Research Paper

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

Simple and Efficient Document Image Binarization Technique For Degraded Document Images

Engineering

KEYWORDS : Document Image Processing, Document Image binarization, Degraded document image, Adaptive Image Contrast

Manju Joseph — PG Scholar, Department of Computer Science and Engineering, Vedavyasa Institute of Technology, University of Calicut, Kerala, India

Jijina K.P — Assistant Professor, Department of Computer Science and Engineering, Vedavyasa Institute of Technology, University of Calicut, Kerala, India

ABSTRACT

Document Image binarization converts an acquired gray-scale document image to binary format, the objective of binarization is to automatically choose a threshold that separates the foreground and background information. Document image binarization is a process that is usually carried out in the pre-processing stage of document image processing. Primary aim of this document image binarization is to extract the foreground text from the document background. In the case of degraded document images this text extraction or segmentation is a difficult task. In this paper, we propose a simple and efficient document image binarization technique it makes use of the adaptive image contrast and some of the noise reduction methods. In the proposed technique, first input degraded document image is normalized to improve the quality of output binarized document image. Second, an adaptive image contrast map is constructed for the normalized image .Third, adaptive image contrast map is binarized and combined with Canny's edge map to identify the text stroke edge pixels. Then the document text is segmented by a local threshold that is estimated based on the intensities of the detected text stroke edge pixels. Finally, the output document image is filtered to reduce noise. The proposed method requires only minimum number of parameters. This method shows superior performance over various datasets interms of various performance measures.

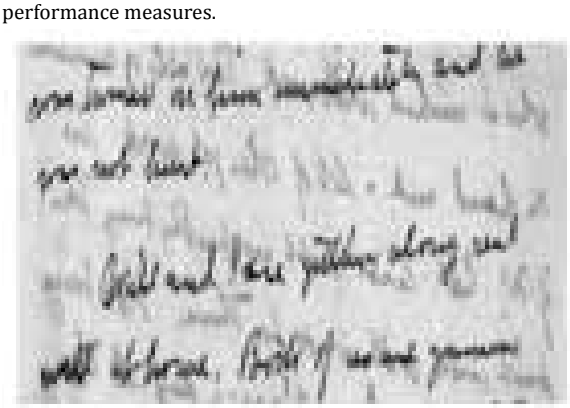
I. INTRODUCTION

In recent years, the field of document image processing has increasingly widespread applicability and powerful growth.

Document Image Binarization is usually performed in the preprocessing stage of Document image processing. Frequently, binarization is used as a pre-processor before Optical Character Recognition (OCR). Image binarization converts an image of up to 256 gray levels to a black and white image Document Image Binarization converts a gray-scale document image into binary document image. The main of this document image binarization technique is to segment or extract foreground text from the document background. In the case of degraded document images this foreground text extraction is a challenging task due to variations in the document image properties. By degradations we mean every sort of less-than ideal properties of a real document image, example coarsening of document image, ink or toner drop-outs, smear, thinning and thickening, geometric deformations etc. Handwritten text within document images also shows some variations in stroke width, stoke connection etc. In addition historical document images are often degraded by bleed-trough.

Many document image binarization techniques [1]-[10] have been reported for binarization of degraded document images .In the year 1998, a recursive thresholding thresholding technique for image segmentation [1] has been proposed .This approach is only applicable to gray-scale images specifically for real-life bank checks. Performance analysis indicates that this method is more efficient to segment darkest object in a given image. An Iterative multimodel subimage binarization technique [2] has been proposed for handwritten document images in the year 2004.This approach can be used for different type of handwritten document images where we do not do not have prior knowledge about noisiness of document image .In the year 2005, an image binarization technique [3] has been proposed for degraded historical document images .This approach is mainly based on a decompose algorithm. Main drawback of this approach is that the algorithm does not works well on document images with big pattern or pictures. In order to give best results on heavily degraded document images a document image binarization technique using Markov field model [4] has been proposed .This method is more effective to detect text than other local thresholding methods. An improved document image binarization technique [5] has been proposed in the year 2008 .This method is mainly based on the combination of different document image binarization technique and efficient edge information about gray scale images. A document image binarization using background estimation and stroke edges [6] has been proposed in the year 2010. The proposed document thresholding method still has several limitations. One of the drawback is that the proposed technique is worked for the binarization of scanned document images that have no or weak slanting. Another approach for document image binarization is using local maximum and minimum filter [7].The main drawback of this paper is that the problem due to over-normalization. A Robust document image binarization technique for degraded document image [8] based on adaptive image contrast has been proposed in the last year .The main drawback of this paper is the problem due to over-binarization .This method also have some limitation related to performance over various datasets and some noise still remains in the image.

In this paper we propose a document image binarization technique that extends the previous paper based on adaptive image contrast. This method requires only minimum number of parameters related to a document image. This technique also makes use of the adaptive image contrast which is a combination of local image contrast and local image gradient. In addition, some noise reduction technique is used here to improve the quality of output document image and to extract most of the text information. Proposed method addresses over-binarization problem of the previous paper .At the same time, it shows improved performance over various datasets interms of various performance measures.



(a)

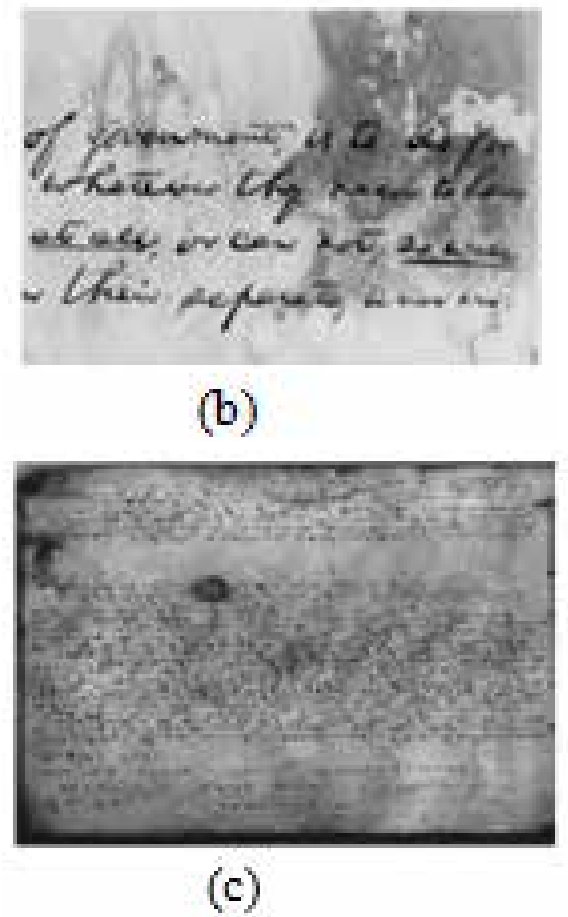


Fig 1.Three degraded document images (a)-(b) are taken from DIBCO datasets and (c) from Bickley diary dataset

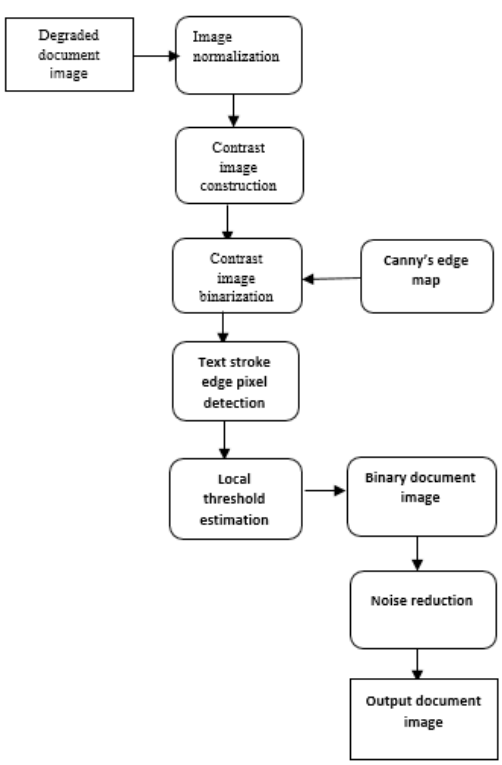


Fig.1. Block diagram of proposed system

This section describes the proposed methodology. In the proposed document image binarization technique a degraded document image is given as input.First,the input degraded document image is normalized to improve the quality of output binarized document image.Second,an adaptive image contrast map is constructed for the normalized image.Third,adaptive image contrast map is binarized and combined with canny's edge map to detect text stroke edge pixels. Then foreground text is extracted from the document background by a local threshold that is estimated based on the intensities of the text stroke edge pixels. Finally, the output binarized image is filtered to reduce noise.

A.Construction of contrast image

Primary aim of the contrast image construction is to detect text stroke edge pixels properly. In prior to the construction of adaptive image contrast map for the input degraded document image input image is normalized to improve the quality of output binarized image. Adaptive image contrast is a combination of local image contrast and local image gradient.

Adaptive image contrast [8] is given as follows:

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

Where *C(i,j)* denotes the local contrast that is calculated as follows :

$$C(i,j) = \frac{I_{\max}(i,j) - I_{\min}(i,j)}{I_{\max}(i,j) + I_{\min}(i,j) + \epsilon}$$

$(I_{max}(i,j)-I_{min}(i,j))$ refers to the local image gradient. Where ε is a positive infinitely small number that is added in case the local maximum is equal to zero.α is the weight between local image contrast and local image gradient that is estimated using a power function.

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

Where *std* refers to the document image intensity standard deviation .In the previous paper γ is arbitrarily close to 1.When γ is 1, the histogram will be linear .And the image will become brighter if γ is brought closer to the histogram peak. In this paper we first find the histogram and then adjust γ bringing it closer to peak (either setting γ>1 or γ<1,which ever bring the curve to edge of the histogram peak) making image brighter which removes background details and then histogram equalization is applied. This will enhance the edge contrast of strokes eliminating the background.

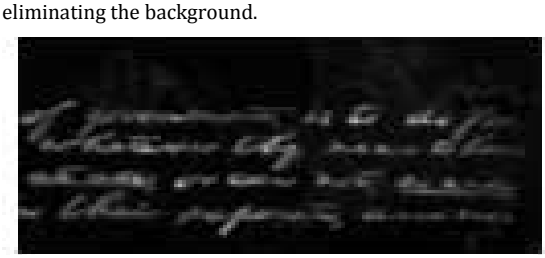


Fig 2: Contrast image constructed using proposed method of the sample document image in Fig 1(b)

B. Detection of text stroke edge pixels

We can extract the foreground text from the document background once the high contrast edge pixels are detected properly. Text Stroke edge pixels can be detected easily by using, previously constructed contrast image. Adaptive image contrast computed at the text stroke is considerably higher than that computed within document background. Contrast map is then binarized using a global thresholding method which can extract the stroke edge pixels properly. Local image contrast and local image gradient are calculated within a local window and are evaluated by the difference between the maximum and minimum intensities in a local window.

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

Binarized contrast map is then combined with Canny's edge map .Canny's edge detector can detect the edges close to the real edge locations in the detecting image. The combined map consists of only those pixels that appear within both high contrast image pixel map and Canny's edge map .This helps to extract stroke edge pixels accurately.

C.Estimation of Local Threshold

After high contrast text stroke edge pixels are detected properly, we can segment the foreground text from the document background by a local threshold that is estimated based on the intensities of the detected text stroke edge pixels. If we analyze different kinds of document images we can observe that the text pixels are close to the detected text stroke edge pixels and there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The foreground text [8] can be extracted based on the intensities of the detected text stroke edge pixels as follows:

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

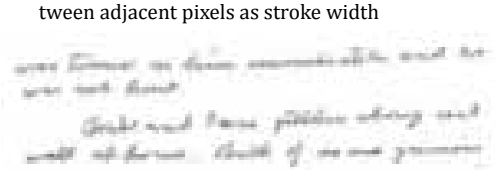
Where E_{mean} and E_{std} denote the mean and standard deviation of the intensity of the detected text stroke edge pixels within a local neighborhood window, respectively. The size of the local window should be larger than the stroke width. Therefore, we can set the local window size based on the stroke width *EW* of the document image. For that, we use an edge width estimation algorithm. By using this algorithm we calculate the most frequently occurring distance between two adjacent edge pixels.

Algorithm 1 Edge Width Estimation

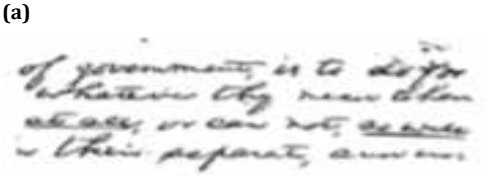
Input: Degraded document image *I* and corresponding Binary Text Stroke Edge Image *Edg*

Output: Edge Width *EW*

1. Get the *width* and *height* of *I*
2. In the edge image, for each *Row i*=1 to height in *Edg* do
3. Scan from left to right, edge pixel candidates are selected as follows:
   a) Its label is 0 (background)
   b) The next pixel is labeled as 1(edge)
4. Examine the intensities of the selected pixels. Remove those pixels that have 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, then calculate the distance between pixels in the pair.
6. End for
7. Construct a histogram of calculated distances
8. Use this estimated most frequently occurring distance between adjacent pixels as stroke width



(a)



(b)

Fig 3.Combined edge map corresponding to Fig 1(a) and 1(b) respectively.

D. Reduction of noise

Extraction of foreground text from the document background requires binarizing the image, which discard most of the noise and replace the character and background pixels with binary 0 and 1 respectively. Then the binarized image is filtered to reduce noise. A document to be scanned can itself be contaminated with noise .Sometimes, scanning itself introduces some noise. Noise may be due to dust, spot, degeneration, ageing etc. In order to improve the quality of output binarized document image, scanned document image is to be freed from existing noise. This can be done by contrast adjustment, noise suppression and many others. For noise reduction we can use smoothing operations in document image. Smoothing operations can be used to reduce the noise or to straighten the edges of the characters in the binary document image. For example, to fill the small gaps or to remove small bumps in the edges of the character. Smoothing and noise removal can be done by filtering. In filtering operation, value of any pixel in the output image is determined by applying some algorithm in the values of the pixels in the neighborhood of the corresponding input pixel.

III.RESULT AND ANALYSIS

The proposed paper discusses a document image binarization technique. It involves only minimum number of parameters. It has been tested on various images in DIBCO datasets and Bickley diary datasets .The proposed document image binarization technique combines the local image contrast and local image gradient .The binarization performance of the proposed method is evaluated in terms [9] –[10] of F-measure, PSNR, Negative Rate Matric (NRM) and Miss classification Penalty Metric (MPM).PSNR of proposed method is considerably higher than the previous methods. Hence the proposed method extracts the text better than previous methods. Value of F-measure, MPM and NRM are more close to the previous best performing methods. The proposed method also solve the over-binarization problem in the previous paper.

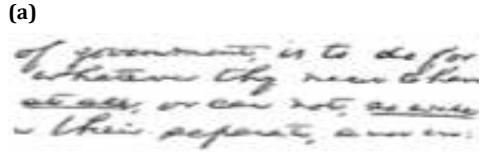
Table 1 Evaluation results of the dataset of dibco 2009

Methods	F-Measure(%)	PSNR	NRM($\times 10^{-2}$)	mpm($\times 10^{-2}$)
LMM	91.06	18.5	7	0.3
Based on Adaptive image contrast	93.5	19.65	3.74	0.43
Proposed method	93.8	20.15	3.33	0.39

TABLE 2 Evaluation results of the bickley diary dataset

Methods	F-Measure	PSNR	NRM($\times 10^{-2}$)	mpm($\times 10^{-2}$)
LMM	66.44	10.76	17.50	72.08
Based on Adaptive image contrast	78.54	13.15	12.92	16.71
Proposed method	79.15	15.11	12.87	16.70

(a)



(b)

Fig 4.Output document images corresponding to Fig 1(a) and (b)

IV. CONCLUSION

This paper presents a document image binarization technique

based on adaptive image contrast .The output binary document image is further improved by filtering this image to reduce noise. Proposed technique is simple and efficient, only minimum number of parameters are involved. Proposed method has been tested on various datasets and the method outperforms previous methods in term of the PSNR, F-measure, MPM, and NRM. This method can be applied to any kind of degraded document images.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support and facilities provided by Department of CSE, Vedavyasa Institute of Technology. Authors also extend their thanks to the Head of the Department for her immense help during the course of the project.

REFERENCE

[1] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding tech- nique for image segmentation," in Proc. IEEE Trans. Image Process., Jun. 1998, pp. 918–921. | | [2] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Iterative multimodel subimage binarization for handwritten character segmentation," IEEE Trans. Image Process., vol. 13, no. 9, pp. 1223–1230, Sep. 2004. | | [3] Y. Chen and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," IEE Proc. Vis., Image Signal Process., vol. 152, no. 6, pp. 702–714, Dec. 2005 | [4] T. Lelore and F. Bouchara, "Document image binarisation using Markov field model," in Proc. Int. Conf. Doc. Anal. Recognit., Jul. 2009, pp. 551–555.. | | [5] B. Gatos, I. Pratikakis, and S. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in Proc. Int. Conf. Pattern Recognit., Dec. 2008, pp. 1–4. | [6] 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. | | [7] 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 | | [8] Bolan Su, Shijian Lu, and Chew Lim Tan,"Robust document image binarization for degraded document images" ieee transactions on image processing, vol. 22, no. 4, april 2013 | | [9] 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. | [10] 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.. |