BINARIZATION TECHNIQUE MAKES DEGRADED DOCUMENT IMAGE INTO ROBUST DOCUMENT IMAGE

F.Anto Mary Jenith[1] , Dr. S. Prasanna[2]

[1]*Research Scholar,School of Computing Sciences, VISTAS(Vels University),Pallavaram, Chennai,(India)*
[2]*Assoc. Prof, M.C.A. Dept, School of Computing Sciences, VISTAS(Vels University), Chennai,(India)*

ABSTRACT

Segmentation of text from badly degraded document images is a very demanding task due to the high inter/intra-variation between the text background and the foreground text of different text images. In this paper, we propose a novel document image binarization technique that addresses these issues by using adaptive image inconsistency. The adaptive image contrast is a combination of the local image disparity and the local image gradient that is tolerant to text and background variation caused by different types of document degradations. In the proposed system, an adaptive contrast map is rest constructed for an input degraded document image. The contrast map is then binaries and combined with Canny's edge map to classify the text stroke edge pixels. The document text is further segmented by a local entry that is estimated based on the intensities of detect text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum constraint tuning. It has been experienced on three public datasets that are used in the recent document image binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010 and achieves accuracies of 93.5%, 87.8%, and 92.03%, respectively that are significantly higher than or close to that of the best-performing methods reported in the three contests. Experiments on the Beckley diary dataset that consists of several challenging bad quality document images also show the greater performance of our proposed method, compared with other technique.

Keywords -Adaptive Image Contrast, Document Analysis, Document Image Processing, Degraded Document Image Binariza-Tion, Pixel Classification

I. INTRODUCTION

Document Image Binarization is performed in the preprocessing stage for document analysis and it aims to segment the foreground text from the document background. A quick and accurate document image binarization technique is important for the resulting document image processing tasks such as optical character recognition (OCR).Though document image binarization has been studied for many years, thresholding of degraded document images is still an unsolved problem due to the high inter/intra-variation between the textstroke and the document background across different document images. As illustrated in Fig. 1, 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-through as illustrated in Fig. 1(a) and (c) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in Fig. 1(e). 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 recent Document Image Binarization Contest (DIBCO) [1], [2] 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) [3] 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 performs the best among entries of algorithms submitted from international research groups. We also participated in the H-DIBCO 2010 and our local maximum-minimum method 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. The proposed method is simple, strong and capable of handling different types of degraded document images with minimum para-meter 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 mini-mum algorithm. At the same time, the parameters used in the algorithm can be adaptively estimated. The rest of this paper is organized as follows. Section IIfirst reviews the current state-of-the-art binarization tech-niques. Our proposed document binarization technique is described in Section III. Then experimental results are reported in Section IV to demonstrate the superior perfor-mance of our framework. Finally, conclusions are presented in Section V.

II. 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 non-stroke edges from the background of degraded document that often contains certain image variations due to the noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to com-pensate 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 . In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar of 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 normal-ization 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 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 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:

$$Ca(i, j) = \alpha C(i, j) + (1 - \alpha)(Imax(i, j) - Imin(i, j))$$

where $C(i, j)$ denotes the local contrast 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. deadly, 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 bi-narization 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 bi-narization technique relies more on image radiant and avoid the over normalization problem of our previous method. We model the **(10)** mapping from document image intensity variation to α by a power function as follows:

α= St d γ

128

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,∞], where the power function becomes a linear function when γ = 1. Therefore, the local image gradient will play the major role when γ is large and the local image contrast will play the major role when γ is small. The setting of parameter γ shows the contrast map of the sample document images in and that are created by using local image gradient, local image contrast and the sample document with a complex document background in the use of the local image contrast produces a better result as compared with the result by the local image gradient because the normalization factors helps to suppress the noise at the upper left area But for the sample document in 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 whereas the use of local image gradient is capable of preserving those light text strokes. As a comparison, the adaptive combination of the local image contrast and the local image gradient can produce proper contrast maps for document images with different types of degradation In particular, the local image contrast in 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 in

B. Text Stroke Edge Pixel Detection

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly.The constructed contrast image has a clear bi-modal pattern, where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background

III. LOCAL THRESHOLD ESTIMATION

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

IV. INDENTATIONS AND EQUATIONS

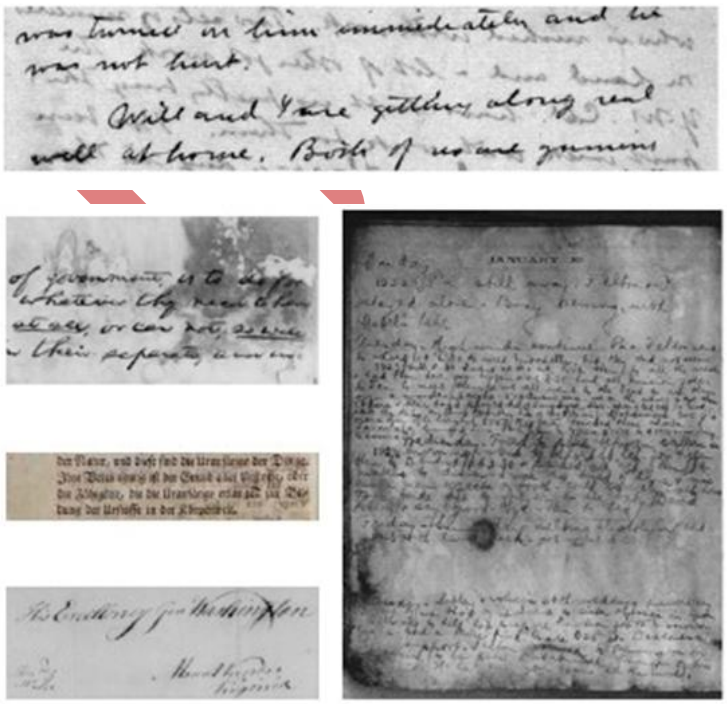
The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to the noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image

gradient needs to be normalized to com-pensate 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 . In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar of 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 normal-ization 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 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 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:

$$Ca(i, j) = \alpha C(i, j) + (1 - \alpha)(Imax(i, j) - Imin(i, j))$$

where $C(i, j)$ denotes the local contrast 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. deadly, 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 bi-narization 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 bi-narization technique relies more on image radiant and avoid the over normalization problem of our previous method.

V. FIGURES AND TABLES



VI. 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, NRM, MPM and DRD.

REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest(DIBCO 2009)," International Conference on Document Analysis and Recognition, pp. 1375–1382, July 2009.

[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," International Conference on Document Analysis and Recognition, September 2011.

[3]"H-DIBCO 2010 handwritten document image binarization competition," International Conference on Frontiers in Handwriting Recognition, pp. 727–732, November 2010.

[4] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," International Journal on Document Analysis and Recognition, vol. 13, pp. 303–314, December 2010.

[5] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter,"International Workshop on Document Analysis Systems, pp. 159–166, June 2010.

[6] G. Lethem, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images,"