ISSN:0975-9646

Savita Borole et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1759-1763

Text Extraction Using Adaptive Thresholding

Savita Borole, Minal Thobde, Reshma Hore, Shraddha Shinde

Computer Department, University of Pune
Alard College Of Engineering, Hinjewadi, India.
XploreTech India
Hinjewadi, Pune, India.

***Abstract* - Segmentation of text from poorly documented images is a very difficult task due to high mutation between the document background and foreground text of various document images. In this paper, a binarization technique is significantly designed for historical document images. This existing binarization technique points either on finding an appropriate global threshold for each area in order to remove strains, smear and uneven illuminations. In binarization process an adaptive contrast map is first constructed for an input degraded document image. Adaptive image contrast is a combination of local image contrast and local image gradient. This contrast map is then binarized and combined with Canny's edge map to detect text stroke edge pixels. The document is further segmented by local threshold that is estimated based on the intensities of detected text stroke edge pixel within that local window. This method is simple, robust and includes minimum parameter tuning. Our approach applies a global threshold and detects image areas that are more likely to still contain noise. Each of these areas is reprocessed separately to achieve better quality of binarization.**

***Keywords* - global threshold, local image contrast, local image gradient, Canny's edge map, adaptive image contrast, text stroke edge pixel.**

I. Introduction

Documents are the most essential part of any organization today whether they are stored electronically or stored on papers. When we store them electronically it is the most convenient way as we can preserve them efficiently. In case of storing them on papers, it becomes quite difficult to preserve them for a long time.

We are going to introduce a method to preserve these documents more effectively as well as more efficiently. Image binarization technique is the mostly used and accepted method to extract information from degraded images. Though it is studied for many years, still there is a need of progress. It becomes quite difficult to extract text from the degraded documents due to degradation issues such as stains, blood seeps or ink seeps, smears, noise etc. In such cases, accuracy of extracting text lacks somewhere. In case of handwritten text, it is again a big headache to maintain the accuracy of the extracted text. There are some figures to illustrate the actual causes of degradation (Refer Fig 1).

This paper represents a document binarization technique which is simple, robust and efficient. This method is capable of handling different kinds of degraded images with minimum parameter tuning. It makes use of adaptive image contrast that combines the local image contrast and the local image gradient so that it can work efficiently on different types of degraded documents.

We have organized this paper as follows. Section II, describes the current techniques and works related to image binarization techniques. Section III, describes our approach to perform image binarization more effectively. In Section IV, we have concluded all the study related to this approach.

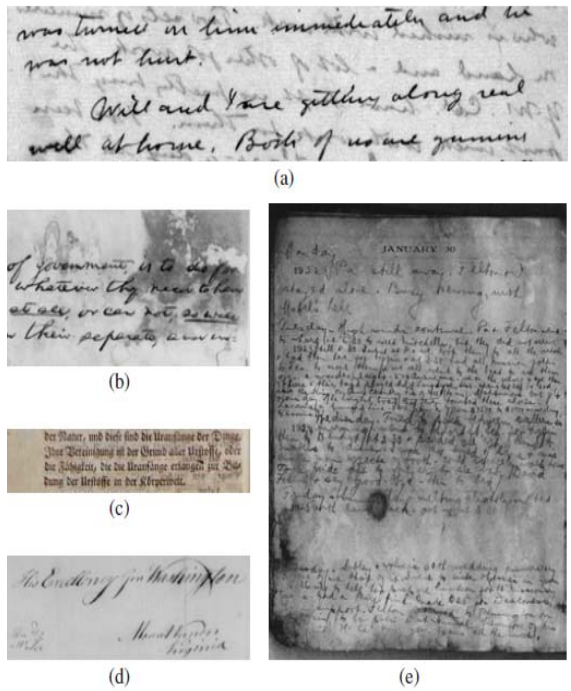

Fig 1 Examples of some degraded images. (a)-(d) are taken from DIBCO series and (e) is taken from Bickley Diary Dataset.

II. Related Work

We have seen most of the binarization techniques which works on the global thresholding concept. As per our study, global thresholding technique is not a suitable approach to handle degraded documents binarization. Instead of that adaptive thresholding technique calculates local threshold for each pixel of the degraded document so it is the most suitable approach among all. Mainly adaptive thresholding technique is suitable for all kinds of degraded documents.

In window based adaptive thresholding technique local threshold is calculated using the mean and standard variations of image pixels within a neighbourhood window. This method is a traditional one. As the thresholding is heavily dependent on window size, performance is heavily dependent on character stroke width so it is the main drawback of this technique. There are some techniques too such as background subtraction, recursive method, self-learning, laplacian energy, contour completion, decomposition method, matched wavelet, Markov Random field, cross section reference graph analysis, matched wavelet, texture analysis and combination of binarization. These methods are related to different domain knowledge and skills and works on different types of information. These methods are quite complex too.

Local image contrast and local image gradient are the most useful features to segment text from the document background because document text has certain image contrast to the neighbouring document background. Due to their effectiveness, they are mostly used in many document image binarization techniques. As per Bernsen's paper, local image contrast is defined as:

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

where C (i, j) denotes the contrast of an image pixel (i, j), Imax (i, j) and Imin (i, j) denotes the maximum and minimum intensities within a local neighbourhood windows of (i, j) respectively. The pixel is set as a background directly if the local contrast c(i, j) is smaller than threshold. Otherwise it will be classified into text or background by comparing with the mean of Imax (i, j) and Imin (i, j). This method is simple but cannot work properly on a complex degraded document background.

The local image contrast is further modified as:

$$C\,(i, j) = \frac{Imax\,(i,j) - Imin\,(i,j)}{Imax\,(i,j) + Imin\,(i,j) + \varepsilon} \qquad (2)$$

where ε is a positive but infinitely small number hat is added in case of local maximum is equal to zero. Equation 2 introduced the normalization factor in denominator to reduce the image variation within the document background [1].

III. Proposed Method

In this section we are going to describe our approach to segment text from degraded images. For a given degraded document image at first an adaptive contrast map is constructed. Then by combining this adaptive contrast map and Canny's edge map, the text stroke edges are detected. Based on local threshold which is calculated from the detected text stroke edge pixels, text is segmented.

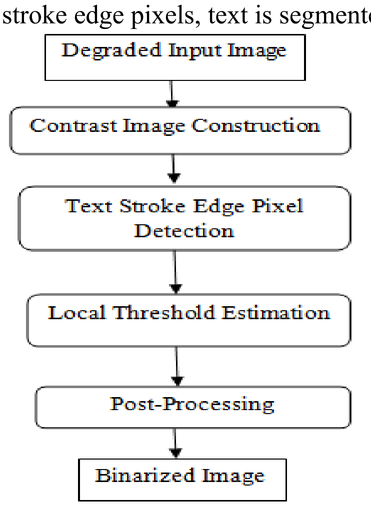

Fig. 2 Block diagram

A. *Contrast Image Construction*

The image gradient has been used to great extent for edge detection [11] and detects the text stroke of the document images effectively which have the uniform document background.

Then again, it often notices many non-stroke edges from the background of degraded documents that frequently contains certain image variations due to noise, uneven lighting, bleed-through, atmospheric conditions, etc. To draw out only the stroke edges properly, the image gradient is required to be normalized to compensate the image variation within the document background.

In this method [10] the local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as depict in equation 2. Actually, the numerator contains the local image difference that is similar to the traditional image gradient [11]. The denominator is a normalization component that subdues the image variation within the document background. Large normalization factor will be produced for the pixels in the bright regions to neutralize the numerator and accordingly result in a relatively high image contrast. However, the image contrast in equation 2 will be large but the numerator will be small, so to overcome this over-normalization issue we have combined the local image contrast with the local image and have driven an adaptive local image contrast as follows[1]:

$$Ca\,(i, j) = \alpha\, C(i, j) + (1 - \alpha)\,(Imax\,(i, j) - Imin\,(i, j)) \qquad (3)$$

where C(i,j) denotes the local image contrast in equation 2 and (Imax (i, j) – Imin (i, j)) refers to the local image gradient that is normalized to [0,1]. The local window size is set to 3 theoretically. Α is the weight between local contrast and gradient that is controlled based on the document image statistical information. Possibly, the image contrast will be assigned with a high weight (i.e. : large α) when the document image has significant intensity variations. So that the proposed binarization technique depends on more on the local image contrast that can capture the intensity variation well and hence produce good results. Or else, the local image gradient will be assigned with a high weight.

We form the mapping from document image intensity variation to α by a power function as follows:

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

where Std represents document image intensity standard derivation, and γ is pre-defined parameter. The power function has a fine property in that it monotonically and easily increases from 0 to 1 and its shape can be easily controlled by different γ. This γ can be selected from [0, ∞], where the power function becomes a linear function when γ=1. Hence, the local image gradient will play the major role in equation 3 when γ is large and the local image contrast will play the major role when γ is small.

B. *Noise Detection*

The Gaussian noise algorithm [13] is fast and efficient with less computational complexity for removing noise. This initially estimates the amount of noise disruption from the noise corrupted image. In the later stage, the central pixel is replaced by mean value of some surrounded pixels based on threshold value. Removing noise with edge preservation and computational complexity are two conflicting images which are used to detect noise. This method is an optimum solution for noising and de-noising. This method makes it easy to implement in hardware. The noise deviation of image is estimated by using Immerker's fast technique [14]. The difference between the central pixels and surrounding pixels in the filtering process is obtained by subtracting each element in the process with central pixels. This difference will be very large when the image is highly corrupted. The difference is compared with the thresholding technique.

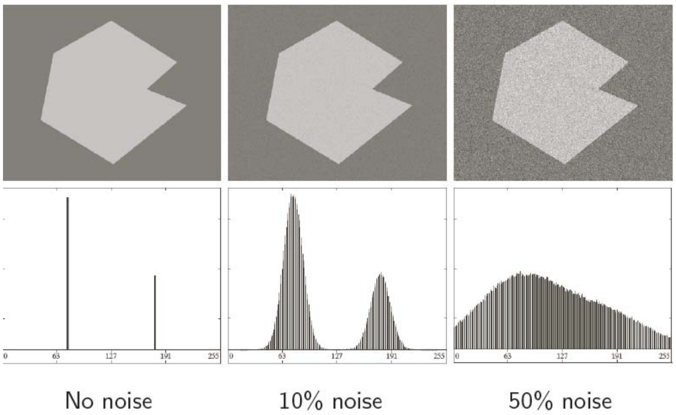

Fig 3. Noise detection

C. Text Stroke Edge Pixel Detection

The purpose of this module is to find the stroke edge pixels of the document text. The constructed contrast image has a clear bi-modal pattern [10] where adaptive image contrast computed at text stroke edge is larger than that computed within that document background. Therefore we detect the text stroke edge pixel candidate by using Otsu's global thresholding algorithm.

Algorithm 1: Global Thresholding

Step 1: Initial estimate of T.

Step 2: Segmentation using T:

. G1, Pixels greater than T;

. G2, Pixels darker than (or equal to) T.

Step3: Computation of the average intensities m1 and m2 of G1 and G2.

Step 4: New threshold value:

$$T_{new} = \frac{m1 + m2}{2}$$

Step 5: If |T - T new| > ΔT, Back to step 2, otherwise stop.

For the contrast images in fig shows a binary map by Otsu's algorithm that derives the text stroke edge pixels. As the local image contrast and local image gradient are calculated by the difference between maximum and minimum.

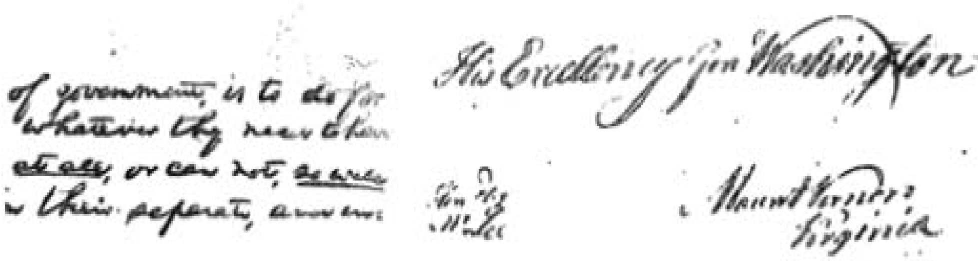

Fig 4. Binary contrast map

Fig 5. Canny's edge map

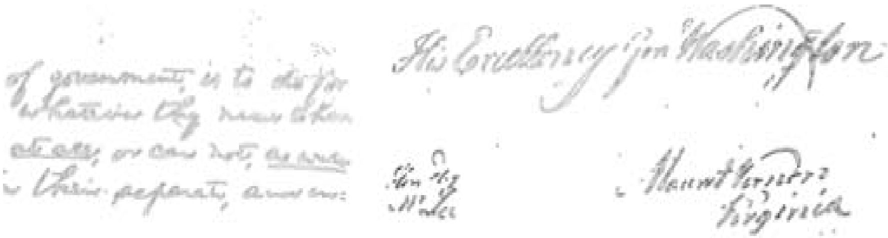

Fig 6. Combined edge map

D. Local Threshold Estimation

The text can then be taken out from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics that can be observed from different kinds of document images [10]: Firstly, the text pixels are close to the detected text stroke edge pixels. Secondly, 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 based on the detected text stroke edge pixels as follows [1]:

$$R(x,y) = \begin{cases} 1 & I(x,y) \leq Emean + \frac{Estd}{2} \\ 0 & otherwise \end{cases} \quad (5)$$

Here Emean and Estd are mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighbourhood window W, respectively.

The neighbourhood window should be at least larger than the stroke width in order to contain stroke edge pixels. So the size of the neighbourhood window W can be set based on the stroke width of the document image under study, EW, which can be estimated from the detected stroke edges Since, we do not need a precise stroke width, we just need to 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. Firstly, the edge image is scanned horizontally row by row and the edge pixel candidates are selected.

Algorithm 2: Edge width evaluation

Require: Input Degraded Document Image I, Corresponding Binary Text Stroke Edge Image Edg
Ensure: The Evaluated Text Stroke Edge Width EW
Step 1: Get width and height of I.
Step 2: For each row assign i=1 to height in edge do
Step 3: Scan from left to right which fulfils the following criteria:
    Label is 0 then set to background
    Next pixel labelled as 1 which is edge
Step 4: Calculate intensities of I of those pixels selected in Step 3
    Lower intensity pixels are then removed which presents next to it in same row
Step 5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
Step 6: end for
Step 7: Construct histogram of those calculated distances.
Step 8: Use the most frequently occurring distance as The Evaluated Stroke Edge Width EW.

Now the edge pixels, which are labelled 0 (background) and the pixels next to them are labelled to 1(edge)in the edge map(Ed g), are properly detected, they should have higher intensities than the following few pixels (which should be the text stroke pixels). 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. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width EW can then be roughly estimated by using the most frequently occurring distances of the adjacent edge pixels.

E. Post Processing

Once the initial binarization is obtained from local threshold estimation as described, the binarization result can be further implemented by algorithm 3 [1]. First the separated foreground pixels that do not connect with each other are removed to make edge pixels set precisely. Second the neighbouring pixel pair those co-insides on both the side of text stroke edge pixel should belong to different classes. One pixel of the pixel pair is then labelled to the other category if both of the two pixels belongs to the same class finally some single pixel along the text stroke boundaries are filtered out by using various logical operators.

Algorithm 3: Post-Processing Procedure

Require: Input Degraded Document Image I, Initial Binary Result B And Corresponding Binary Text Stroke Image Edg.
Ensure: The Final Binary Result BF
Step1: Find all the connected components of the Text Stoke Edge Pixels in Edg.
Step2: Remove those pixels that do not connect with other pixels.
Step3: For each remaining edge pixels (i,j) : do
Step4: Get its neighbourhood pairs: (i-1,j) and (i+1,j);
    (i, j-1) and (i, j+1)
Step5: If the pixels in the same pairs belong to the same class (both text and background) then
Step6: Assign the pixels with lower intensity to foreground class (text), and other to background class.
Step7: end if
Step8: end for
Step9: Remove the single pixel Artifacts along the Text Stroke boundaries after the document Thresholding.
Step10: Store the new binary result to BF.

Table I. Comparative Results

Rank	Method	F-measure
1	Adaptive thresholding for maximum and minimum method	93.9%
2	Adaptive degraded document method	90.3%
3	Adaptive logical method	80.0%
4	Bernsen's adaptive method	86.3%
5	Savoula's adaptive method	85.5%
6	Otsu's global method	82.0%

IV. CONCLUSION

The proposed method makes use of adaptive image contrast so it is capable of handling different types of degraded images. This method is simple, robust and effective as it takes few parameters to perform the desired task effectively. There are some other techniques too to which works on local thresholding as well as global thresholding but the proposed method is the convenient one. As per study and experiments, there is a need of enhancement to achieve more accuracy.

ACKNOWLEDGEMENT

The Authors would like to thank our principal, Dr. Annasaheb Goje, our Head of department, Mrs. Kalpana Saharan and other college staff members for their valuable guidance and support.

REFERENCES

[1] Bolan Su, Shijian Lu, And Chew Lim Tan,(2013). *Robust Document Image Binarization Technique For Degraded Document Images*. [Online].Available:http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6373726

[2] Yibing Yang, Hong Yan, (2000). *An Adaptive Logical Method for Binarization of Degraded Document Images.* [Online]. Available: http://www.researchgate.net/publication/222662311_An_adaptive_logical_method_for_binarization_of_degraded_document_images

[3] Otsu, N (1979). *A Threshold Selection Method From Gray-Level Histograms.*[Online].Available:http://www.citeulike.org/group/1650/article/1116982

[4] Bernsen, J, (1986). *Dynamic Thresholding Of Grey-Level Images*. [Online].Available:http://academic.research.microsoft.com/Paper/2042781.aspx

[5] Niblack, W (1986). *An Introduction To Digital Image Processing*. [Online]. Available: http://dl.acm.org/citation.cfm?id=4901

[6] Sauvola, J., Pietikainen, (2000*), Adaptive document image binarization*, [Online].Available:http://www.sciencedirect.com/science/article/pii/S0031320399000552

[7] Couasnon, B., Camillerapp, J., Leplumey, I (2004). *Making handwritten archives documents accessible to public.* [Online]. Available:http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?tp=&arnumber=1263255&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F8926%2F28251%2F01263255.pdf%3Farnumber%3D1263255

[8] Marinai, S., Marino, E., Cesarini, F., Soda, G (2004). *A general system for the retrieval of document images from digital libraries.* [Online].Available:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.4515

[9] B. Su, S. Lu, and C. L. Tan (2010). *Binarization of historical handwritten document images using local maximum and minimum filter.* [Online]. Available: http://dl.acm.org/citation.cfm?id=1815351

[10] D. Ziou and S. Tabbone (1998). *Edge detection techniques.* [Online]. Available: http://wwwmath.tau.ac.il/~turkel/notes/Maini.pdf

[11] Canny, (1986). *A computational approach to edge detection.* [Online]. Available:http://dl.acm.org/citation.cfm?id=11275

[12] V.R.Vijaykumar, P.T.Vanathi, P.Kanagasabapathy (2010). *Fast and Efficient Algorithm to Remove Gaussian Noise in Digital Images.* [Online].Available:http://connection.ebscohost.com/c/articles/48459116/fast-efficient-algorithm-remove-gaussian-noise-digital-images

[13] R. Garnett, T. Huegerich and C. Chui (2005). *Noise Removal Algorithm With An Impulse Detector.* [Online]. Available: http://www.cis.rit.edu/~cnspci/references/dip/filtering/garnett2005