**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

***IJCSMC, Vol. 4, Issue. 3, March 2015, pg.383 – 388***

**RESEARCH ARTICLE**

# RETRIEVING TEXT from DEGRADED DOCUMENT IMAGES using IMAGE BINARIZATION TECHNIQUE

1st **Vitthal D. Walimbe**, 2nd **Rahul H. Gehlot**, 3rd **Swati P. Pukale**, 4th **Vrushali S. Kudale**

[1] *Computer Engineering, Savitribai Phule Pune University, Trinity Academy of Engineering, Pune, India – 411038*

[2] *Computer Engineering, Savitribai Phule Pune University, Trinity Academy of Engineering, Pune, India – 411038*

[3] *Computer Engineering, Savitribai Phule Pune University, Trinity Academy of Engineering, Pune, India – 411038*

[4] *Computer Engineering, Savitribai Phule Pune University, Trinity Academy of Engineering, Pune, India – 411038*

1st vithal.walimbe06@gmail.com, 2nd rgehlot143@gmail.com,
3rd pukale12swati@gmail.com, 4th vrushalikudale11@gmail.com

***Abstract***

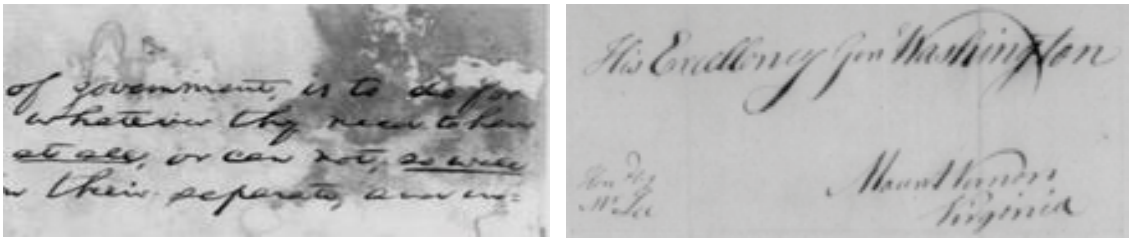
***Extracting the text from poorly damaged document images is a very difficult task as there are very high inter/intra-variations present in between the documents background and the foreground text . In this paper, we present a robust document image binarization technique which deal with these issues with the help of gray scale image contrast inversion. The image contrast inversion is performed by initially changing the provided image to evert image and then searching the contrast of the image (inverted) to distinguish context and backdrop differences caused by various forms of script damages. The following modifying technique, a flexible contrast map is first generated for the image. The grayscale image is generated from the contrast map so as to clearly identify the strokes of text from the pixels of background and foreground. Further disjoining of document text is performed by a local threshold that is calculated based on the strengths of detected strokes of text edge pixels within a bounded window. The following method is easy, powerful, and contains slightest criterion attune. Various confronting lousy trait document images also show the extraordinary work of our presented method, as to other approaches.***

***Keywords : Image contrast, gray scale image, document image processing, degraded document image binarization, pixel classification.***

## I. INTRODUCTION

Document Image Binarization is performed in the pre-processing stage for document analysis and it aims to segment the text in foreground from the background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as optical character recognition (OCR). Image binarization is been studied for several years, but then also thresholding of damaged artefact images is today also an problem not solved due to the huge inter/intra-variation between the stroke of text and the background across various images. As illustrated in Fig. 1, the hand scripted text from within the damaged documents mostly displays a certain quantity of changes in the terms of the stroke width, stroke brightness, stroke connection, and document background.



**(a)** **(b)**

Fig.1 Degraded Document Images

In addition, historical documents are often degraded by the bleed through as shown in Fig. 1(a) and (b) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artefacts as shown in Fig. 1. These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-art techniques. This paper presents a document binarization technique that extends our previous local maximum-minimum method and the method used in the latest DIBCO 2011[6]. The proposed method is simple, robust and capable of managing all sorts of damaged document images with minimum criterion attuned. It makes use of the inversion image contrast that converts the input image into invert image and then converts the invert image into contrast image and therefore is tolerant to the variations in text and background due to variety of document degradations. Precisely, the presented method deals with the excess-normalization issue due to local maximum minimum algorithm. Simultaneously, the criterion in the algorithm can be adaptively estimated. The rest of this paper is organized as follows. Section II first reviews the current state-of-the-art binarization techniques. Our proposed document binarization technique is described in Section III. Then experimental results are reported in Section IV to demonstrate the superior performance of our framework. Finally, conclusions are presented in Section V.

## II. RELATED WORK

Many thresholding techniques [1]-[3] have been reported for document image binarization. Loads of devalued documents don't have a fine bimodal pattern, overall thresholding generally aren't an efficient approach for the deteriorated document binarization. Adjusting thresholding, which estimates a parish threshold for each and every pixel in image, is usually an improved way to handle variety of variations within damaged document images. For example, previously the image pixels mean and the standard variation from within an insular neighbourhood window was used by the window-based adaptive thresholding techniques. In the window-based thresholding techniques window size matters a lot which is a main drawback because the thresholding performance depends heavily on it and hence the stroke width of character. Alternative ways reported such as, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance, and combination of binarization techniques. Image information of different types and knowledge of domain are combined in these which are generally complex. The local image contrast and gradient are very helpful miens for separating the text from the background as the text generally has particularly contrast of image to the neighbouring document background. Many document image binarization techniques have their usage and are very sufficient. The local contrast is defined as follows in Bersen's paper:

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

where contrast of an image pixel *(i, j) is denoted as C(i, j),* the maximum and minimum intensities from-within a local neighbourhood windows of *(i, j)* are denoted by Imax*(i, j)* and *I*min*(i, j ),*respectively. If the threshold is greater than local contrast *C (i, j)* , the pixel is fixed as background directly. It is classified either as text or background by comparing with the *I*max *(i, j)* and *I*min *(i, j) mean*. Bersen's method is simple, but cannot work properly on degraded document images with a complex document background. We have previously presented a unique document image binarization process with the help of the contrast of local image that is calculated as follows:

$$C(i, j) = \frac{\mathbf{I}_{max}(i, j) - \mathbf{I}_{min}(i, j)}{\mathbf{I}_{max}(i, j) + \mathbf{I}_{min}(i, j) + \epsilon} \qquad (2)$$

here € is a +ve but boundlessly timid no. that is included if the value of local maximum is equal to 0. Compared with Bersen's contrast in Equation 1, in Equation 2 due to local image contrast a factor of normalization (the denominator) to compensate the image variation within the document background is introduced. Take the text within shaded document areas such as that in the sample document image in Fig. 1(a) 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

The devised document image binarization technique is explained in this section. Provided a devalued document image, an invert contrast map is constructed initially and thus stroke edges of text are then recognised by the conversion of contrast image to grayscale. Disjoining of text is done on the basis of local threshold, calculated from the text stroke edge pixels that are detected. Little after-processing is applied further to enhance the binarization quality of document.

### *A. Contrast Image Construction*

The gradient of image is been extensively used particularly for edge detection of the document images effectively that have an orderly background of document. Most often it detects the background containing non-stroke edges of degraded document that contain main image variations as a result of uneven lighting, bleed-through, noise etc. Only to extract the proper stroke edges, the normalization (inversion) of the image needs to be done to compensate the variation within the document background. Previous method, evaluated local contrast with the help of local image maximum and minimum that is used to curb the variation in background as explained in 2nd Equation. Mainly, the local image difference i.e. numerator (i.e. the discrepancy in between the local maximum and minimum) is captured which is akin to the conventional image gradient. The normalization factor i.e. denominator that curbs the variations of image within the background of document. Considering image pixels present in bright regions, a large normalization factor is produced so that the numerator is neutralised and accordingly result is obtained in a relatively low image contrast. Considering the image pixels present within dark regions, a less denominator is produced and it results in a relatively high image contrast accordingly. However, there is one typical limitation in image contrast of Equation 2 which is handling document images with the bright text properly may be is impossible. It happens so as result of a weak contrast is estimated for bright text stroke edges wherein the Equation 2 the numerator will be small and the denominator will be large. This over-normalization problem is swamped, by converting the input devalued image into invert image where the image colour pixels are inverted according to 256 bitmap colours. The conversion of inverted image takes place into contrast image to get clarification in between background and foreground pixels.

The contrast is found as follows:

$$Ca(i,j) = \alpha C(i,j) + (1 - \alpha)(I\max(i,j) - I\min(i,j)) \quad (3)$$

The proposed binarization technique relies more on inverted image and deflect the extra normalization issue of the old methodology.

Section IV displays the map of contrast of the sample images in Figure.1 that are achieved using gradient and, contrast of local image as well as of our proposed method in Equation 3, resp. For the illustration document in Fig. 1 with a mosaic document background, using the local image contrast produces an improved result as displayed in Fig. 2(a) as compared to the result by the gradient of local image (because in Equation 2 the normalization factors help to overcome the upper left area noise of Fig. 2(a)). In the sample document of Fig. 1(b) having variations of small intensity within the background but variations of high intensity within the strokes of text, the local image contrast is used to eradicate many light strokes present in the contrast map improperly as illustrated in Fig. 2(b) but the use of local image gradient preserves those light strokes of text as shown in Fig. 2(b). As a result, the attune combination of the local image contrast as well as the image gradient of Equation 3 is capable of producing exact map of contrast for document images with various types of deterioration as illustrated in Fig. 2(b). Peculiarly, the contrast of local image in Equation 3 brings a huge weight for the document image in Fig. 1(a) with large intensity variation from within the document backdrop whereas the gradient of local image achieves a large weight for the image in Fig. 1(b).

(a) (b)

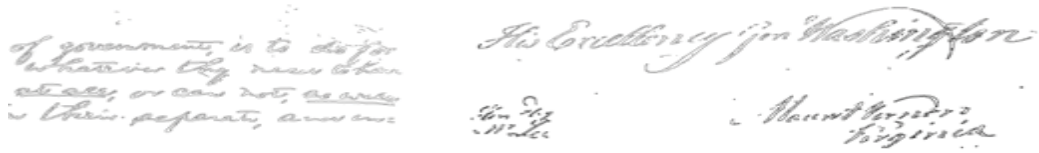
Fig.2 contrast map of the sample document images

### *B. Text Stroke Edge Pixel Detection*

Aim of the image contrast development is to find the edge pixels of stroke of the text correctly. The developed contrast image has a crystal clear bi-modal pattern, where the image inversion contrast calculated at edges of text stroke is certainly huge than the computed from within the background of document. Thus we spot the edge pixel of text stroke with the Otsu's overall

thresholding way. The images contrast in Fig. 2(b), a binary map by Otsu's algorithm shown in fig 3(a) which excerpts the edge pixels stroke appropriately. As the local image contrast and gradient are revised through the change in-between the intensities of maximum and minimum in a local window, the both sides pixel of the text stroke will be chosen as the high contrast pixels. Thus, prior to implementing the Otsu's global thresholding algorithm, initially we proselyte the image into grayscale because the gray scale version of the image contrast have the excellent variation in-between pixels of background and foreground. Finally edge image map of text stroke, we maintain solo pixels that occur within the image pixels of high contrast in the image of grayscale. The conversion of grayscale aid to cite the edge pixels of text stroke meticulously as depicted in Fig.3 (a) & (b) .



(a) (b)

Fig.3 Text Stroke Edge Pixel Detection

*C. Local Threshold Estimation*

Extraction of text from the document background pixels takes place once the stroke edge pixels of high contrast are detected properly. Two characteristics can be observed from different kinds of document images. Initially, the detected text stroke edge pixels are close to the text pixels. Secondly, the stroke edge pixels of high contrast and the background pixels bordering have a distinct intensity variations. The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

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

here the mean and standard deviation are "*E*mean" and "*E*std" of the intensities of the detected text stroke edge pixels from-within a neighbourhood window *W*, respectively. The stroke width must be smaller than the neighbouring window in order to hold pixels of stroke edge. Based on the stroke width of the document image the size of the neighbourhood window *W* can be set under study. From the detected stroke edges EW can be estimated [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 labelled 0 (background) and the pixels next to them are labelled to 1 (edge) in the edge map (*Edge*), 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. Construction of histogram is done so that the occurrence of the span in-between two adjoining competitor pixels is recorded. The stroke edge width *EW* can then be approximately estimated by using the most frequently occurring distances of the 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: read width and height of I
2: for each n every Row where x = 1 to height in Edge do
3: Check from left to right to search edge pixels that fit the following criteria:
a) The label is 0 (background);
b) The succeeding pixel is named as 1(edge).
4: Check the intensities in I of that pixels elected in Step 3, and discard that pixels which have a lessened intensity than the successive pixel next to it in the equivalent row of I.
5: In the same row into pairs match the rest neighboring pixels, and determine the length the two pixels have in between them.
6: for ends
7: Generate a histogram of the calculated distances.
8: adopt the most regularly occurring span as the approximated stroke edge width EW.

*D. Post-Processing*

Once the initial binarization result is derived from Equation5 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 neighbourhood 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 labelled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artefacts along the text stroke boundaries are filtered out by using several logical operators as described in [4].

**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 Edge.
2: Remove those pixels that do not connect with other pixels.
3: for Each remaining edge pixels (x, y ): do
4: Get its neighborhood pairs: (x − 1, y) and (x + 1, y );
(x, y − 1) and (x, y + 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* .

## IV. EXPERIMENTS AND DISCUSSION

Designing of some experiments is done to perform the effectiveness and robustness of our proposed method. Analysing the working of the recommended approach on public datasets for parameter selection is performed first. Some datasets lack the ground truth data, each and every metrics are not applied on each and every single image.

*A. Parameter Selection*

The γ increases from $2^{-10}$ to $2^{10}$ exponentially and monotonically. In particular, $\alpha$ is close to 1 when $\gamma$ is small and the local image contrast *C* dominates the adaptive image contrast *Ca* in Equation 3. On the other hand, *Ca* is mainly influenced by local image gradient when $\gamma$ is large. At the same time, the variation of $\alpha$ for different document images increases when $\gamma$ is close to 1. Under such circumstance, the power function becomes more sensitive to the global image intensity variation and appropriate weights can be assigned to images with different characteristics. Data set improves significantly when $\gamma$ increases to 1. Therefore the proposed method can assign more suitable $\alpha$ to different images when $\gamma$ is closer to 1. Parameter $\gamma$ should therefore be set around 1 when the adaptability of the proposed technique is maximized and better and more robust binarization results can be derived from different kinds of degraded document images. Fig. 4 shows the thresholding results when *W* varies from *EW* to 4*EW*.

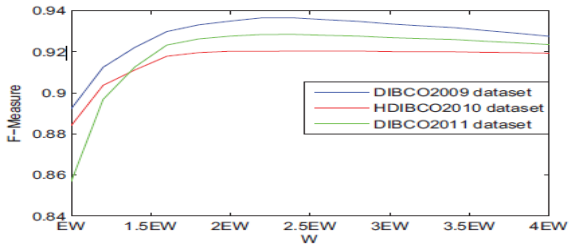



Fig.4 Threshold result when *W* varies from *EW* to 4*EW*

Generally, a larger local window size will help to reduce the classification error that is often induced by the lack of edge pixels within the local neighbourhood window. In addition, the performance of the proposed method becomes stable when the local window size is larger than 2*EW* consistently on the three datasets. *W* can therefore be set around 2*EW* because a larger local neighbourhood window will increase the computational load significantly.

**Fig.5 Thresholding results**

*B. Discussion*

Considering the description in preceding sections, method stated contains various criterions, many of them could be certainly predicted on the basis of enumerations of the document image given. Thus our proposed tactic attains higher stability and is user-friendly for various genres of image degradation. The exceptional pursuance displayed by our recommended method can be elucidated by distinct aspects. Basically, the local image contrast and gradient is joined in our stated method which helps in the backdrop variation suppression and the extra-normalization of images having lower variation is averted. Subordinately, the merger with edge map benefits in producing a particular map of edge of text stroke. Triennially, the edges of text stroke are used which help to excerpt the forepart from the artefact backdrop precisely.

## V. CONCLUSION

An adaptive contrast of image based document image binarization technique is presented in this paper which is lenient to various sorts of document damages namely irregular beam and document blur. Presented approach is easy and sturdy, just hardly any criterions are present. On the top of it, it executes for various genres of damaged document images. The contrast of local image which is calculated on the basis of the insular maximum and minimum is used in our technique. The presented mechanism is been used to scrutinise different datasets.

## ACKNOWLEDGEMENT

We are deeply indebted to our Head of the Department Prof. S.N. Maitri and to our Project Co-ordinator as well as guide Prof. C.P. Kedia for their unwavering moral support and motivation during the entire course of the project.

## REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.