International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Segmentation of Text from Degraded Document Images by Local Threshold Method

D. Dharani[1], D. Saraswathi[2]

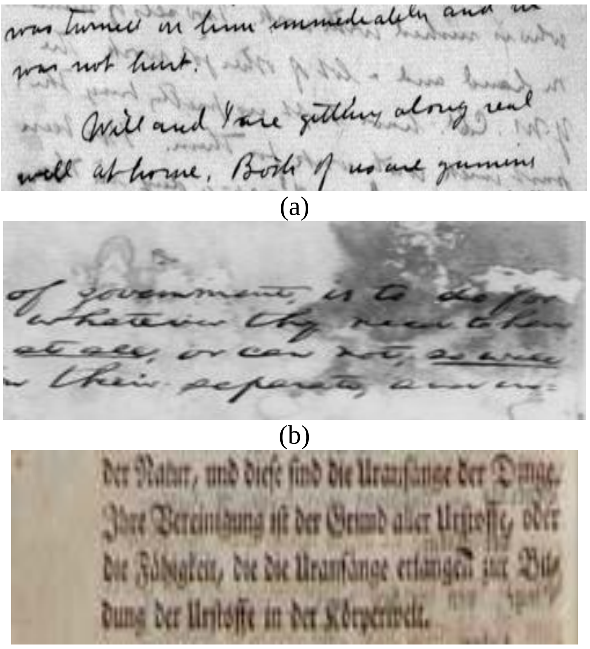
[1,2]Department of ECE, Manakula Vinayagar Institute of Technology, Puducherry, India

Abstract: ***Text segmentation from badly degraded document images is a very interesting task due to the high inter/intravariation of different document images. In this paper, a document image binarization technique that reports these problems by adaptive image contrast. An adaptive contrast map is first constructed for an input degraded document image by combination of the local image contrast and local image gradient. The contrast map is then binarized and combined with Canny's edge map to find the text str oke edge pixels. The text is further segmented by a local threshold method. Some post-processing is further applied to increase the document binarization quality. The proposed method is simple and involves minimum parameter tuning. To improve the quality of the text in the degraded document image using two thresholding techniques. One is OTSU with several edge detection (i.e. canny, sobel, and total variation) techniques applied to the degraded document image. Another is Adaptive threshold with several edge detection (i.e. canny, sobel, and total variation) techniques applied to the degraded document image. The qualities of these output images evaluated by PSNR and MSE. The best combination of threshold and edge detection techniques is selected by testing several degraded documents.***

Keywords: binarization, thresholding, pixel classification, adaptive image contrast, document analysis, degraded document image.

1. Introduction

Document Image Binarization is made in the pre-processing stage for document analysis and its use to segment the foreground text from the background of the document. An accurate and fast document image binarization technique is important for the succeeding document image processing. Though document image binarization has been studied for various years, the thresholding of degraded document is still an unresolved problem due to the high intra/intervariation between the background and text of the document across different document images. As shown in Fig. 1, the handwritten text often shows a definite amount of variation interms of the stroke brightness, width, connection and background of the document.


(a)

(b)

(c)

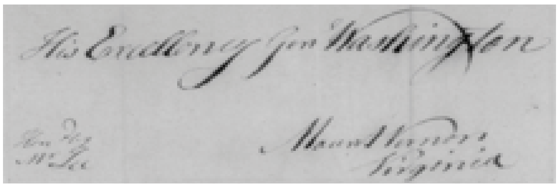

(d)

Figure 1: Degraded document image examples (a)–(d) are taken from DIBCO series datasets

In addition, historical documents are frequently degraded by the bleed through as shown in Fig. 1(a) and (c) where the ink of the other side leaks over to the front. These different forms of document degradations tend to induce the document thresholding mistake and create degraded document image binarization a big challenge to most state-of-the-art techniques.

2. Related Work

Several thresholding techniques [6]–[9] have been stated for document image binarization. As many degraded documents do not have a perfect bimodal pattern, global thresholding [10]–[13] is regularly not an appropriate approach for the degraded document binarization. Estimate a local threshold for each document image pixel by adaptive thresholding and it is better approach to deal with different variations within degraded document images [14]-[20].

The main drawback of window-based thresholding techniques is that the performance depends heavily on the window size and the width of the stroke. The local image gradient and contrast are very suitable features for segmenting the text from the document background [5], [18], [19]. The local contrast is defined in Bernsen's paper as follows [14]:

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

where $C(i,j)$ denotes the contrast of an image pixel $C(i,j)$, $I_{\max}(i,j)$ and $I_{\min}(i,j)$ denote the maximum and minimum intensities $C(i, j)$ within a narrow neighborhood windows respectively. If the local contrast is smaller than a threshold, the pixel is set as background openly. Otherwise it will be classified into passage or surroundings by comparing with the mean of $I_{\max}(i, j)$ and $I_{\min}(i, j)$.

Bernsen's method is simple and cannot work accurately on degraded document images that have a complex document background. This method have prior planned a novel image binarization method [5] by local image contrast that is calculated as follows [21]:

$$C(i, j) = \frac{I_{\max}(i, j) - I_{\min}(i, j)}{I_{\max}(i, j) + I(i, j) + e} \quad (2)$$

where $C(i, j)$ is a small and positive number that is added if the local maximum value is 0. The local image contrast in Equation 2 introduce a normalization factor (as the denominator) to compensate the image distinction compared with Bernsen's Equation 1. 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.

3. Proposed Method

This section defines the proposed document image binarization techniques. An adaptive contrast map is first built and the text stroke edges are then detected through the mixture of the adaptive contrast map and the canny edge map for a given degraded document. Then the text is segmented based on the local threshold value that is predictable from the detected text stroke edge pixels. Finally, post-processing is further applied to increase the document binarization quality.

3.1 Contrast Image Construction

The image gradient has been commonly used for edge detection [22] and it can be used to detect the text stroke edges of the document images efficiently. On the other hand, it often detects many nonstroke edges from the background that frequently contains some image variations due to noise, uneven lighting, bleed-through, etc. To properly extract the stroke edges, gradient of the image needs to be normalized.

In the earlier method [5], the local contrast evaluated by maximum and minimum value of the local image is used to suppress the background variation as defined in Equation 2. In particular, the numerator captures the difference of local image that is similar to the old image gradient [22]. The denominator is a normalization factor. Though, the image contrast in Equation 2 has one limitation that it may not handle the document images with the bright text properly. Combine the local image contrast with the local image gradient and develop an adaptive local image contrast to overcome this problem as follows:

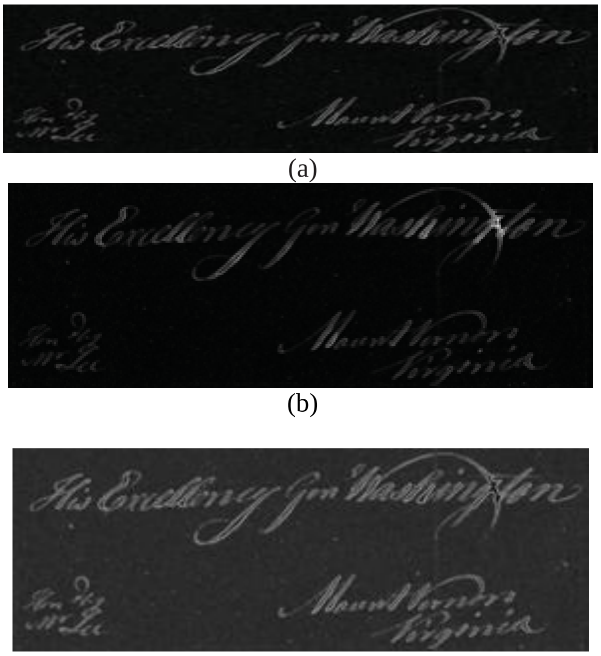
$$C_a(i, j) = \alpha C(i, j) + (1 - \alpha)(I_{\max x}(i, j) - I_{\min}(i, j)) \quad (3)$$

where $C(i, j)$ denotes the local contrast in Equation 2 and $(I_{\max x}(i, j) - I_{\min}(i, j))$ refers to the local image gradient that is normalized to [0, 1]. The local windows size is fixed to 3. The proposed method depends on image gradient and avoid the over normalization problem of the previous method [5]. In this model, the mapping from intensity variation of the document to α using a power function as follows:

$$\alpha = \left(\frac{Std}{128}\right)^{\gamma} \quad (4)$$

where Std indicates the 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

Fig. 2 shows the contrast map of the sample document images in Fig. 1(d). For the sample document with a complex document background in Fig. 1(d), the local image contrast yields a superior result as shown in Fig. 2(b) compared with the result by the local image gradient as shown in Fig. 2(a). The adaptive combination of the local image contrast and the local image gradient in Equation 3 can produce proper contrast maps.



(a)

(b)

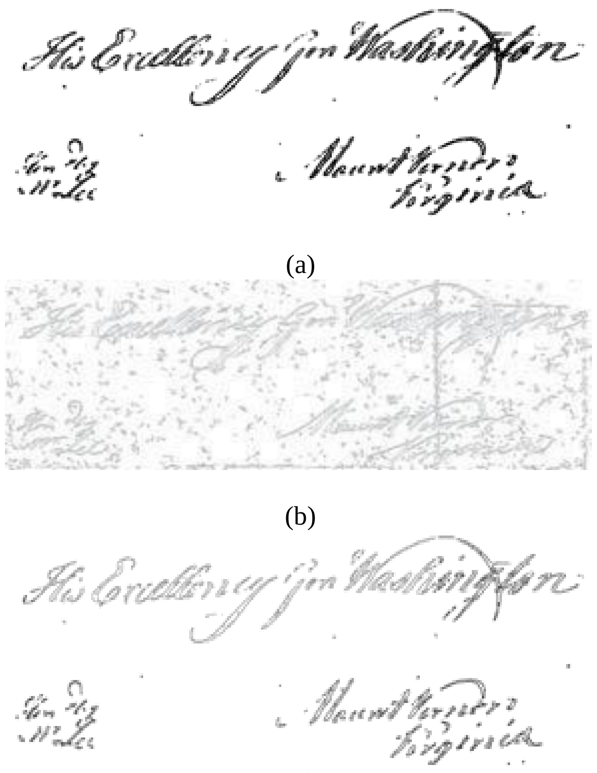
(c)

Figure 2: Contrast Images made using (a) local image gradient [22], (b) local image contrast [5], and (c) proposed method of the sample document image in Fig.1 (d), respectively.

The adaptive combination in Equation 3 can produce accurate contrast maps for document images with different kinds of degradation as shown in Fig. 2(c) for comparison. In specific, the local image contrast in Equation 3 gets a high weight for the document image in Fig. 1(a) with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image in Fig. 1(b).

3.2 Text Stroke Edge Pixel Detection

The use of the contrast image construction is to detect the stroke edge pixels of the document text properly. The assembled contrast image has a clear bi-modal pattern [5], where the adaptive image contrast calculated at edge stroke text is clearly larger than that calculated within the document background. Therefore detect the text stroke edge pixel candidate by Otsu's global thresholding method. For the contrast images in Fig. 2(c), Fig. 3(a) displays a binary map by Otsu's algorithm. The binary map can be further improved over the mixture with the edges by Canny's edge detector [23], because Canny's edge detector has a good localization property that it can spot the edges close to real edge locations in the detecting image.



(a)

(b)

(c)

Figure 3: (a) Binary contrast map, (b) canny edge map, and their (c) joint edge maps of the sample document image in Fig. 1 (d) respectively.

It should be noted that Canny's edge detector by itself repeatedly extracts a great amount of non-stroke edges as shown in Fig. 3(b). In the joined map, keep only pixels that look within both the canny edge map and high contrast image pixel map. The mixture helps to extract the text stroke edge pixels perfectly as shown in Fig. 3(c).

3.3 Local Threshold Estimation

In local threshold estimation, once the high contrast stroke edge pixels are detected properly, the text can be extracted from the document background pixels. Two features can be detected from different kinds of document images [5]: First, the pixels of the text are close to the detected text stroke edge pixels. Second, there is a different intensity difference between the high contrast stroke edge pixels and the neighboring background pixels. The text image document can be extracted based on the identified text stroke edge pixels as follows:

$$R(x,y)=\begin{cases}1 & I(x,y)\le E_{mean}+\frac{E_{Std}}{2}\\ 0 & otherwise\end{cases} \qquad (5)$$

where E_{mean} and E_{std} are the mean and standard deviation of the identified text stroke edge pixels within a neighborhood window W, respectively. The neighborhood window should be greater than the stroke width in order to contain stroke edge pixels. So the neighborhood window size W can be fixed based on the stroke width of the document image as stated in Algorithm 1.

First the edge image is scanned horizontally row by row and the edge pixel applicants are nominated as defined in step 3. If the edge pixels, which are labeled 0 (background) and the pixels next to them are labeled to 1 (edge) are correctly identified, and must have higher intensities than the next few pixels (which should be the text stroke pixels). So those inappropriately detected edge pixels are removed in step 4. In the left over edge pixels in the same row, the two adjacent edge pixels are expected the two sides of a stroke, so these two nearby edge pixels are matched to pairs and the distance between them are calculated in step 5. Next, a histogram is built that registers the frequency of the distance between two neighboring applicant pixels. The edge width *EW* can be approximately valued by the most repeatedly occurring distances of the adjacent edge pixels.

Algorithm 1 Edge Width Estimation
Require: The Input Document Image I and Equivalent Binary Text Stroke Edge Image Edg
Ensure: The Estimated Text Stroke Edge Width EW

```
1: Get the height and width of I
2: for Each Row i = 1 to height in Edg do
3: Scan from left to right to find edge pixels that see the following conditions:
   a) its label is 0 (background);
   b) the next pixel is labeled as 1(edge).
4: Examine the intensities in I of those pixels nominated in Step 3, and remove those pixels that have a lower intensity than the succeeding pixel next to it in the same row of I.
5: Match the remaining adjacent pixels in the same row into pairs, and estimate the distance between the two pixels in pair.
6: end for
7: Construct a histogram of those calculated distances.
8: Use the most repeatedly occurring distance as the estimated stroke edge width EW.
```

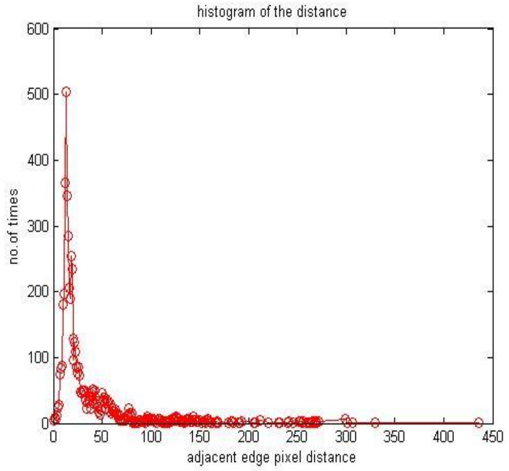


Figure 4: Histogram of the distance between adjacent edge pixels.

3.4 Post Processing

In post processing, the binarization result derived from Equation 5 as described in previous subsections, can be further improved by including certain domain knowledge as described in Algorithm 2.

Algorithm 2 Post-Processing procedure
Require: Image Document Input I, Initial Binary Result B and Corresponding Binary Text Stroke Edge Image
Ensure: The Final Binary Result B f

```
1: Find out all the connect components of the edge stroke pixels in Edge.
2: Eliminate those pixels that do not connect with other pixels.
3: for each left over edge pixels (i, j): do
4: Get its adjacent pairs: (i − 1, j ) and (i + 1, j ); (i, j − 1) and (i, j + 1)
5: if the pixels in the identical sets belong to the same class then
6: Assign the pixel with lower intensity to forefront class and the other to background class.
7: end if
8: end for
9: Remove artifacts beside the boundaries of text stroke after the document thresholding.
10: Store the new binary result to Bf.
```

First, the foreground pixels that do not attach with other foreground pixels are filtered out to make the edge pixel set specifically. Second, the adjacent pixel pair that lies on symmetric sides of a text stroke edge pixel must belong to different classes. One pixel of the pixel pair is marked to the other class if both of the two pixels belong to the same class. Lastly, some artifacts along the text stroke boundaries are filtered out by some logical operators.

4. Experimental Results

Table I shows that performance measures of various threshold and edge detection techniques. The result shows that proposed method performed best document binarization methods in term of PSNR and MSE. The best combination of threshold and edge detection techniques is selected by testing several degraded documents.

Table 1: Various performance measures using different threshold and edge detection techniques

METHOD	*PSNR*	*MSE*
OTSU with canny	49.789	0.123
OTSU with sobel	48.456	0.567
OTSU with TV	49.678	0.256
Adaptive with canny	50.567	0.768
Adaptive with sobel	50.234	0.567
Adaptive with TV	50.456	0.654

5. Conclusion

In this paper, an adaptive image contrast based document image binarization technique that is tolerant to different kinds of document degradation such as document smear and uneven illumination. The proposed technique is simple and only limited factors are involved. Additionally, it works for different kinds of degraded document images. The proposed method makes usage of the local image contrast that is estimated based on the local maximum and minimum. The proposed method has been verified on the several datasets. The results show that the proposed method outperforms most reported document binarization methods interms of PSNR and MSE. The best combination of threshold and edge detection techniques is selected by testing several degraded documents.

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, 1992.

[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, ―Anew 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, ―Afast 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.

[19] W. Niblack, ―An Introduction to Digital Image Processing", Englewood Cliffs, NJ: Prentice-Hall, 1986.

[20] J.-D. Yang, Y.-S. Chen, and W.-H. Hsu, ―Adaptive thresholding algorithm and its hardware implementation," Pattern Recognit. Lett, vol. 15, no. 2, pp. 141–150, 1994.

[21] M. van Herk, ―A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels," Pattern Recognit. Lett, vol. 13, no. 7, pp. 517–521, Jul. 1992.

[22] D. Ziou and S. Tabbone, ―Edge detection techniques—an overview," Int. J. Pattern Recognit. Image Anal., vol. 8, no. 4, pp. 537–559, 1998.

[23] J. Canny, ―A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 6, pp. 679–698, Jan. 1986.

[24] T. Lindeberg, ―Edge detection and ridge detection with automatic scale selection," Int. J. Comput. Vis., vol. 30, no. 2, pp. 117–156, 1998.