RESEARCH ARTICLE OPEN ACCESS

Binarization of Document Image

G.silpalatha, K.S.Raghavendra Reddy and B.Rajani Kumar Reddy
[1,2,3]Department of ECE, YSR Engineering College of YVU, Proddatur, Kadapa.

ABSTRACT
Documents Image Binarization is performed in the preprocessing stage for document analysis and it aims to segment the foreground text from the document background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as optical character recognition (OCR). Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra variation between the text stroke and the document background across different document images. The handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleed. Documents are often degraded by different types of imaging artifact. These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-art techniques. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm. At the same time, the parameters used in the algorithm can be adaptively estimated.
Keywords **-**Binarization, OCR, Thresholding, Adaptive

I. INTRODUCTION

The recent Document Image Binarization Contest (DIBCO) held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 & 2011 and the Handwritten Document Image Binarization Contest (H-DIBCO) held under the framework of the International Conference on Frontiers in Handwritten Recognition show recent efforts on this issue. We participated in the DIBCO 2009 and our background estimation method [4] performs the best among entries of 43 algorithms submitted from 35 international research groups. We also participated in the H-DIBCO 2010 and our local maximum-minimum method [5] was one of the top two winners among 17 submitted algorithms. In the latest DIBCO 2011, our proposed method achieved second best results among 18 submitted algorithms. This paper presents a document binarization technique that extends our previous local maximum-minimum method and the method used in the latest DIBCO 2011.

Many thresholding techniques [6]–[9] have been reported for document image binarization. As many degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. For example, the early window-based adaptive thresholding techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. The main drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Other approaches have also been reported, including background subtraction, texture analysis recursive method, ecomposition method, contour completion, Markov Random Field matched wavelet [33], ross section sequence graph analysis self-learning, Laplacian energy user assistance and combination of inarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The local image contrast and the local image gradient are very useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques. In Bernsen's paper, the local contrast is defined as follows:

$$C(i,j)=Imax(i,j)-Imin(i,j) \quad (1)$$

where $C(i, j)$ 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. 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 $Imin(i, j)$. Bernsen's method is simple, but cannot work properly on degraded document images with a complex document background. We have earlier proposed a novel document image binarization method [5] by using the local image contrast that is evaluated as follows [41]:

$$C(I,j)=(Imax(I,j)-Imin(I,j))/(\ Imax(I,j)+Imin(I,j)+c) \quad (2)$$

where _ 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. Take the text within shaded document areas such as that in the sample document image in Fig. 1(b) as an example. The small image contrast around the text stroke edges in Equation 1 (resulting from the shading) will be compensated by a small normalization factor (due to the dark document background) as defined in Equation.

II. METHODOLOGY:

CONTRAST IMAGE CONSTRUCTION

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many nonstroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background. In our earlier method, The local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation (2). In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image
difference that is similar to the traditional image gradient.

The denominator is a normalization factor that suppresses the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator and accordingly result in a relatively high image contrast.

However, the image contrast in Equation (2) has one typical limitation 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. To overcome this over-normalization problem, we combine the local image contrast with the local image gradient and derive an adaptive local image contrast as
follows:

$$C_a(I,j)=\alpha C(I,j)+(1-\alpha)(\ Imax(I,j)-Imin(I,j)) \quad (3)$$

where $C(i, j)$ denotes the local contrast in Equation (2) and $(Imax(i, j) - Imin(i, j))$ refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 empirically. α is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large α) when the document image has significant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem of our previous method. We model the mapping from document image intensity variation to α by a power function as follows:

$$\alpha= (std/128)^{\gamma} \quad (4)$$

where *Std* denotes the document image intensity standard deviation, and γ is a pre-defined parameter. The power function has a nice property in that it monotonically and smoothly increases from 0 to 1 and its shape can be easily controlled by different γ .γ can be selected from $[0,\infty]$, where the power function becomes a linear function when $\gamma = 1$. Therefore, the local image gradient will play the major role in Equation (3) when γ is large and the local image contrast will play the major role when γ is small. The setting of parameter γ will be discussed in Section IV. Fig. 1shows the contrast map of the sample document images that are created by using local image gradient, local image contrast and our proposed method in Equation 3, respectively. For the sample document with a complex document background, the use of the local image contrast produces a better result as shown in Fig. 1(b) compared with the result by the local image gradient

as shown in Fig. 1(a) (because the normalization factors in Equation 2 helps to suppress the noise at the upper left area of Fig. 1(a)). But for the sample document that has small intensity variation within the document background but large intensity variation within the text strokes, the use of the local image contrast removes many light text strokes improperly in the contrast map as shown in Fig. 1(b) whereas the use of local image gradient is capable of preserving those light text strokes as shown in Fig. 1(a).

As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3 can produce proper contrast maps for document images with different types of degradation as shown in Fig. 1(c). In Particular , the local image contrast in Equation (3) gets a high weight for the document image with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image.

Fig 1: Contrast images constructed using (a) local image gradient, (b) local image contrast
(c) Proposed method of the sample documents

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images : First, the text pixels are close to the detected text stroke edge pixels. Second, there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$R(x,y)=1$ if $I(x,y)<=E_{mean}+E_{std}/2$ otherwise 0

where *E*mean and *E*std are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window *W*, respectively. The neighborhood window should be at least larger than the stroke width in order to contain stroke edge pixels. So the size of the neighborhood window *W* can be set based on the stroke width of the document image under study, *EW*, which can be estimated from the detected stroke edges [shown in Fig. 3(b)] as stated in Algorithm 1.

Since we do not need a precise stroke width, we just calculate the most frequently distance between two adjacent edge pixels (which denotes two sides edge of a stroke) in horizontal direction and use it as the estimated stroke width. First the edge image is scanned horizontally row by row and the edge pixel candidates are selected as described in step 3. If the edge pixels, which are labeled 0 (background).

III. RESULT

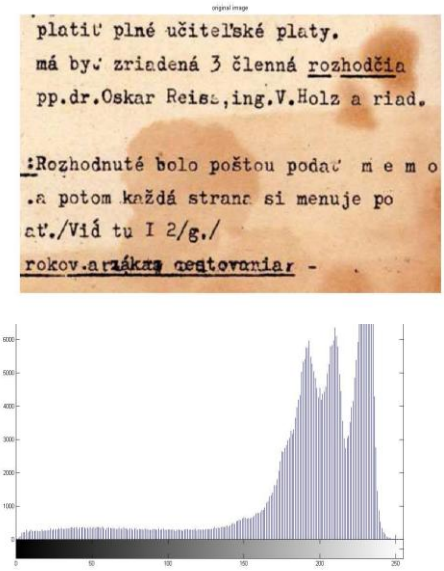

Fig 3: Histogram of Original document image

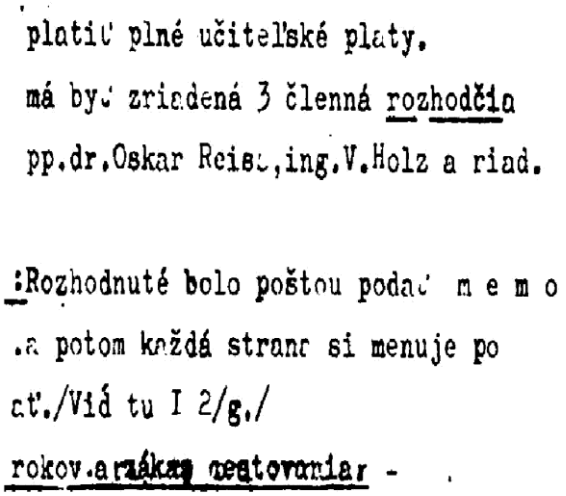

Fig 4: Output document image

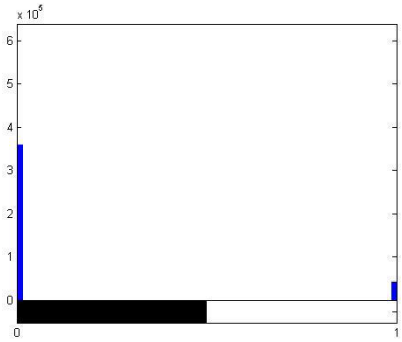

Fig 5: Histogram of Output document image

PSNR----

26.4251

RMSE--

12.1698

ENL--

0.0347

IV. CONCLUSION

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds 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. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, RMSE, ENL.

REFERENCES:

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 documentimage binarization contest (DIBCO 2009)," in *Proc. Int. Conf. DocumentAnal. Recognit.*, Jul. 2009, pp. 1375–1382.

[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 documentimage binarization contest (DIBCO 2011)," in *Proc. Int. Conf. DocumentAnal. Recognit.*, Sep. 2011, pp. 1506–1510.

[3] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwrittendocument image binarization competition," in *Proc. Int. Conf.Frontiers Handwrit. Recognit.*, Nov. 2010, pp. 727–732.

[4] S. Lu, B. Su, and C. L. Tan, "Document image binarization using backgroundestimation and stroke edges," *Int. J. Document Anal. Recognit.*,vol. 13, no. 4, pp. 303–314, Dec. 2010

[5] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwrittendocument images using local maximum and minimum filter," in *Proc.Int. Workshop Document Anal. Syst.*, Jun. 2010, pp. 159–166.

[6] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparisonof some thresholding algorithms for text/background segmentatioin difficult document images," in *Proc. Int. Conf. Document Anal.Recognit.*, vol. 13. 2003, pp. 859–864.

[7] M. Sezgin and B. Sankur, "Survey over image thresholding techniquesand quantitative performance evaluation," *J. Electron. Imag.*, vol. 13,no. 1, pp. 146–165, Jan. 2004.

[8] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarizationmethods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12,pp. 1191–1201, Dec. 1995.

[9] O. D. Trier and T. Taxt, "Evaluation of binarization methods fordocument images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17,no. 3, pp. 312–315, Mar. 1995.

[10] A. Brink, "Thresholding of digital images using two-dimensional entropies," *Pattern Recognit.*, vol. 25, no. 8, pp. 803–808, 1992.

[11] J. Kittler and J. Illingworth, "On threshold selection using clusteringcriteria," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 5,pp. 652–655, Sep.–Oct. 1985.

[12] N. Otsu, "A threshold selection method from gray level histogram," *IEEETrans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62–66, Jan. 1979.

BOOKS:

DIGITAL IMAGE PROCESSING BY C.GONZALEZ PEARSON EDUCATION.