International Journal of Image Processing and Vision Science

Volume 1 | Issue 2 Article 2

October 2012

HYBRID BINARIZTION TECHNIQUE FOR HISTORICAL MANUSCRIPTS

Rajan Goyal
D.C.S.E Department, Punjabi University, Patiala,India, er.rajangoyal@gmail.com

Amandeep Kaur
D.C.S.E Department, Punjabi University, Patiala, India, aman_@gmail.com

Recommended Citation

Goyal, Rajan and Kaur, Amandeep (2012) "HYBRID BINARIZTION TECHNIQUE FOR HISTORICAL MANUSCRIPTS," *International Journal of Image Processing and Vision Science*: Vol. 1 : Iss. 2 , Article 2.
DOI: 10.47893/IJIPVS.2012.1014
Available at: https://www.interscience.in/ijipvs/vol1/iss2/2

HYBRID BINARIZTION TECHNIQUE FOR HISTORICAL MANUSCRIPTS

Rajan Goyal
Mtech Student, D.C.S.E Department
Punjabi University,
Patiala,India
er.rajangoyal@gmail.com

Amandeep Kaur
Assistant Professor, D.C.S.E Department
Punjabi University,
Patiala,India
Aman_@gmail.com

***Abstract*— This paper presents a new hybrid approach for the binarization and enhancement of Historical Manuscript. This paper deals with degradations which occur due to shadows, non-uniform illumination, low contrast and strain. We follow two distinct method of Binarization with a pre-processing procedure using a adaptive Wiener filter, a rough estimation of foreground regions and a background surface calculation by interpolating neighboring background intensities. Further logical anding of the calculated background surface with compliment of second method result, performing final thresholding and post-processing in order to improve the quality of text regions. After extensive experiments, our method demonstrated superior performance against some well-known techniques on numerous degraded document images as well as on Historical Manuscript in both manners qualitatively and quantitatively.**

Keywords- Historical manuscripts; Binarization

I. INTRODUCTION *(HEADING 1)*

Binarization plays a key role in degraded document processing. Its performance affects quite effectively in the degree of success in character segmentation and recognition. When processing degraded document images, binarization is not an easy task. Actually, image binarization refers to the conversion of a gray-scale image into a binary image. It is the initial step of most document image analysis. Usually, it distinguishes text areas from background areas, so it is used as a text locating technique. Mainly, degradations appear frequently and may occur due to several reasons such as the appearance of variable background intensity caused by non-uniform intensity, strains, shadow effects and low contrast. Hybrid binaraization is a technique used to improve the quality of the degraded document as well as Historical Manuscripts in a visual manner.

Historical manuscripts collections are a valuable resource for human history. There is a huge collection of historical documents that have invaluable knowledge about the history, culture and religion of punjab. Typically, only a small group of people are allowed access to such collections, because the preservation of the material is of great concern. These document have deteriorated due to age and lack of preservation facilities. 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 avaliable to large number of interested indivisuals.

In general, image processing is very wide area for scientific research. Many algorithms for degraded document image binaraization are available. These algorithms are categorizes in groups according to the information they are exploiting. These categories are: 1. Histogram shape-based methods where the peaks, valleys and curvatures of the smoothed histogram are analyzed. 2. Clustering-based methods where the gray level samples are clustered in two parts as background and foreground (object). 3. Entropy-based methods result in algorithms, for example, that use the entropy foreground-background regions, the cross-entropy between the original and binarized image etc. 4. Object attribute-based methods search a measure of similarity between the gray-level and binarized images, such as fuzzy similarity, shape, edges, number of objects etc. 5. The spatial methods use the probability mass function models taking into account correlation between pixels on a global scale. 6. Local methods do not determine a single value of threshold but adapt the threshold value depending upon the local image characteristics.

General approaches that are mostly used for document image binarization are either global or local. In a global approach, threshold selection leads to a single threshold value for the entire image. Global thresholding [1] has a good performance in the case that there is a good separation between the foreground and the background. Unlike global approaches, local area information may guide the threshold value for each pixel in local (adaptive) thresholding techniques [2-4]. These techniques have been widely used in document image analysis because they have a better performance in extracting the character strokes from an image that contains spatially uneven gray levels due to degradations. We mainly considered three groups of algorithm's: Cluster-Based Thresholding, Local Or Adaptive Method and Entropy-Based Method [5].

Our main focus is to digitized the historical Manuscripts by applying Hybrid Binarization Algorithm to suppress the background noise and to retain text information without any distortion for the multimedia format and futher image processing.

The proposed scheme consists of six basic steps. The first step is dedicated to a denoising procedure using an adaptive Wiener filter based on local statistics. In the second step, we compute the final image by entropy based method and also we obtain rough estimation of foreground regions. Next, as a third step, we compute the background surface of the image by interpolating neighboring background intensities into the foreground areas that result from the previous step. In the fourth step, we proceed to the image by logical anding of the Compliment of entropy based method image with the calculated background surface estimation. Further applying final thresholding and post-processing to obtain the resultant image. The proposed method has been extensively tested with a variety of degraded image documents and Historical manuscripts and has demonstrated superior performance against well-known techniques.

The paper is organized as follows. Section 2 briefly reviews the state-of-the-art with particular emphasis on local adaptive methods and entropy based method used during our experiments for comparison purposes. In Section 3, our methodology is described in detail while in Section 4 we discuss our experimental results. Finally, conclusions are drawn in Section 5.

II. Related Work

In literature, the global thresholding technique today named as Otsu method [1]. A non-parametric and unsupervised method of automatic threshold selection for picture segmentation. A method to select an optimal threshold form the 3 discriminant criterion: mainly by maximizing the discriminant measure η (total variance of level) as it is independent of k (levels) as other two are dependent and based on first order variance (class mean) and second order variance (class variance). This maximization has been 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. This method leads to stable and automatic selection of threshold based on integration of the histogram. As a result this method has been recommended as the most simple and standard method, while for the local methods (adaptive thresholding), local area information guides the threshold value for each pixel [7,8]. Most of the image binarization algorithms rely on statistical methods, without taking into account the special nature of document images.

Algorithms that are considered as the current state-of-the-art. These algorithms have been used for the comparison and evaluation of our approach. Niblack [2] introduced an algorithm that calculates a pixel wise threshold by shifting a rectangular window across the image. The threshold *T* for the center pixel of the window is computed using the mean *m* and the variance *s* of the gray values in the window

$$T = m + ks, \tag{1}$$

Where, *k* is a constant set to −0.2. The value of *k* is used to determine how much of the total print object boundary is taken as a part of the given object. This method can distinguish the object from the background effectively in the areas close to the objects. The results are not very sensitive to the window size as long as the window covers at least 1–2 characters. However, noise that is present in the background remains dominant in the final binary image. Consequently, if the objects are sparse in an image, a lot of background noise will be left.

Sauvola and Pietikainen [3] propose a method that solves this problem by adding a hypothesis on the gray values of text and background pixels (text pixels have gray values near 0 and background pixels have gray values near 255), which results in the following formula for the threshold:

$$T = m + (1 - k(1 - s/R)), \tag{2}$$

Where, *R* is the dynamics of the standard deviation fixed to 128 and *k* takes on positive values (usually set to 0.5). This method gives better results for document images.

The Shannon entropy[5] parametrically dependent upon the threshold value *T* for the foreground and background is formulated as:

$$H_f(T) = -\sum_{g=0}^{T} p_f(g) \log p_f(g)$$

$$H_b(T) = -\sum_{g=T+1}^{G} p_b(g) \log p_b(g)$$

(3)

$$H(T) = H_f(T) + H_b(T)$$

When the entropy is calculated over the input image distribution p(g) (and not over the class distributions), then obviously it does not depend upon the threshold T and hence is expressed simply as H. For various other definitions of the entropy in the context of thresholding, with some abuse of notation, we will use the same symbols of $H_f(T)$ and $H_b(T)$ and further ROC(Receiver Operating Characteristics are used for improvement of threshold value and result are achieved by measures such as precision, recall, accuracy, F-measure, etc.

III. METHODOLOGY

Proposed methodology for the degraded and Historical manuscripts is described in following steps:

A. Pre-processing

A pre-processing stage of the grayscale source image is essential for the elimination of noisy areas, smoothing of back- ground texture as well as contrast enhancement between background and text areas. The use of wiener filter has been proved efficient for aforementioned goals. It is commonly used for image restoration. We proposed a module of adaptive wiener filter[9] based on statistics estimated from a local neighborhood around each pixel. Filter gray scale image I_F according to following formula:

$$I(x,y) = \mu + \left(\sigma^2 - v^2\right)(I_s(x,y) - \mu)/\sigma^2, \quad (4)$$

Where, μ is the local mean, σ^2 the variance in 5x5 neighborhoods around each pixel and v^2 is the average of all estimated variances of each pixel in the neighborhood.

B. Rough estimation of foreground regions

At this step, we obtain a rough estimation of foreground (text) regions. Our intention is to proceed to an initial segmentation of foreground and background regions that will provide us a superset of the correct set of foreground pixels. This is refined at a later step (Section E). Sauvola's approach for adaptive thresholding [3] using the above equation (2) with k = 0.2 suitable for this case. At this step we process original image to get the binary image S(x,y), where 1's correspond to the rough estimated foreground regions.

C. Foreground regions by entropy based method

At this step, we obtain I_R resultant image formed by entropy method [5] similar to PUN method and ROC curves are used for the adjustment of the threshold value achieved by the entropy algorithm. That technique has the disadvantage of being very slow for the high resolution images used in historical documents applications and its speed is improved in this paper by using the percentage of black. This is a very simple algorithm which works only with the image histogram; so it is very efficient.

$$B(x,y) = \begin{cases} I(x,y) & if\ S(x,y) = 0 \\ \dfrac{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (I(ix,iy)(1-S(ix,iy))}{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (1-S(ix,iy))} & if\ S(x,y) = 1 \end{cases} \quad (5)$$

D. Background surface estimation

At this stage, we compute an approximate background surface *B(x, y)* of the image *I (x, y)*. A similar approach has been proposed for the binarization of camera images [10]. Background surface estimation is guided by the valuation of S*(x, y)* image. For pixels that correspond to 0's at image *S(x, y)*, the corresponding value at *B(x, y)* equals to *I (x, y)*. For the remaining pixels, the valuation of *B(x, y)* is computed by a neighboring pixel interpolation, as described in:

The interpolation window of size d*x*×d*y* is defined to cover at least two image characters for this case it is 5x5. This might occurs when the region of the window is very dark and a division by zero is faced. This special case is faced when:

$$\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (1 - S(ix,iy)) = 0 \quad (6)$$

E. Combining & Final thresholding

At this step, *B(x,y)* combine with I_R image obtained by compliment of entropy based method by logical anding and obtain *B(x,y)* In this step, the final thresholding *T(x,y)* for each pixel at *(x, y)* is computed by combining the calculated background surface ***B(x, y)*** with the preprocessed image ***I(x, y)*.** Text areas are located if the distance of the preprocessed image ***I (x, y)*** from the calculated background ***B(x, y)*** exceeds a threshold ***d (B(x, y))*.**

$$T(x,y) = \begin{cases} 1, & if\ B(x,y) - I(x,y) < 1.5 * d(B(x,y)) \\ 0, & else \end{cases} \quad (7)$$

The threshold 1.5* ***d (B(x, y))*** must change according to the gray-scale value of the background surface ***B(x, y)*** in order to preserve textual information even in very dark background areas. For this reason threshold ***d (B(x, y))*** has smaller values for darker regions.

$$b = \frac{\sum_x \sum_y (B(x,y))(1 - S(x,y))}{\sum_x \sum_y (1 - S(x,y))}. \quad (8)$$

To simulate this requirement, we use the following logistic sigmoid function which is recommended by [8].

$$d(B(x,y)) = q\delta\left(\frac{(1-p_2)}{1+\exp\left(\frac{-4B(x,y)}{b(1-p_1)} + \frac{2(1+p_1)}{(1-p_1)}\right)} + p_2\right). \quad (9)$$

For the case of degraded and poor quality document images, we assumed the values of the parameters q, P1 and P2 to be 0.65, 0.55 and 0.85 respectively.

F. Post-processing

Post-processing step is applied on the resulting binary image in order to eliminate noise, improve the quality of text regions. In Phase I: Shrink filter is used to remove the remaining noise from the background. Each foreground pixel in the image is examined, and an N X N sliding window (5 X 5 window is recommended) around each foreground pixel in the image is selected. If number of background pixels in the sliding window (Psh) is larger than a fixed value (Ksh), then this pixel is changed to background. The above process can be calculated easily using Eq. 10 and Eq. 11.

$$\text{Psh} = \sum_{y=0}^{\text{Window Width}} \sum_{x=0}^{\text{Window Height}} \text{Bg\_Pixel}[x,\ y] \quad (10)$$

$$\sum_{y=0}^{\dim_y} \sum_{x=0}^{\dim_x} \text{Fg\_Pixel}[x,\ y] = \begin{cases} \text{if Psh} > \text{Ksh} & \text{Central\_Pixel} = 0 \\ \text{else} & \text{Central\_Pixel} = 1 \end{cases} \quad (11)$$

In our project a 5 X 5 window is found experimentally to be the most appropriate sliding window and (Ksh) = **v** (**4**x0.6) ^{2}, where **v** is a variable ranges between 0.5 and 1. In Phase II: morphology operation is performed to suppress the noise. We apply structure element square of 3x3 and perform dilation on image obtained from Phase I and obtained the resultant image.

IV. EXPERIMENTS

The proposed algorithm was tested using degraded document images which belong historical handwritten & typed documents. All images have varying resolutions, stroke sizes, illumination contrast, background complexity and noise levels. The historical handwritten documents were selected from the council held up in 1763. All historical handwritten document images are of poor quality and have shadows, non-uniform illumination, smear and strain. In some of these documents, ink sipped through the pages and characters on the reverse side become visible and interfere with the characters on the front side. Example historical handwritten & typed documents used for testing are shown in Figure 1 and 2 respectively.

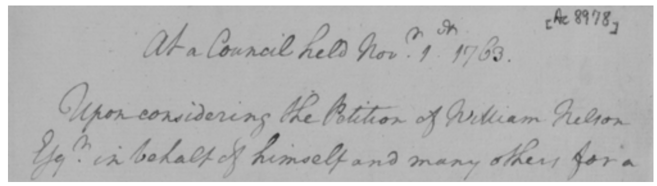

Figure1. Handwritten historical document

Figure2. Typed Historical Document

We compared the performance of our algorithm with few well-known binarization techniques. We evaluated the following: Niblack's adaptive thresholding method [2], Sauvola adaptive thresholding method, entropy based threshold method and adaptive threshold method by B.Gatos [4]. the binarized images are stored in TIFF file format and occupy 70 Kbytes in average. Figure 3 and 4 presents two samples of the application of the algorithm.

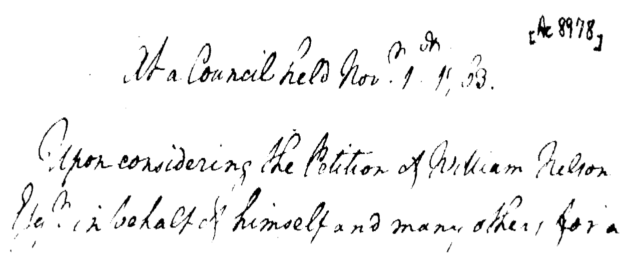

Figure3. Binary Handwritten Historical Document

Figure4. Binary Typed Historical Document

Figure 3 presents the result of the application of the algorithm in the image presented in figure 1. It is one of the documents with low contrast. Figure 4 presents the result of the application of the algorithm in the image presented in figure 2. It is one of the documents with non uniform illumination.

In order to evaluate quantitatively the performance of the algorithm, ‘‘clean’’ images are generated manually by adaptive filter’s. Images are bi-level, with the threshold defined non-automatically by visual inspection. These clean images are used for comparison with the images produced by the thresholding algorithms. We evaluated true positives (TP e number of ink pixels correctly classified as ink), true negatives (TN e number of pixels correctly classified as paper), false positive (FP e number of pixels that are part of the paper, but are misclassified as ink) and false negative (FN e number of ink elements classified as paper). With these values, the following metrics can be found [28]:

$$P = \mathrm{TP} + \mathrm{FN}; \quad N = \mathrm{FP} + \mathrm{TN};$$

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$$

$$\mathrm{Recall} = \mathrm{Sensitivity} = \mathrm{TP}/P;$$

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(P + N); \quad \mathrm{Specificity} = \mathrm{TN}/N$$

Table 1 presents the values of precision, recall, accuracy and specificity for the image of Figure 1 (the one with low contrast) after being cleaned and compared to its version created by every algorithm briefly defined in Section II. The best algorithm must have the greater values of F-measure and accuracy from these four metrics as they are calculated by using precision, recall, specificity.

TABLE I. MEASURES FOR FIGURE 1

Algorithms	Figure 1				
	Precision	*Recall*	*F-measure*	*Accuracy*	*Specificity*
Niblack	0.990	0.564	71.90	0.839	0.997
Sauvola	0.828	0.920	87.22	0.949	0.956
Entropy	0.499	0.999	66.60	0.896	0.884
B.Gatos	0.905	0.810	85.52	0.936	0.970
New Alg	0.953	0.883	91.73	0.964	0.988

Table 2 presents the values of precision, recall, accuracy and specificity for the image of Figure 2 (the one with non uniform illumination) after being cleaned and compared to its version created by every algorithm briefly defined in Section II. The best algorithm must have the greater values of F-measure and accuracy from these four metrics as they are calculated by using precision, recall, specificity.

TABLE II. MEASURES FOR FIGURE 2

Algorithms	Figure 2				
	Precision	*Recall*	*F-measure*	*Accuracy*	*Specificity*
Niblack	0.994	0.314	47.73	0.854	0.997
Sauvola	0.573	0.910	70.38	0.967	0.956
Entropy	0.998	0.636	77.73	0.961	0.884
B.Gatos	0.706	0.769	73.62	0.966	0.979
New Alg	0.810	0.756	78.26	0.969	0.986

We must also analyze precision value equal to 1 for some of the algorithms. It can be noticed that when an image is completely black (as the one created by the application of the algorithms of Brink or Johannsen) the comparison with the cleaned image results that all black pixels are correctly classified. For a real analysis of the performance of the algorithm we must check the other metrics. The new algorithm has a better answer for F-measure and accuracy than any other. As there is not much difference in accuracy where as f-measure have seen differences. Though sauvola has been close difference in figure1 but the visual result are not much accurate as shown in figure 5.

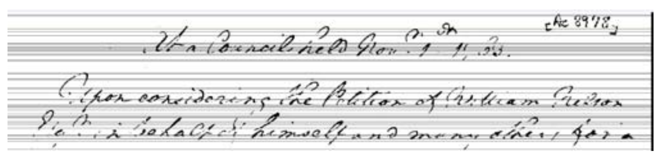

Figure5. Sauvola result for Table 1

V. CONCLUSION

This paper proposed a hybrid binarization technique consist of two method’s ie. Local method and entropy based method and with preprocessing and postprocessing lead to improve the image result as by decreasing the noise at the background. . The proposed method does not require any parameter tuning by the user and can deal with degradations which occur due to shadows, non-uniform illumination, low contrast, and strain.

In order to make these documents more easily available, a conversion of the images from a color to black and white creates files of smaller size. For this a thresholding algorithm was presented to generate bi-level images of historical documents.

Further research will focus on the challenges that merge from the binarization of Punjabi Historical manuscripts found on the Web for digitization and further processing.

REFERENCES

[1] Nobuyuki Otsu, "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man& Cybernetics, Vol. 9(1), PP: 62 - 66, 1979.

[2] W. Niblack., “An Introduction to Digital Image Processing”, N.J.:Prentice Hall, 1986.

[3] 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

[4] B.Gatos, I. Pratikakis and S.J. Perantonis, “Adaptive Degraded Document Image Binarization”, Pattern Recognition, Vol. 39(3), PP: 317 – 327, 2006

[5] Carlos Mello, Angel Sanchez, Adriano Oliveira, Alberto Lopes, “An Efficient Gray-Level Thresholding Algorithm For Historic Document Images”, Journal of Cultural Heritage 9, PP: 109-116, 2008.

[6] Yahia S. Halabi, Zaid SA, Faris Hamdan, Khaled Haj Yousef, “Modeling Adaptive Degraded Document Image Binarization and Optical Character System”, Euro Journals Publishing, Inc., Vol.28 No.1, PP: 14 - 32, 2009.

[7] P.K. Sahoo, S. Soltani, A.K.C. Wong, A survey of thresholding techniques,Computer Vision, Graphics Image Processing 41 (2) (1988) 233–260.

[8] I.-K. Kim, D.-W. Jung, R.-H. Park, Document image binarization based on topographic analysis using a water flow model, Pattern Recognition 35 (2002) 265–277.

[9] A. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989

[10] M. Seeger, C. Dance, Binarising Camera Images for OCR, Sixth International Conference on Document Analysis and Recognition (ICDAR’01), Seattle, Washington, 2001, pp. 54–58.