IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834, p- ISSN: 2278-8735.
PP 58-64
www.iosrjournals.org

Segmentation of Text from Badly Degraded Document Image

Athira T N[1] , Priyanka Udayabhanu[2]
[1](ECE Dept , SNGCE , India)
[2](ECE Dept , SNGCE , India)

Abstract: *This paper proposes a method to improve the contrast of text in a badly degraded document. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.*
Keywords: *Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.*

I. Introduction

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/intravariation between the text stroke 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.

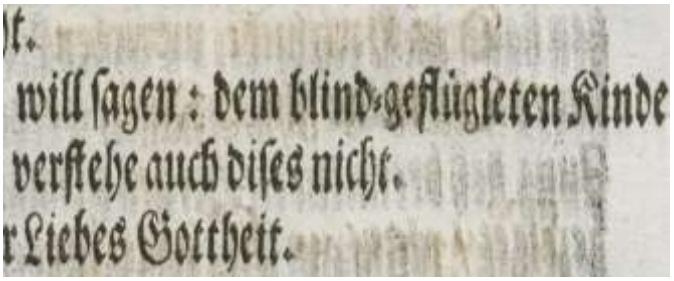

Fig 1.degraded document image example

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 [5]. At the same time, the parameters used in the algorithm can be adaptively estimated. document Image Binarization is performed in the preprocessing stage for document analysis The rest of this paper is organized as follows. Section 2 first reviews the current state-of-the-art binarization techniques. Our proposed document binarization technique is described in Section 3.

II. Related Method

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 [10]–[13] is usually not a suitable approach for the degraded document binarization. Adaptive thresholding [14]–[20], 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 [18], [19] 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 [4], [21], texture analysis [22], recursive method [23], [24], decomposition method [25], contour completion [26]–[28], Markov Random
Field [29]–[32], matched wavelet [33], cross section sequence

graph analysis [34], self-learning [35], Laplacian energy [36]
user assistance [37], [38] and combination of binarization techniques [39], [40]. 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 [5], [14], [18], [19]. In Bernsen's paper [14], the local contrast is defined as follows:

$$C(i, j) = I_{max}(i, j) - I_{min}(i, j) \quad (1)$$

where $C(i, j)$ denotes the contrast of an image pixel (i, j), $I_{max}(i, j)$ and $I_{min}(i, j)$ 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 $I_{max}(i, j)$ and $I_{min}(i, j)$. 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]:

$$C(i, j) = I_{max}(i, j) - I_{min}(i, j)[(I_{max}(i, j) + I_{min}(i, j) + \sum] \quad (2)$$

where $\sum$ 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 2.

III. Proposed Method

This section describes the proposed document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The
text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document
binarization quality.

3.1. Contrast Image Construction

The image gradient has been widely used for edge detection [42] 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 [5], 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 [42]. 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:

$$C_a(i, j) = \alpha C(i, j) + (1 - \alpha)(I_{max}(i, j) - I_{min}(i, j)) \quad (3)$$

where $C(i, j)$ denotes the local contrast in Equation 2 and $(I_{max}(i, j) - I_{min}(i, j))$ 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 [5]. We model the mapping from document image intensity variation to α by a power function as follows:

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

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. 4.
Fig. 2 shows the contrast map of the sample document images in Fig.1 that are created by using local image gradient [43], local image contrast [5] and our proposed method in Equation 3, respectively

Fig. 2. Adaptive contrast map of sample document image

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. 2. In particular, the local image contrast in Equation 3 gets a high weight for the document image in Fig. 1 with high intensity variation within the document background

3.2. 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 [5], where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background.We therefore detect the text stroke edge pixel candidate by using Otsu's global thresholding method. For the contrast images in Fig. 2, Fig. 3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly. As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels.

The binary map can be further improved through the combination with the edges by Canny's edge detector [43], because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading [44]. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3(b) without tuning the parameter

manually. In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Fig. 3(c).

(a)

(b)

(C)

Fig 3. (a) (a) Binary contrast maps, (b) canny edge maps, and their (c) combined edge maps of the sample document images in Fig. 1

3.3. 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 [5]: 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:

$$R(x, y) = 1 \quad I(x, y) \leq E\text{mean} + E\text{std}/2$$

$$= 0 \quad \text{otherwise} \qquad (5)$$

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) and the pixels next to them are labeled to 1 (edge) in the edge map (*Edg*), are correctly detected, they should have higher intensities than the following few pixels (which should be the text stroke pixels). So those improperly detected edge pixels are removed in step 4. In the remaining edge pixels in the same row, the two adjacent edge pixels are likely the two sides of a stroke, so these two adjacent edge pixels are matched to pairs and the distance between them are calculated in step 5. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width *EW* can then be approximately estimated by using the most frequently occurring distances of the adjacent edge pixels as illustrated in Fig. 4.

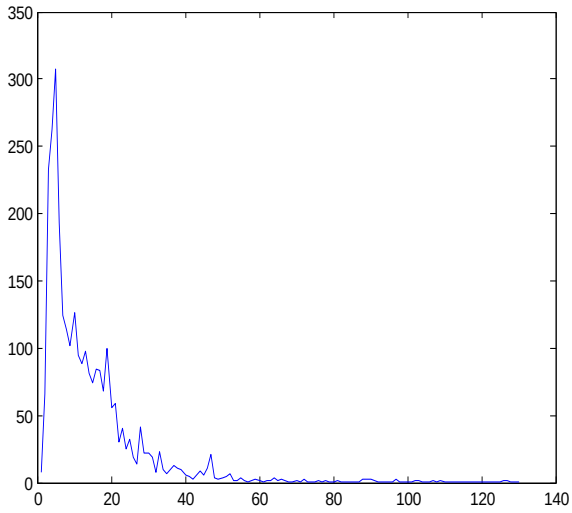

Fig. 4. Histogram of the distance between adjacent edge pixels.

Algorithm 1 Edge Width Estimation
Require: The Input Document Image *I* and Corresponding Binary Text Stroke Edge Image *Edg*
Ensure: The Estimated Text Stroke Edge Width *EW*
1: Get the *width* and *height* of *I*
2: **for** Each Row *i* = 1 to *height* in *Edg* **do**
3: Scan from left to right to find edge pixels that meet the following criteria:
a) its label is 0 (background);
b) the next pixel is labeled as 1(edge).
4: Examine the intensities in *I* of those pixels selected
in Step 3, and remove those pixels that have a lower
intensity than the following pixel next to it in the same
row of *I* .
5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
6: **end for**
7: Construct a histogram of those calculated distances.
8: Use the most frequently occurring distance as the estimated
stroke edge width *EW*.

3.4. Post-Processing

Once the initial binarization result is derived from Equation 5 as described in previous subsections, the binarization result can be further improved by incorporating certain domain knowledge as described in

Algorithm 2. First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class

Algorithm 2 Post-Processing Procedure
Require: The Input Document Image I , Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edg
Ensure: The Final Binary Result B_f
1: Find out all the connect components of the stroke edge pixels in Edg.
2: Remove those pixels that do not connect with other pixels.
3: **for** Each remaining edge pixels (i, j): **do**
4: Get its neighborhood pairs: $(i-1, j)$ and $(i+1, j)$; $(i, j-1)$ and $(i, j+1)$
5: **if** The pixels in the same pairs belong to the same class (both text or background) **then**
6: Assign the pixel with lower intensity to foreground class (text), and the other to background class.
7: **end if**
8: **end for**
9: Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.
10: Store the new binary result to B_f

.

Fig 5 final contrast image

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.

Acknowledgement

I would like to thank Asst.Prof .Priyanka udayabhanu for her valuable guidance .I would also like to thank entire ec department for their cooperation.

References

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 1375–1382.
[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1506–1510.

[3] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document 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 background estimation 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 handwritten document 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, "Comparison of some thresholding algorithms for text/background segmentation in 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 techniques and 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 binarization methods," *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 for document 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, 199[11] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," *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," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62–66, Jan. 1979.
[13] N. Papamarkos and B. Gatos, "A new approach for multithreshold selection," *Comput. Vis. Graph. Image Process.*, vol. 56, no. 5,pp. 357–370, 1994.
[14] J. Bernsen, "Dynamic thresholding of gray-level images," in *Proc. Int. Conf. Pattern Recognit.*, Oct. 1986, pp. 1251–1255.
[15] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 1991, pp. 435–443.
[16] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognit.*, vol. 35, no. 1, pp. 265–277, 2002.
[17] J. Parker, C. Jennings, and A. Salkauskas, "Thresholding using an illumination model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Oct. 1993, pp. 270–273.
[18] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.