Journal of Theoretical and Applied Information Technology
20th July 2014. Vol. 65 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENTBINARIZATION TECHNIQUE FOR HISTORICAL DOCUMENT IMAGES

[1]J.RAMYA, [2]B.PARVATHAVARTHINI

[1]Associate Professor, Department of MCA, St. Joseph's College of Engineering, Chennai, India and Research Scholar, JNTU, Hyderabad, India

[2]Professor & Head, Department of MCA, St. Joseph's College of Engineering, Chennai, India

E-mail: [1]ramsharsha@gmail.com, [2]parvathavarthini@gmail.com

ABSTRACT

Binarization is an important preprocessing step in several document image processing tasks. Binarization of historical document with poor contrast, strong noise, and non-uniform illumination is a challenging problem. A new binarization algorithm has been developed to address this problem. In this paper, we describe a new method which does not utilize statistical properties of the intensity histogram of a gray-scale image to determine a threshold. Our method is based on the "Divide and Conquer" strategy for calculating a threshold value which is based on the list of gray levels in that image. It uses low pass Weiner filter method as a preprocessing step to enhance the image by deblurring the image. It uses Median Filter method as a postprocessing step to reduce noise in the image. Our result shows that this new binarization method produce high quality binary image for historical document than any other global methods such as Otsu's method.

Keywords: *Binarization, Divide and Conquer, Historical Document images, Threshold.*

1. INTRODUCTION

Historical documents are documents that contain important information about a person, place or event. Most of that information will be lost in future since they are either printed on ordinary paper or leaf such as palm leaf or papyrus which has a limited lifespan. The main objective of historical document image binarization is the initial step in most of the document image analysis and also which understands the systems that convert a gray scale image into a binary image aims to distinguish text from background. Binarization plays a vital role in document image processing since its performance affects quite critically the degree of success in subsequent character segmentation and recognition.

In document image processing there is a need of generating a binary image from a gray scale or color image through binarization. Say for example, most of the Optical Character Recognition (OCR) methods are utilizing image binarization as the first step. Document images can be obtained through cameras or scanners. With the wide spread usage of digital cameras on smart phones, user can take picture of a document or a product label bearing text. Moreover there are number of challenges that are unique to camera captured document images they are non-uniform illumination, blur, complex background, degradations etc and so on.

Traditionally, Optical Character Recognition [1] is performed on binary images because character recognition on a gray-scale image or a color image does not produce more accurate results. The problem of binarization of historical document consists of transforming a gray-scale image into a two-color black and white image in which individual pixels are marked as "Object" pixels, if their value is greater than the threshold value otherwise it is marked as "Background" pixels. We assume that the white pixels of a binary image form a background while the black pixels represents the text to be analyzed.

There have been many techniques proposed for binarizing these historical document images. These methods utilize statistical properties of the intensity histogram to determine a threshold value. The threshold for binarization can be selected either locally or globally. For the global methods or global thresholding, a single calculated threshold value is used to categorise image pixels into object or background classes [2-6], while for the local thresholding or adaptive thresholding, local area

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information guides the value of the threshold for each pixel [7-8].

The remaining sections of the papers are designed as follows: Section 2 provides the details of related works, Section 3 shows all stages of the proposed methodology and experimental results are given in Section 4. In section 5, enhancement and future work discussed. Finally, conclusions are drawn in Section 6.

2. RELATED WORK

Global binarization techniques are very fast and they provide good results for scanned documents. For several years, the binarization of grayscale documents was established on the global thresholding algorithms. These approaches can be measured as clustering approaches and are also suitable for converting any type of gray scale image into a binary format but those are inappropriate for complex documents, and even more, for degraded documents. If the brightness over the document is not uniform, for instance in cases of camera-captured documents or scanned books, global binarization techniques have a tendency to produce minimal noise along the page borders.

To overcome these difficulties, local thresholding methods have been proposed for document binarization. These techniques estimate a different threshold for each pixel according to the grayscale information of the neighbouring pixels. The techniques available in this category are Bernsen [9] proposed the average of the maximum and minimum pixel values in the neighbouring region as a local threshold method, Eikvil [10] discussed that a clustering is performed locally and a test is used to decide whether the region consists of one or more classes. After the classification the mean values of the foreground and background are updated. Mardia and Hainsworth [11], Niblack [12] combined the mean and standard deviation of pixel values in the sliding window as a local threshold. White and Rohrer [13] uses a running average of the pixel values as the local threshold. Abutaleb [14] determined the local threshold by maximizing the entropy of each local region. Tsai [15] uses a local threshold to preserve image moments.

There is more refined adaptive binarization methods have been proposed to take into picture the unique features of document images. Gatos et. al [16] utilized background removal to deal with non-uniform illumination, using stroke connectivity preservation as a constraint. Sauvola et. al[17] treated a document image as a collection of background, texts and pictures, and binarized text and non-text regions differently. L.O. Gorman [18] and Liu and Li [19], which combines the information of both global and local thresholds, are called as hybrid techniques.

Finally, there are more than 20 global methods available and more than 11 local methods available. Trier and Jain [20] evaluated the performance of these methods and found that Otsu's method [3] is the best global method and Niblack's method [12] and Sauvola's method [17] are the best local adaptive methods. In any event, no "goodness" of threshold has been evaluated in most of the methods so far proposed.

Relating to local thresholding methods, a goal-directed performance evaluation of most of the popular local thresholding algorithms has been executed for map images [20]. According to this evaluation, for a gradually changing background, local methods work well. However, with a complex background, it appeared that none can be tuned up with a set of operating parameters good for all images.

For document image binarization, global thresholding are not enough because document images are usually degraded and having low quality which includes shadows, non-uniform illumination, low contrast, smear and strain. So a new binarization technique has been introduced to recover the degraded image with high quality.

3. A NEW BINARIZATION TECHNIQUE

The existing global method for binarization such as Otsu's method does not produce better result for historical documents due to strong noice, smear or bad illumination. In this paper, we develop a new binarization method for this type of historical document images. The proposed algorithm consists of 3 parts: (1) Preprocessing using low pass Weiner Filter method (2) Binarization using Divide and Conquer strategy (3) Postprocessing using Median Filter method.

Characters in the historical document images are blurred due to high level of smear. Therefore before biniarization, we deblur the images with Weiner Filter method. Weiner deconvolution can be used effectively when the frequency characteristics of the image and additive noise are known, to at least some degree.

The new binarization method presented in this paper is based on the gray levels available in the image. We retrieve the gray levels in the image by using its histogram. This method does not consider the number of pixels in the image. It purely depends upon the various intensity levels

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

available in the image. With the help of the list of gray levels, we find the threshold value by applying the "Divide and conquer" strategy. The steps of the algorithm are enumerated as follows.

Algorithm

1. Extract the list L of distinct gray levels in the image
2. Calculate the mean of the list L
3. Divide the list L into sub list L1 and sub list L2 based on the mean
4. Calculate the mean of right sub list L2
5. Divide the right sub list L2 into L2.1 and L2.2 based on this mean
6. Repeat step 2 to 5 for sub list L2.1 of this right sub list L2 until all the gray levels are processed.
7. Final mean is considered as Optimum Threshold

In this method, we segment the list of gray levels L into two sub list based on the mean of the list. The gray levels in the left sub list L1 is considered as background while the right sub list L2 is again divided into two sub list. The gray levels in the right most sub list L22 is considered as Object or character. The left sub list L21 of the right sub list L2 is processed again until all the gray levels are processed. Finally we will get a two color black and white image in which pixels related to white color is considered as "Background" and pixels related to black is considered as "Object".

The binary image obtained from this new algorithm contains some noise due to strong noise in the historical document images. This noise is reduced by Median filter method. Median Filter method is a noise reduction technique which operates over an m x m region by selecting the median intensity in the region. The below figure 1 shows the flow of the new binarization technique for the historical document images.

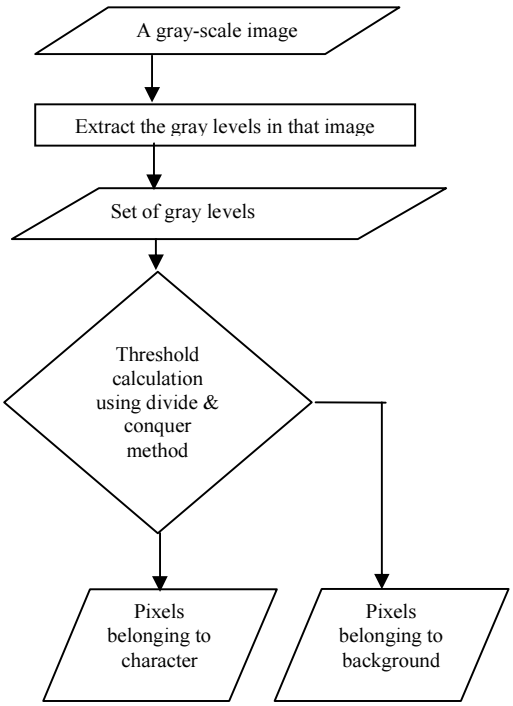

Figure 1: Block diagram of the proposed method

4. EXPERIMENTAL RESULTS

We evaluate our binarization approach on various historical document images such as Palm leaf, Papyrus etc. We select Otsu's method for comparison because it is the best global method. The result of the experiment is shown in table 1. It contains the original gray-scale image, result of Otsu's method and result of the proposed method. It shows that the proposed method is more effective than the Otsu's method for historical documents. Our algorithm also performed significantly better than other binarization algorithms and techniques [21], [22]. Our algorithm maintained low insertion rate because it was able to eliminate the background efficiently.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Original Image, Binarization Output Using Otsu's Method And Binarization Output Using Proposed Method.

Original image Otsu's Method Proposed Method

Comparison of PSNR values of the proposed technique with other best global technique is shown in table 2 and figure 2. The experimental results along with the performance study of our technique prove its effectiveness in the document image binarization. According to these results, the higher the PSNR values, the reconstruction is good. Our proposed method has given the higher values of PSNR like 32.42, 33.83, 30.51, 34.51 and 33.06 for the five samples than the existing method.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Evaluation Results using PSNR

Input images	PSNR Values	
	Otsu's Method	Proposed Method
Sample1	30.64	32.42
Sample2	32.81	33.83
Sample3	29.72	30.51
Sample4	32.5	34.51
Sample5	29.92	33.06

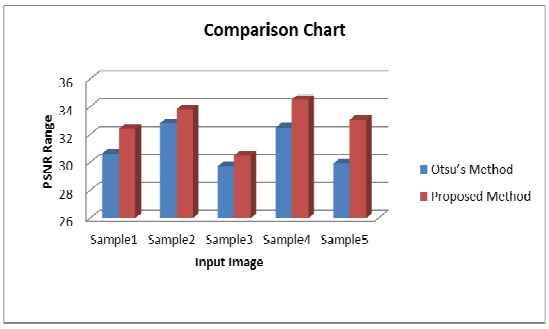

Figure 2: Comparative experimental results

5. CONCLUSION

Historical document images contain lot of smear, low contrast and bad illumination which make it so difficult to recognize the characters. The existing global method such as Otsu's method does not produce accurate results. The proposed method does not depend upon the statistical properties and extract the characters based on the intensity levels. Therefore they provide more efficient results for old Palm leaf images. The advantage of our algorithm is its simplicity and low computational cost. The limitation of the proposed document image binarization is that it does not produce better results for images contain non-uniform illumination and also it does not eliminate the noise completely. Historical and ancient document collections available in libraries throughout the world are of great cultural and scientific importance. The transformation of such documents into digital form is essential for maintaining the quality of the originals while provide scholars with full access to that information. It holds very great value not only in terms of fortune but also significantly to the present way the world is at present. By the way of ensuring that these historical documents are still physically present, one may continue to use these historical document as a reference in making further discoveries about the entire world, and most importantly, in creating necessary actions to make ensure peace, equality and freedom from all over the world. Experimental results show that the algorithm is able to avoid blurring of image and extract object from the old noisy image. Our method can be used for character recognition for historical documents.

REFRENCES:

[1] Trier O.D., Jain A.K., and Taxt T, "Feature extraction methods for character recognition - a survey", *Technical Report, Michigan State Univ.,* 1994.

[2] A. Rosenfeld, A.C. Kak, "Digital Picture Processing", second edition, Academic Press, New York, 1982.

[3] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Systems Man Cybernet,* Vol. 9, No. 1, 1979, pp. 62–66.

[4] J. Kittler, J. Illingworth, "On threshold selection using clustering criteria", *IEEE Trans. Systems Man Cybernet,* Vol. 15, 1985, pp. 652–655.

[5] A.D. Brink, "Thresholding of digital images using two-dimensional entropies", *Pattern Recognition*, Vol. 25, No. 8, 1992, pp.803–808.

[6] H. Yan, "Unified formulation of a class of image thresholding techniques", *Pattern Recognition*, Vol. 29, No. 12, 1996, pp.2025–2032.

[7] P.K. Sahoo, S. Soltani, A.K.C. Wong, "A survey of thresholding techniques", *Computer Vision, Graphics Image Processing*, Vol. 41, No. 2, 1988, pp. 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*, Vol. 35, 2002, pp. 265–277.

[9] J. Bernsen, "Dynamic thresholding of gray-level images", *The 8th International Conference on Pattern Recognition*, 1986, pp. 1251-1255.

[10] Eikvil, L., Taxt, T., and Moen, K. "A fast adaptive method for binarization of document images". *Proc. ICDAR, France*, 1991, pp. 435–443

[11] Mardia, K.V., and Hainsworth, T.J. "A spatial thresholding method for image segmentation", *IEEE Trans. Pattern Anal. Mach. Intell.,* Vol. 10, No. 8, 1988, pp. 919–927.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[12] W. Niblack, "Introduction to Digital Image Processing", *Englewood Cliffs, NJ,* Prentice Hall, 1986.
[13] J.M. White, G.D. Rohrer, "Image thresholding for optical character recognition and other applications requiring character image extraction", *IBM Journal of Research and Development*, Vol. 27, 1983, pp. 400-411.
[14] A.S. Abutaleb, "Automatic thresholding of gray-level pictures using two-dimensional entropy", *Computer Vision, Graphics and Image Processing*, Vol. 47, 1989, pp. 22-32.
[15] W.H. Tsai, "Moment-preserving thresholding: a new approach", *Computer Vision, Graphics and Image Processing*, Vol. 29, 1985, pp. 377-393.
[16] B. Gatos, I. Pratikakis, J.P. Stavros, "An adaptive binarization technique for low quality historical documents", *Lecture Notes in Computer Science*, 3163, 2004, pp. 102-113.
[17] J. Sauvola, M. Pietikakinen, "Adaptive document image binarization", *Pattern Recognition*, Vol. 33, 2000, pp. 225-236.
[18] Gorman, L.O. "Binarization and multithresholding of document images using connectivity", *CVGIP, Graph. Models Image Process*, Vol. 56, No. 6, 1994, pp. 494–506.
[19] Liu, Y., and Srihari, S.N. "Document image binarization based on texture features", *IEEE Pattern Anal. Mach. Intell.,* Vol. 19, No. 5, 1997, pp. 540–544.
[20] O.D. Trier, A.K. Jain, "Goal-Directed evaluation of binarization methods", *IEEE Trans Pattern Anal. Mach. Intell,* 1995, 1191-1201.
[21] Amer Dawoud, Mohamed Kamel "Iterative model-based binarization algorithm for cheque images", *International Journal Document Analysis and Recognition*, 2002, pp 28-38.
[22] Ergina Kavallieratou "A Binarization Algorithm specialized on Document Images and Photos", *8th International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005.