otsu's method [14] is the best known and most widely used thresholding technique. This method mainly assumes that images have two normal distributions with similar variances. The threshold is selected by partitioning the image pixels into two classes at the gray level that maximizes the between-class scatter measurement. Let the pixels of a given image be represented in L gray levels 1, 2, . . . ,L. The number of pixels at level i is denoted by ni, and the total number of pixels by N = n1 + n2 + . . . + ni + ... Then suppose that the pixels were dichotomized into two classes C0 and C1, which denote pixels with levels 1, .., k and k + 1, . . . ,L, respectively.
This method search for the gray level k which maximizes the between-class variance. The between class variance is defined as:

$$\sigma^2_B = \omega_0\, \omega_1\, (\mu_1 - \mu_0)^2 \qquad (2)$$

Where:

- $\omega_0 = \sum_{i=1}^{k} (\boldsymbol{P(i)}$ is the frequency or the probability of the first class occurrence and P(i) = ni/N is the statistical frequency distribution of level i;
- $\omega_1 = 1 - \boldsymbol{\omega_0}$ is the frequency or the probability of the second class occurrence;
- $\mu_0 = \frac{\sum_{i=1}^{k} (\boldsymbol{iP(i)}}{\boldsymbol{\omega_0}}$ is the mean of the first class;
- $\mu_1 = \frac{\sum_{i=k+1}^{L} (\boldsymbol{iP(i)}}{\boldsymbol{\omega_1}}$ is the mean of the second class.

Otsu's method yields good binary results. However, if the text intensities are inseparable from the background intensities, Otsu's method finds an unsuitable threshold value.

ISSN: 2278 – 1323

2. **Local thresholding:**

Local thresholding technique select the target pixels depends into local information. This technique is applying only selected pixels. Local thresholding methods calculate a threshold for each pixel on the basis of the information contained in a neighborhood of the pixel as in global it is for entire image. If a pixel (x, y) in the input image has a higher gray level than the threshold surface evaluate at (x, y) and set to white, otherwise black. Some methods in this class are histogram-based, such as the method of Kopf et al., which after some pre-processing, estimate the dominant text colors based on histograms computed on four luminance bits. They analyze each block between two separators and classify the pixel as text or background to produce the binary image, based on a region-growing algorithm and the colors previously determined. [21]

These types of approaches are window-based, which means that the local threshold for a pixel (x, y) is computed from the gray values of the pixels in a window centered at (x, y). Many researchers proposed various techniques to compute the local threshold based on the minimum and the maximum gray values in each window, some are based on the mean and the standard deviation as follows:

$$T = m + k.s \qquad (3)$$

Where,

- T is the threshold for the central pixel of a rectangular window which is shifted across the image,
- m is the mean,
- s is the variance of the gray values in the window.
- k is a constant.

One drawback of this approach is noise created in areas which do not contain any text, due to the fact that a threshold is created in these cases as well [9].

For solving this type of problem a similar method is proposed as Sauvola method. The researcher assumed text pixel to have gray values near 0 and background pixels to 255. This assumption results, formula for the threshold:

$$T = m \times (1 - k \times (1 - s/R)), \qquad (4)$$

Where, R is the dynamic range of the standard deviation. [12]

This method gives better results for document images than method used by Nibalck , et.al. The performance of all these algorithms depends on the window size, so a suitable window size for a document image is heavily affected by the character size and thickness. If window size is large, the details of the character structure are degraded, otherwise, if window size is too small, the window centered at an inner pixel produces a large mean and a small standard deviation, as most pixels in that window belong to the foreground.

As a result, the local threshold for such a pixel is likely to be assigned a large value, and therefore, the central areas of thick and black regions were often classified into background. Method has been proposed based on three windows to solve above problem. The first window is used for a pre-processing binarization for the analyzing. The second and the third windows are used to determine the actual local thresholds.

Local binarization methods calculate a threshold $t(x, y)$ for each pixel such that

$$\mathbf{b(x,y)} = \begin{cases} \mathbf{0} & \boldsymbol{if\ g(x,y) \le t(x,y)} \\ \mathbf{255} & \boldsymbol{otherwise} \end{cases} \qquad \mathbf{(5)}$$

The threshold $t(x, y)$ is computed using the mean $\mu(x, y)$ and standard deviation $\sigma(x, y)$ of the pixel intensities in a $w \times w$ window centered around the pixel (x, y) in Sauvola's binarization method[6]:

ISSN: 2278 – 1323

$$t(x,y) = \mu(x, y)\left[1 + k\left(\frac{\sigma(x,y)}{R} - 1\right)\right] \qquad (6)$$

Where R is the maximum value of the standard deviation (R = 128 for a grayscale document), and k is a parameter which takes positive values. The formula (Equation 6) has been designed in such a way that, the value of the threshold is adapted according to the contrast in the local neighborhood of the pixel using the local mean $\mu(x, y)$ and local standard deviation $\sigma(x, y)$. So, it estimate appropriate threshold $t(x, y)$ for each pixel under both conditions: high and low contrast. If local high contrast region ($\sigma(x, y) \approx R$), the threshold $t(x, y)$ is nearly equal to $\mu(x, y)$. If quite low contrast region ($\sigma << R$), the threshold goes below the mean value
The parameter k controls the value of the threshold in the local window such that the higher the value of k, the lower the threshold from the local mean $m(x, y)$. [10]

3. **Hybrid thresholding:**

Hybrid thresholding technique is combines the local and global thresholding techniques. Firstly apply the global thresholding technique and then secondly apply Local thresholding only selected pixels. As local thresholding techniques, compute the value of threshold at each pixel in a particular window with some of a set of rules.Whereas, in hybrid method more than one approaches are joined together to compute a better thresholding value which is suitable for the images where these local and global method's are not much efficient.
As an example, M.Valizadeh proposed a contrast independent binarization algorithm that effectively eliminates background and reliably extracts some parts of each character by applying some modification to binarization algorithm based on water flow model. This method extracted the edge pixel and SW of the character. Edge detection has been done by canny edge detector as it can extract efficiently weak edges and to measure SW they uses run length histogram. Beside all advantages, this method produces broken character that reduces the efficiency of character recognition algorithms. Dealing with this problem, they have combined binarization algorithm with the Niblack's method which complements algorithm. The main idea is to select the better algorithm in each part of document image. Since Niblack's method effectively distinguishes the text from the background in the areas closed to the text, researcher use proposed algorithm to find these areas and binarize them by Niblack's method. The regions far from the text are labeled as background by proposed algorithm. After extensive experiment, the proposed binarization algorithm demonstrate superior performance against four well-know binarization algorithms on a set of degraded document images captured with camera. [15], [26].
As a result in hybrid method we mainly using two or more binarization techniques in combination and form the result which are more efficient than the global or local method separately applied.

Methods used:

1. Bernsen's Technique

In Bernsen's technique, the local contrast is defined as follows:

$$C(i, j) = I_{max}(i, j) - I_{min}(i, j) \qquad (14)$$

whereC(i, j) denotes the contrast of an image pixel(i, j), Imax(i, j) and Imin(i, j) denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively.

ISSN: 2278 – 1323

If the local contrast C(i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of Imax (i, j) and Imin(i, j).

2. Algo 2 of modified bernsen's technique

Proposed of a novel document image binarization method by using the local image contrast is evaluated as follows:

$$C(i, j) = Imax(i, j) - Imin(i, j) / Imax(i, j) + Imin(i, j) + Pos \quad \textbf{(15)}$$

Where pos is a positive but infinitely small number that is added in case the local maximum is equal to 0. Compared with Bernsen's contrast in Equation 1, the local image contrast in Equation 2 introduces a normalization factor (the denominator) to compensate the image variation within the document background.

3. Algo 3 again modified bersen's technique

However, the image contrast in Equation 2 has one typical disadvantage that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small.

4. Otsu Method Algorithm

Otsu Thresholding method, as proposed in [1], is based on discriminate analysis. In this method,the threshold operation is regarded as the partitioning of the pixels of an image into two classesC0 and C1, (e.g., objects and background) at gray level t.

5. Sauvola Algorithm

Sauvola method introduced by Jaakko Sauvola [7] is an efficent image binarization technique which works well for cases in which background contains light texture as gray values of unwanted details easily exceed threshold value which was the problem in niblack method.

6. Proposed Techniques

In our technique we have consider the base paper of Bolan su who has implemented a technique of his known as Bolan Su by his name. We have proposed with new implementation as follows:

Step 1: Load the images on which we have to apply the proposed technique.

Step 2: First we apply the contrast Image Construction on loaded image by using formula mentioned below:

$$C_a(i,j) = \frac{\alpha C(i,j) + (1-\alpha)\left(I_{max}(i,j) - I_{min}(i,j)\right)}{\left(I_{max}(i,j) + I_{min}(i,j)\right)} \quad (32)$$

Where,

- $\left(I_{max}(i,j) + I_{min}(i,j)\right)$, This equation is named as normalized in which 0 or 1 is considered.
- $\alpha = \left(\frac{Std}{128}\right)^{\gamma}$, $\gamma = 2^{-10}$

ISSN: 2278 – 1323

Step 3: Now, we compute the canny edge detection by using threshold value i.e 0.4 and width is approximate 10 to 12.

Step 4: After this, we applied ostu on step 2 and gets the result in Binary map.

Step 5: Canny edge detection and step 4 results values are if same means 0 (Zero) then keep those values else discard the values.

Step 6: Now we considered combined Binary map resulted after step 5 in which if values founds as Zero (edge), we precede original image intensity else if not zero then we consider as 1 (background). New image is constructed with edge intensity known as E.

Step 7: We now compute Edge mean as $\boldsymbol{Emean}$ and Edge Standard Derivation $\boldsymbol{EStd}$ from edge image E.

Step 8: We compute the Local threshold by using below given formula,

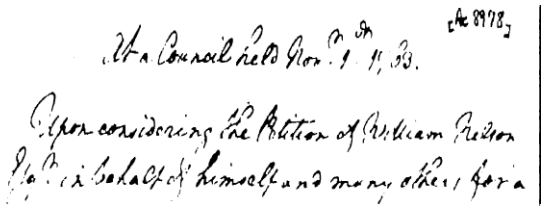
$$S(x,y) = \begin{cases} 1 \Big| I(x,y) \leq \frac{Emean}{4} + Estd * 3 \\ otherwise \end{cases} \quad (33)$$

Step 9: At last, we applied Post Processing technique for removing smaller objects area which is less than 15.

In our proposed work Bulan Su method has been applied with standard estimation Emean divided by 4 and standard derivation multiplying with three. In Post processing, we used to find connected components and their area, on which we applied threshold value i.e fixed 0.4 and area less than 15 has been removed for all standard dibco 2009 images.

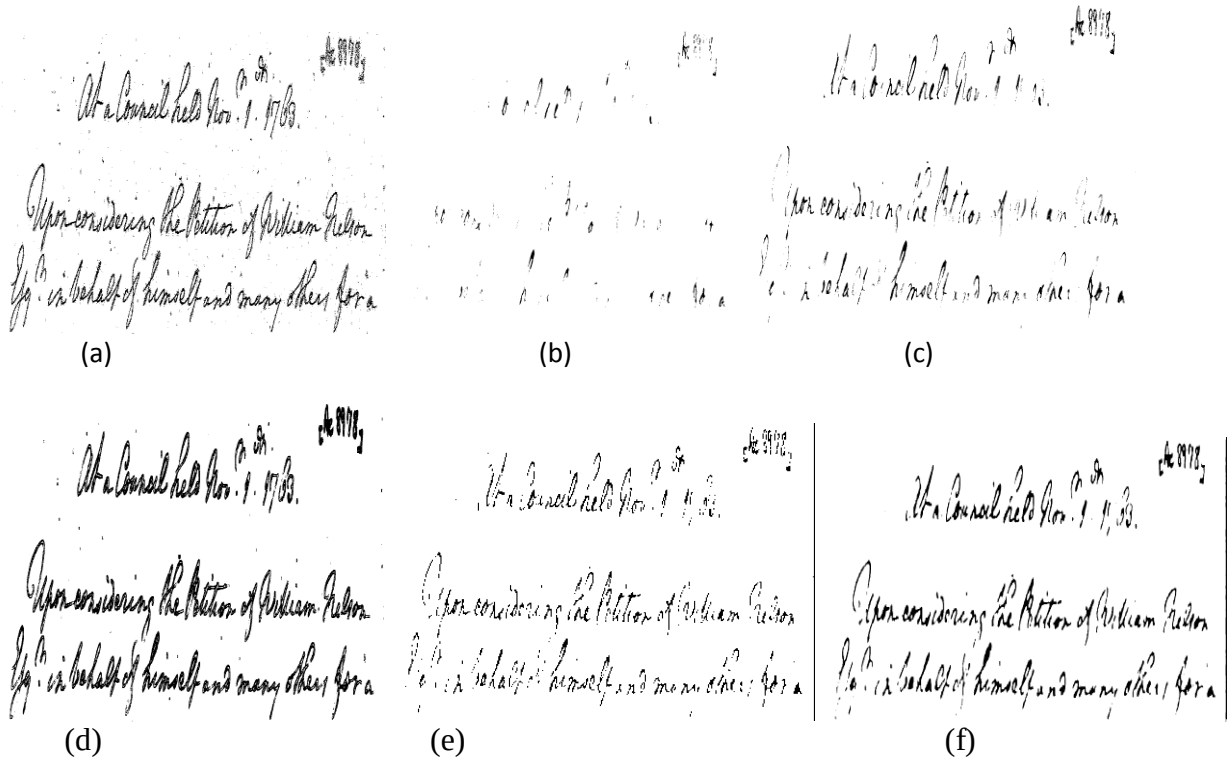
4. Results and Conclusion

1. Handwritten Original Images:



2. Experimental Results:

ISSN: 2278 – 1323



(a) (b) (c)

(d) (e) (f)

Comparison results of handwritten first dataset image (a) p01bersen (b) p01 bersen1modi (algo1) (c) p01 bersen2modi (algo2) (d) p01 bersen2otsu (e) p01sauvola1 Techniques (f) Proposed Technique

Table (Result of Handwritten image dataset)

Techniques	Precision	Recall	Fmeasure	Sensitivity	Specificity	Accuracy
Bersen	0.79194829	0.847009323	81.85539126	0.847009323	0.985155169	0.976515389
Bersen1 Modi	0.10401719	1	18.84340073	1	0.93964849	0.940068394
Bersen2 Modi	0.38721361	0.998257528	55.79891114	0.998257528	0.957919378	0.958965977
Bersen2 Otsu	0.80312641	0.058617575	10.92605778	0.058617575	0.842371094	0.12409436
Sauvola1	0.57377561	0.91026064	70.38714204	0.91026064	0.970235199	0.967706486
Proposed technique	0.7220027	0.768979456	74.47510257	0.768979456	0.980158892	0.966896192

ISSN: 2278 – 1323

In above table shows the different parameters that is calculated from the above figure. In this table different techniques are applied for quality and quantity calculation. Proposed technique results are best as compared to other techniques as shown in above mentioned parameters such as , precision, F-measure etc.

Conclusion

Development of a system for recognizing any fonts and characters based on handwritten Documents are presented in this paper. A survey of general methodology in recent research is summarized at the beginning of this report. We conclude that this paper has contributed to the field of handwritten documents for degradation. A degraded documented image contrast based on binarization technique that is tolerant for dissimilar types of document degradation such as uneven enlightenment and document smear. The proposed technique is simple enough in which only few parameters are involved. Moreover, it works on different i.e handwritten documents of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets that experimentally shows the proposed method outperforms best than previous document binarization methods in term of the performance evaluated by using Parameters such as PSNR, Precision, Recall, F-Measure, accuracy, Sensitivity, Specificity. . It has been found that each technique has its own benefits and limitations; no technique is best for every case.

6. REFERENCES

- Bolan Su, el.at, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL **2013.**
- Jagroop Kaur, Dr.Rajiv Mahajan, "A Review of Degraded Document Image Binarization Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, **2014**.
- Y.H. Chiu et al. "Parameter-free based two-stage method for binarizing degraded document images" in Pattern Recognition 45 (2012), 4250–4262. Ying Liu and Sargur N. Srihari. "Document Image. Binarization Based on Texture Features." IEEE. Trans. PAMI, Vol. 19(5), PP: 540 - 544, **1997**.
- Vikul J.Pawar, "An Improved Binarization Technique for Degraded Document Images using Local Thresholding Method", IJPRET, Vol: 2 (8), PP: 71-82, **2014**.
- B.Gatos, I. Pratikakis and S.J. Perantonis, "Adaptive Degraded Document Image Binarization", Pattern Recognition, Vol. 39(3), PP: 317 – 327, **2006**.
- B. Gatos, I. Pratikakis and S.J. Perantonis, "An adaptive binarization technique for low quality historical documents", IARP Workshop on Document Analysis Systems, Lecture Notes in Computer Science (3163), PP: 102 - 113, **2004**.
- J.J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen, "Adaptive Document Binarization", In International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, PP.147 – 152, **1997**.

ISSN: 2278 – 1323

- J. Kittler and J. Illingworth, "Minimum Error Thresholding," Pattern Recognition, Vol. 19(1), PP: 41 - 47, **1986**.
- Joao Marcelo Monte da Silva, Rafael Dueire Lins, "A New And Efficient Algorithm To Binarize Document Images Removing Back-To-Front Interface", JUCS, Vol. 14(2), PP: 229 - 313, **2008**.
- Jaap Oosterbroek, Dr. Marco A. Wiering, Dr. Michael H.F. Wilkinson, "Using Max-Trees with Alternative Connectivity Classes in Historical Document Processing", **2012**.

ISSN: 2278 – 1323