International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March-April 2016 **ISSN 2278-6856**

An Improved Binarization Technique for Degraded Document Images Using Adaptive Image Contrast

Prof.Sushilkumar N Holambe r[1], Dr.Ulhas B Shinde.[2] Bhagyashree S Choudhari [3]

[1] PG Coordinator ME(CSE) , College of Engineering, Osmanabad

[2].Principal ,Chh.Shahu College of Engineering,Aurangabad

[3]. Persuing ME(CSE), College of Engineering, Osmanabad

Abstract

This paper presents a new adaptive approach to deal with degraded documents by using binarization method. The proposed method is mostly used due to its capability to work without any requirement of parameter tuning by the user. And this method can deal with documents degradations which normally occur due to non-uniform illumination, shadows, large signal-dependent noise, low contrast, strain and smear. This improved document image binarization technique focus on improving quality of text from poorly degraded document images. And this can be achieved with good efficiency and accuracy. As today's world is moving towards digitization, binarization method comes in a demand, to achieve the preserving target of historical documents. to deal with the degraded documents this proposed method uses the adaptive contrast map as the combination of local image contrast and the local image gradient.

Keywords: adaptive contrast map, binarization, contrasts, degraded document.

1. INTRODUCTION

The Document image binarization technique focuses on to the image segmentation. And it's aim is achieved by conversion of a grayscale image into a binary image. Document image binarization plays key role in document image processing. This is mainly used in preprocessing stages of document image processing related applications such as optical character recognition (OCR) and document image retrieval [1], [2].preservation of logical and semantic content in document image is required for thresholding. Though document image binarization has been studied for many years, the thresholding of images is still a challenging task due to the high variation between the text stroke and the document background. For the inputted images, some preprocessing stages need to be performed before the text extraction. Binarization is one of them. As the processing result of binary image gives better result than the grayscale image as shown in fig.1, we convert grey-scale image into a binary image in this stage.

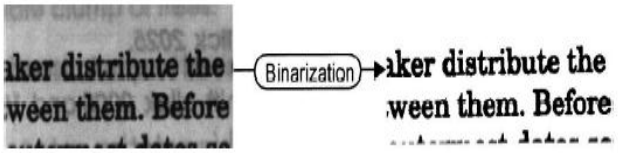
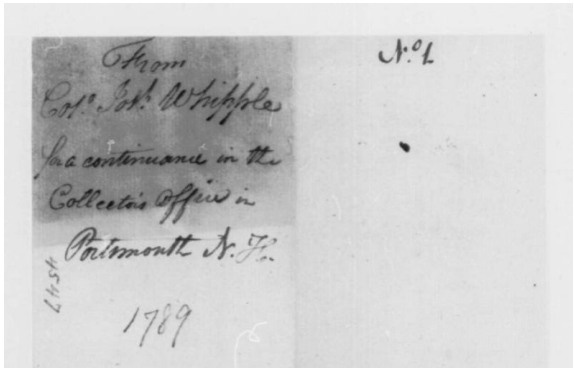
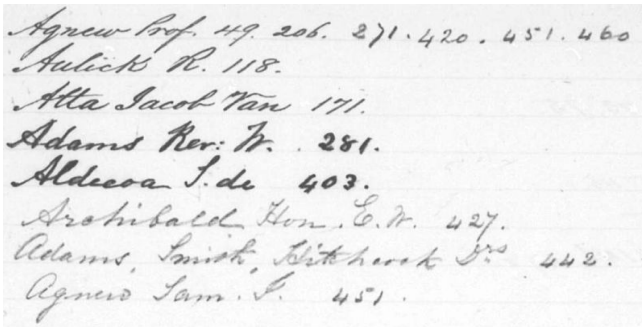


Fig. 1 Binarization example

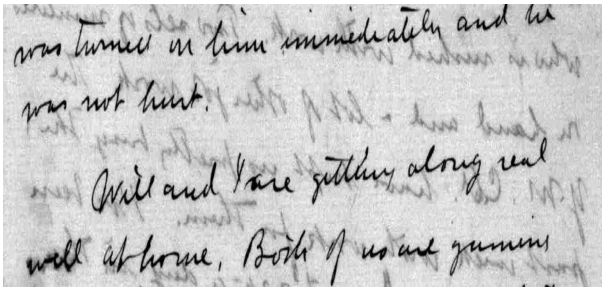
The thresholding techniques of degraded document images is a big issue due to document degradation such as uneven illumination, change in image contrast, old age, smear and bleeding-through that exist in many DIBCO document images as illustrated in Fig. 2.



(a)



(b)



(c)

Fig.2 (a), (b), (c) Degraded document Images from DIBCO dataset.

To deal with grayscale documents the binarization techniques can be categories in two groups as

- Global binarization technique
- Local binarization technique.

Considering the single threshold value for the whole document image is done in global binarization method. Separation of each and every pixel to the foreground or background is totally depend on its gray value. Global binarization methods works good for the typical scanned documents. but to work with the applications such as camera-captured documents or scanned book pages where the illumination over the document is nonuniform and applications where image intensities can change significantly within a document such as historical documents global binarization method fails to give accurate result because ,global binarization methods tend to produce marginal noise along the page borders[3]. Local binarization [4], [5], [6] method was introduced to overcome on these problems of global thresholding method.local thresholding method consider information from the local neighborhood of the pixel, and for each and every pixel it computes threshold individually.in this paper we proposes a novel locally adaptive thresholding technique to maintain the originality of document by binarizing and improving the quality of degraded document images.

2. PROPOSED SYSTEM

This section describes the proposed document image binarization techniques. For a given a degraded document image, construction of an adaptive contrast map is done first afterwards by combining binarized adaptive contrast map and the canny edge map, the text stroke edges are detected. Threshold is further estimated from the detected text stroke edge pixels,and this threshold is used to segment the text.to improve final document binarization quality some post processing is applied.

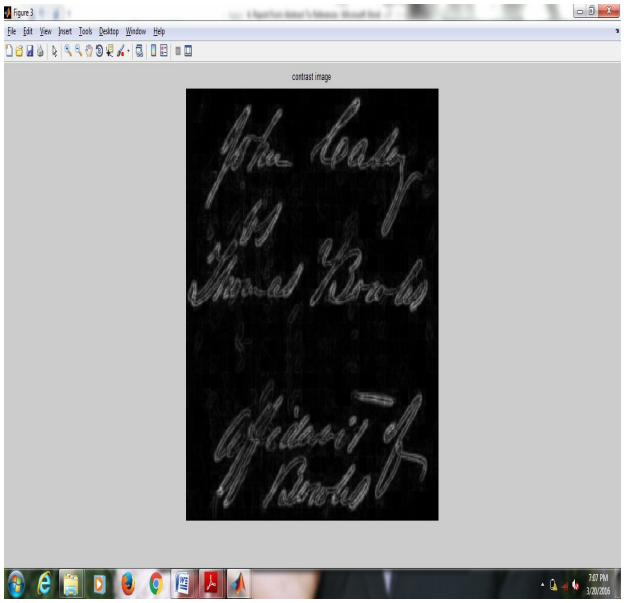
2.1. Contrast Image Construction.
2.2. Text stroke edge pixel detection.
2.3 .Local Threshold Estimation.
2.4. Post Processing.

2.1. Contrast Image Construction

The main purpose of image gradient is to detect edges and it can also be used to detect the text stroke edges as well[7]. Degraded document images with variations in the background caused due to uneven lighting, bleed-through , noise, etc. can also be used as to detect their non-stroke edges also. the image gradient needs to be normalized to compensate the image variation within the document background, To extract only the stroke edges properly.They are used in many document image binarization techniques [8] [9] and are very effective. When document image has noticeable intensity variations, need to work with the image contrast with high weigh (i.e. Large α To overcome this over-normalization problem [10], we combine the local image contrast with the localimage gradient and derive an adaptive local image contrast as follows:

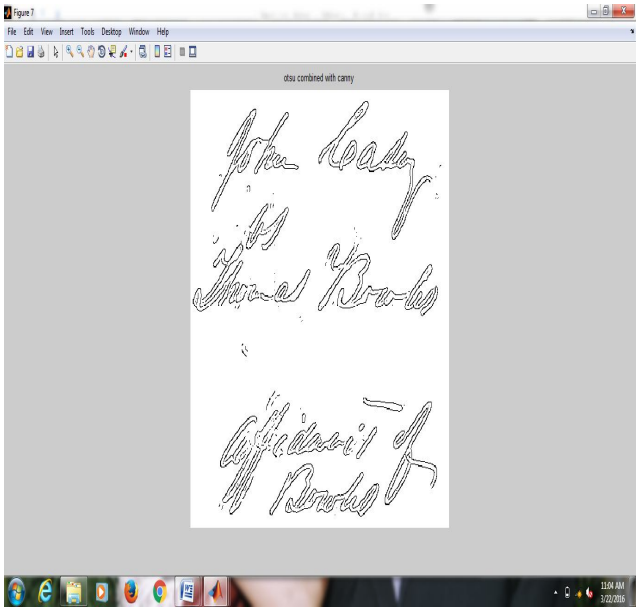
$$C\alpha(i,j) = \alpha C(i,j) + (1-\alpha)(I\max(i,j) - I\min(i,j)) \quad \ldots (1)$$

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]



2.2 Text stroke edge pixel detection

To detect the text stroke edges of pixel candidate properly we often use Otsu's global thresholding method. The binary map can be further improved through the combination with the edges by Canny's edge detector [11], 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 [12]. Many times Canny's edge detector without tuning the parameter manually, so often extracts a large amount of non-stroke edges we in combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map.



2.3 Local Threshold Estimation.

Two characteristics can be observed from different kinds of document images [13]: First, the text pixels are as close as 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:

$$R(x, y) = \begin{cases} 1 & I(x, y) \leq Emean + \frac{Estd}{2} \\ 0 & otherwise \end{cases} \quad (2)$$

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 larger enough to contain stroke edge pixels. Based on the stroke width of the document, neighborhood window size need to be set and the corresponding estimation can be done as stated in algorithm 1.

Algorithm 1: To estimate Edge Width *EW*

1: calculate the width and *height* of Input Document Image *I*

2: **for** Each Row *i* = 1 to *height* in Binary Text Stroke Edge Image *Edg* **do**

3: from left to right scan the edge pixel those who satisfies the criteria as:

a) Edge pixel with label as 0 (background);

b) The next pixel with label as 1(edge).

4: Examine the intensities 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: remaining adjacent pixels need to be paired that are in the same row, and calculate the distance between the two pixels in pair.

6: **end for**

7: calculated distances are used to construct the histogram.

8: The most frequently occurring distances is nothing but the estimated stroke edge width *EW*.

2.4 Post Processing

The binarized document image can be further improving the result. This can be explained in algoritham.2

Algorithm 2 : Post-Processing Procedure (Final Binary Result B)

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: In an edge look for all the connect components of the stroke edge pixels

2: eliminate the pixels that are not connect with other pixels.

3: **for** Each remaining edge pixels *(i, j)*: **do**

4: Get its neighborhood pixel pairs: *(i* − 1, *j)* and *(i* + 1, *j)*; *(i, j* − 1*)* and *(i, j* + 1*)*

5: **if** The same paired pixels belongs to the same class (both text or background) **then**

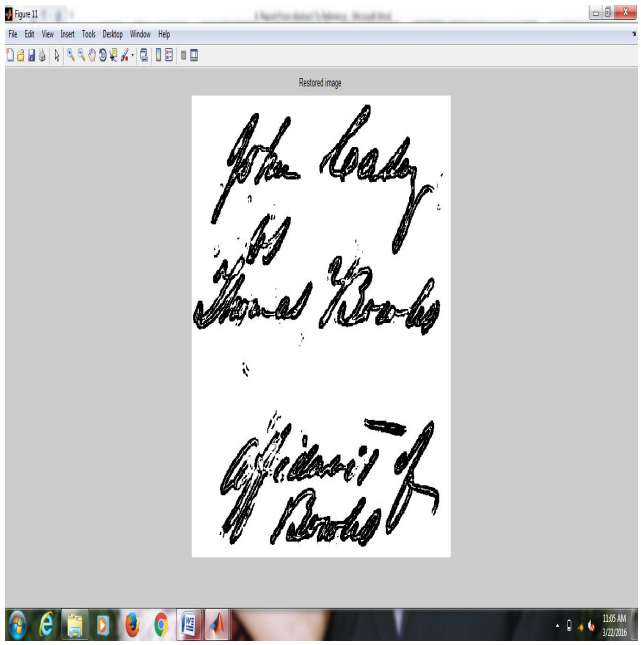
6: pixels with lower intensity are assigned to foreground class (text), and the other to background class.

7: **end if**

8: **end for**

9: Remove single-pixel artifacts [14] along the text stroke boundaries after the document thresholding.

10: Store the new binary result to *B f.*



3.CONCLUSION

This paper gives a new approach to deal with degraded document image known as adaptive image contrast based document image binarization technique. This technique is tolerant to documents degradations which normally occur due to non-uniform illumination, shadows, large signal-dependent noise, low contrast, strain and smear. With the involvement of few parameters only, this technique is easy and robust. the local image contrast used in proposed technique is evaluated based on the local maximum and minimum. Testing of the same can be done on various datasets.

References

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images" IEEE TRANS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.

[2] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[3] Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Page frame detection for marginal noise removal from scanned documents," in 15th Scandinavian Conference on Image Analysis,pp. 651–660, (Aalborg, Denmark), June 2007.

[4] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition 33(2), pp. 225–236, 2000.

[5] J. Bernsen, "Dynamic thresholding of gray level images," in Proc. Intl. Conf. on Pattern Recognition, pp. 1251–1255, 1986.

[6] W. Niblack, "An Introduction to Image Processing", Prentice-Hall, Englewood Cliffs, NJ,1986.

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

[8] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit, Jul. 2009, p.1375–1382.

[9] I. Pratikakis, B. Gaos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in Proc. Int. Conf. Document Anal. Recognit, Sep. 2011, pp.1506–1510

[10] Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, IEEE, "Robust Document Image binarization Technique for Degraded Document Images", IEEE Transactions on Image Processing, Vol. 22, No. 4, April 2013

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

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

[13] 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.

[14] 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.