Image Binarization for Degraded Document Images

Sushilkumar N. Holambe
Persuing Ph.D.
At Department of CSE, Dr. B A.M. University,
Aurangabad, India

Ulhas B. Shinde, PhD
Dean
Faculty of Engineering & Technology, Dr. B. A. M. University,
Aurangabad, India

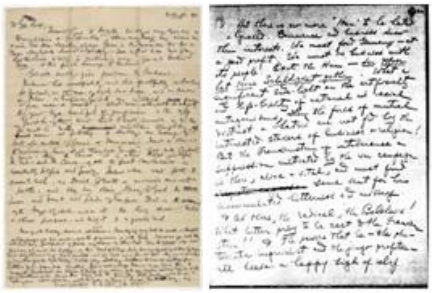
Bhagyashree S. Choudhari
Persuing ME (CSE), T.P.C.T's College of Engineering,
Osmanabad, India

ABSTRACT
Image binarization is the separation of each pixel values into two collections, black as a foreground and white as a background. Thresholding technique is used for document image binarization. Image binarization plays vital role in segmentation of text from the document images that are badly degraded due to the high inter\intra variations between the foreground text of document images and document background. This paper, proposes technique to address the issues of degraded images using adaptive image contrast. The adaptive image contrast technique is a combination of the local image contrast and the local image gradient. And they are tolerant to variation of text and background. Such variations are caused by number of document degradations. The proposed technique, constructs adaptive contrast map for degraded image .the contrast map is combined with Canny's edge map, for the identification of text stroke edge pixels. Thresholding technique can be applied as global technique and local technique. Global thresholding is suitable for a document where there is uniform contrast delivery of background and foreground. However global thresholding fails to the applications where difference in contrast, Extensive background noise and difference in brightness exists. in such circumstances categorization of many pixels as a foreground or as a background is not so easy. Local thresholding plays significant role in such cases. Local thresholding technique uses local threshold t; w.r.t .local window to segment the document image .this local threshold t is estimated based on the intensities of detected text stroke edge pixels. The proposed method is simple, robust, and involves minimum parameter tuning. It has been tested on three public datasets that are used in the recent document image binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010.

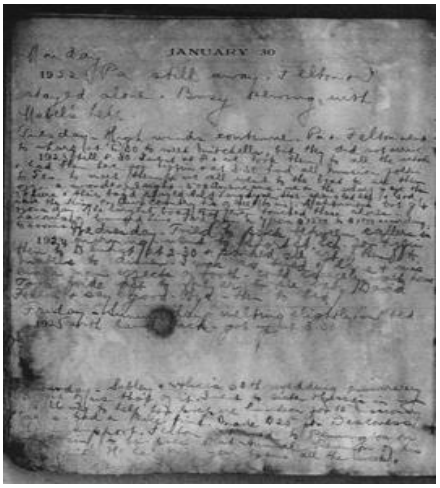
General Terms
Binarization, Degraded historical documents

Keywords
Adaptive image contrast, document analysis, document image processing, degraded image, image binarization, pixel classification, Contrast Image, Canny Edge Detector, Local threshold Segmentation.

1. INTRODUCTION
Historical Documents are degraded due to paper quality, aging, storage conditions. The information contained in historical documents must be preserved therefore efforts are taken at National and international levels .this can be done by using a technique known as binarization binary image is such a digital image that has just two values for every pixel. Two colors are used to represent these two values, i.e. black and white however any other colors can also be used. Therefore ground color is mainly representing the object whereas background color rest of the image. Separation foreground and background of documents images is the pre-processing step for the document analysis, carried out by Binarization. Gray-scale document image is converted into a binary document image i.e. with two values only. In document image processing like applications, binarization technique which is fast and accurate is very important to ensure correctly processing tasks of document images ,such as optical character recognition (OCR).calculation of thresholding for degraded document images is an unsolved problem due to, high intra/inter variation between the document background and text stroke across different document images. As illustrated in Figure 1(a), Historical documents are degraded due to bleed through. In addition, handwritten text documents are degraded due to a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background as illustrated in Figure 1(b). In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in Figure1(c). Thresholding errors are introduced due to such different types of degradations in document images. The recent Document Image Binarization Contest (DIBCO)[1] [2], [3] 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) [4] 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 [5] performs the best among entries of 43 algorithms submitted from 35 international research groups. We also participated in the H-DIBCO 2010 and our local maximum-minimum method [6] 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 [6] and the method used in the latest DIBCO 2011. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning.



(a) **(b)**



(c)

Fig 1: Degraded document image examples (a), (b), (c). (c) Is taken from Bickley diary dataset

In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm [6]. At the same time, the parameters used in the algorithm can be adaptively estimated.

2. EXISTING SYSTEM

The Adaptive Document image binarization uses early window-based adaptive thresholding technique. In this window-based technique estimation of the local threshold is done with the help of mean and the standard variation of image pixels within a local neighborhood window. The local contrast method proposed in Bernsen's "Dynamic thresholding of gray-level images," is simple and depends upon the maximum and minimum intensities within a local neighborhood windows of an image pixel (i, j) respectively

2.1 Thresholding Algorithm Comparison

To perform the Comparison of some thresholding algorithms for text/background segmentation in difficult document images [7] this paper, introduces two new thresholding techniques. These techniques are compared with some existing thresholding algorithms. Document images with background noise or illumination/contrast variations are used for the evaluation of algorithms. The thresholding quality was assessed from resultant words in the background using Precision and Recall analysis of the same. Though all types of images are not handled by any single algorithm, each algorithm definitely works better than others for particular types of images. Appropriate algorithm(s) are combined to do task better.

2.2 Color Thresholding

Color Thresholding Method for Image Segmentation of Natural Images [11] is done with the color values in natural images. The grey level thresholding algorithm with slight modifications is used for color thresholding. Multilevel thresholding has been conducted to the RGB color object. To study color information no of natural images are used. The result is supposed to be achieved if, by using the selected threshold, object is separated from the background.For document image binarization many thresholding techniques [7]–[10] have been reported. Global thresholding [12]–[14] is usually not a suitable approach as many degraded documents do not have a clear bimodal pattern. So the local thresholding estimator for each document image pixel is suitable approach. And this can be achieved through Adaptive thresholding [14], to deal with different variations within degraded document images.

3. CLASSIFING SUB-BLOCK

Sub-block classification and thresholding [15] consist of the below listed three feature vectors to test the local regions. These local regions are classified into three types:

1. Heavy strokes
2. Faint strokes
3. Background (No stroke).

Content information is not included in the background. Lower values of edge strength and variance are covered in background. Small mean-gradient value is associated with a background though it is considered as noise-free. Strokes that are difficult to distinguish from the background are nothing but faint strokes. Heavy stroke areas have strong edge strength, more variance and larger mean-gradient value. The proposed weighted gradient thresholding method is applied to the different classes of sub block.

3.1 Faint handwritten image

3.1.1 Enhancement

To proceed further with degraded document images enhancement of faint strokes is necessary. Wiener filter was applied to avoid the noise enhancement. The stroke faint enhancement can be done as:

Step 1: Enhancing the image can be done by finding the

Maximum and minimum grey value in the 3x3

Window.

Step 2: Mini =min (Total No. Of window elements)

Maxi = max (Total No. Of window elements)

Compare "pixel – mini" and "maxi – pixel", where "pixel" is the pixel-value. If the former is greater, the "pixel" is closer to the highest grey value than the lowest value in this window; hence the value of "pixel" is set to the highest grey value ("pixel" = "maxi"). If the former is smaller, then the value of "pixel" is set to the lowest grey value ("pixel"= "mini").

3.1.2 Thresholding

To Threshold faint stroke, weighted method is used. And this method is based on mean gradient. Various directional strokes are normally available with Western-style or Handwritten English scripts.

4. PROPOSED METHOD

The proposed document image binarization technique works in no of steps and described in this section.

A. Contrast Image Construction.

B. Text stroke edge pixel detection.

C. Local Threshold Estimation.

D. Post Processing.

The proposed method can be implemented as shown in fig.2 Contrast image construction is done in the preprocessing stage.canny's edge detection method is used for the edge detection. Separation of text from the image is done through

local thresholding method.post processing shows its importance in order to improve image quality.

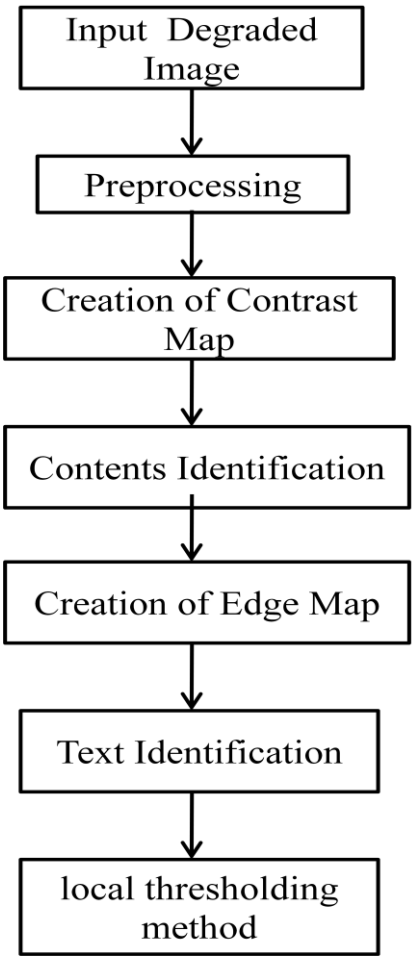


Fig 2: Proposed Method

4.1 Contrast Image Construction

Main aim of binarization is segmenting the document text from the document background. And this can be accomplished with the image features: local image contrast and local image gradient. Because the document text usually has certain image contrast to the neighbouring document background. They are used in many document image binarization techniques [2] [3] 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 over-normalization problem [1] we derive an adaptive local image contrast as:

. $C\alpha(i,j) = \alpha C(I,j) + (1-\alpha)(Imax(I,j) - Imin(I,j))$. *(1)*

The adaptive combination of the local image contrast and the image gradient in above equation can produce proper contrast maps for document images with different types of degradations.

4.2 Text stroke edge pixel detection

The contrast image construction detects the stroke edge pixels of the document text. At later stage edges are detected through canny edge detection algorithm. In this algorithm it smoothes the noise in the image then pixel at both sides of the text stroke will be selected as the high contrast pixel.

4.3 Local Thresholding Segmentation

Once the text stroke edges are detected, we calculate the most frequent distance between two adjacent pixels that are on the edge. We perform it in the horizontal direction and use it as the estimated stroke width as shown in fig.3.

4.4 Post Processing Procedure

Binarization result can be improved by using Post- Processing method. By using the algorithm of post processing we remove single pixel artifacts along the text stroke boundaries after the document thresholding as shown in fig.4.

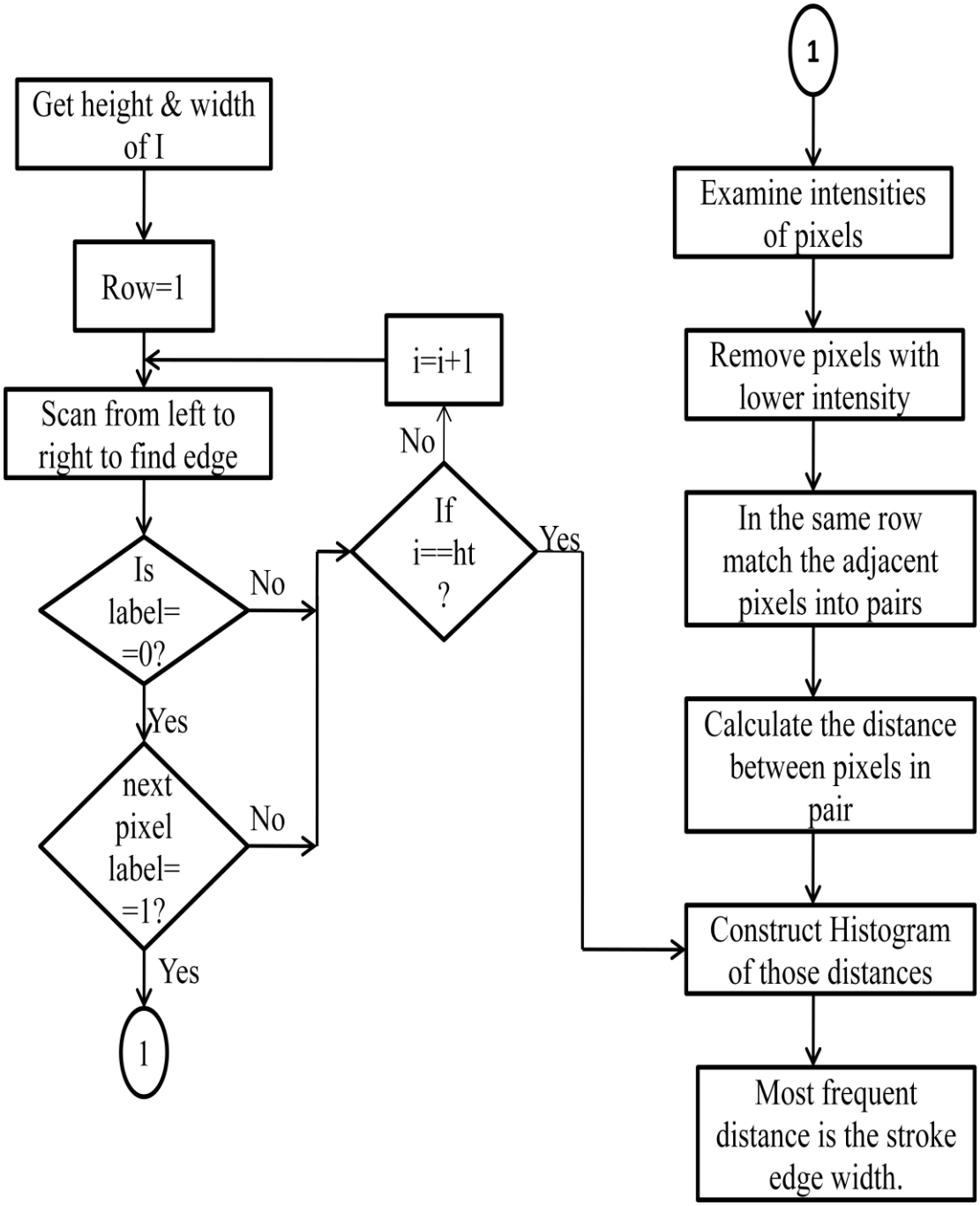


Fig 3: Local Thresholding segmentation

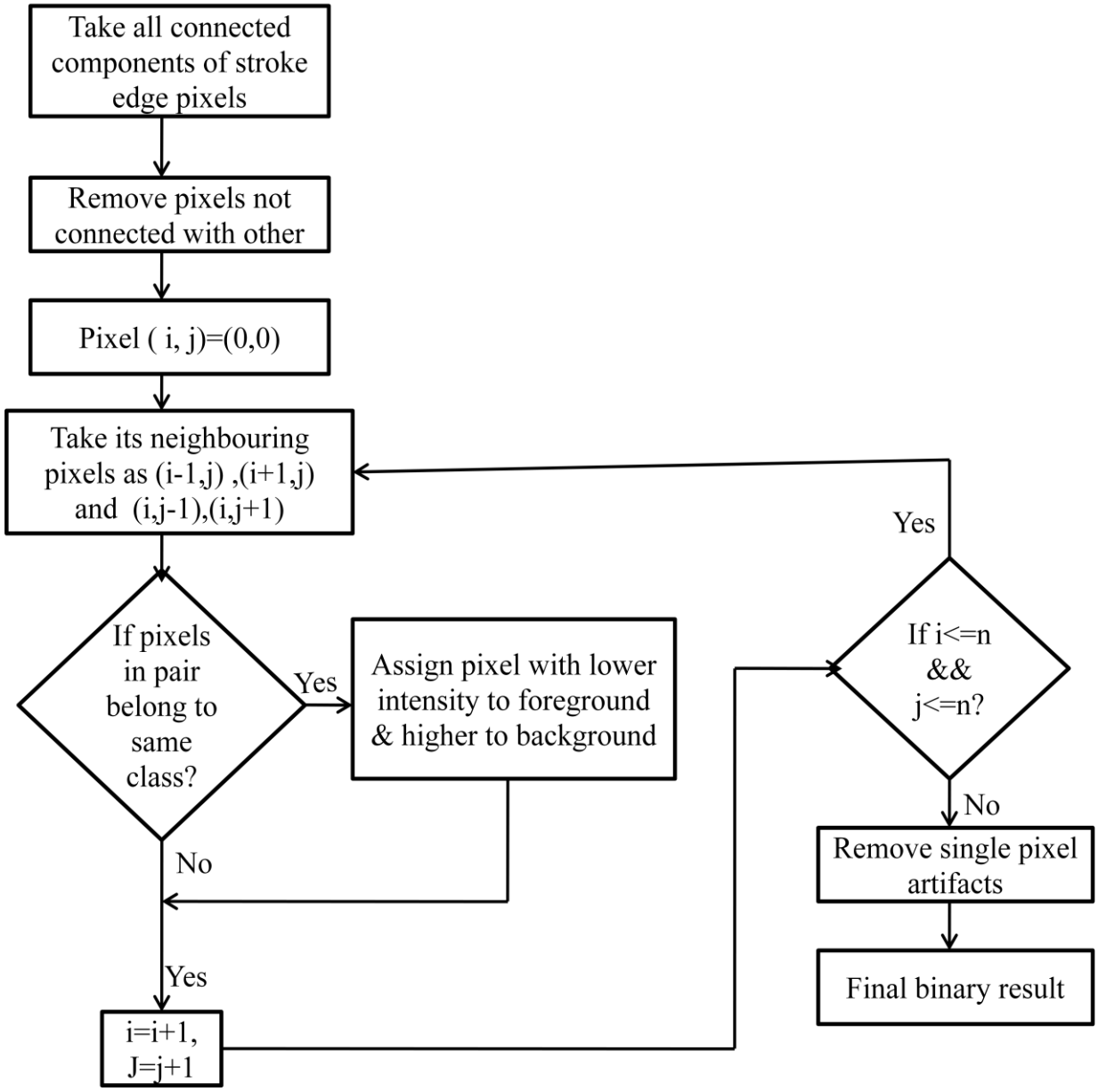


Fig 4: Post Processing

5. CONCLUSION

To deal with uneven illumination proposed method is based on adaptive image contrast. Due to the inclusion of few parameters this method is simple. Robustness of this method is advantageous to work properly with document images of different types which are degraded. This system can be applicable in all the areas that are concerned with preserving the historical documents and managing the degraded documents.

6. REFERENCES

[1] 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

[2] 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.

[3] 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

[4] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Hand writ. Recognit, Nov. 2010, pp. 727–732.

[5] 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.

[6] 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.

[7] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Main, "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.

[8] 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.

[9] 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.

[10] 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.

[11] Nilima Kulkarni,"Color Thresholding Method for Image Segmentation of Natural Images", IJIGSP, vol.4, no.1, pp.28-34, 2012.

[12] A. Brink, "Thresholding of digital images using two-dimensional entropies," Pattern Recognit., vol. 25, no. 8, pp. 803–808, 1992.

[13] 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.

[14] N. Otsu, "A threshold selection method from gray level histogram," IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.

[15] Sayali Shukla, Ashwini Sonawane, Vrushali Topale,Pooja Tiwari,"Improving Degraded Document Images Using Binarization Technique"Vol.3,pp.333-338,May.2014