ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Review on Text Extraction Using Adaptive Thresholding

Savita Borole

Assistant Professor, Dept. of C.S.E., AISSMS IOIT, Pune, India

ABSTRACT: Extraction 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 In this paper, a binarization technique is significantly designed for historical document images.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 method is simple, robust and requires 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, text stroke edge pixel

I. INTRODUCTION

This method is useful to preserve old documents more effectively. 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.

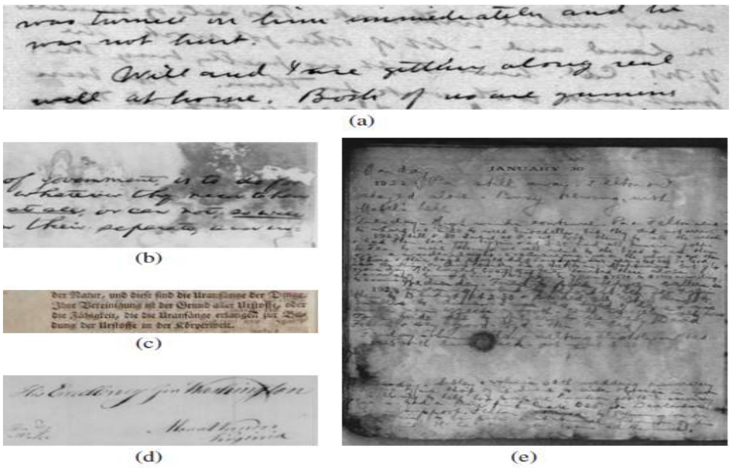

Fig- 1: Examples of some degraded images. (a)-(d) are taken from DIBCO series and (e) is taken from Bickley Diary Dataset

II. RELATED WORK

We have seen most of the binarization techniques which works on the global thresholding concept. As per my study, global thresholding technique is not a suitable approach to handle degraded documents. 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 neighborhood 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 neighboring 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:

$$C\,(i, j) = Imax\,(i, j) - Imin\,(i, j) \qquad (1)$$

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:

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

here 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 ALGORITHM

In this section here 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.

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. A is the weight between local

DOI: 10.15680/IJIRCCE.2016. 0409103

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 more on the local image contrast that can captures the intensity variation well and hence produces good results. Or else, the local image gradient will be assigned with a high weight.

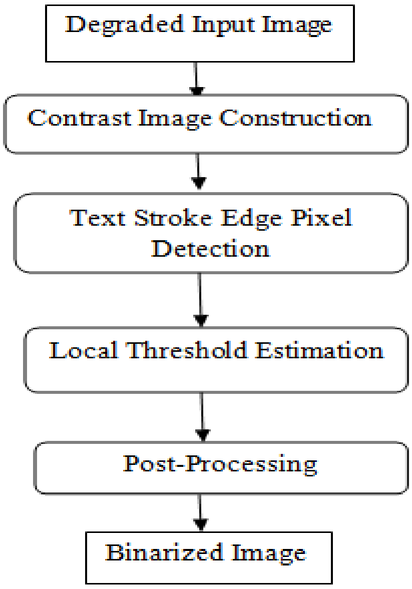
Degraded Input Image
Contrast Image Construction
Text Stroke Edge Pixel Detection
Local Threshold Estimation
Post-Processing
Binarized Image

Fig- 2: Flow of project

B. 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 document background. Therefore, we detect the text stroke edge pixel candidate by using Otsu's global thresholding algorithm [3].

```
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:
        $Tnew = \frac{m1 + m2}{2}$
Step 5: If |T - T new| > ΔT, Back to step 2, otherwise stop.
```

For the contrast images, 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.

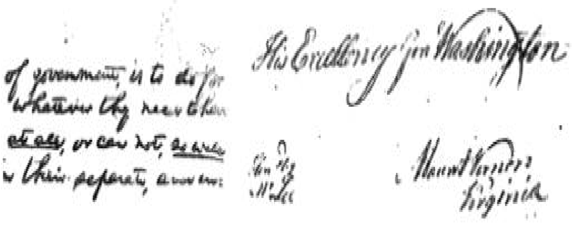

Fig- 3: Binary contrast map

DOI: 10.15680/IJIRCCE.2016. 0409103

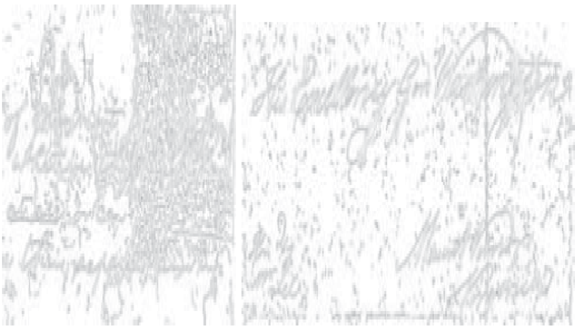

Fig- 4: Canny's edge map

C. LOCAL THRESHOLD ESTIMATION

The text can be taken out from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics those 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 is based on the detected text stroke edge pixels as follows [1]:

$$R(x,y) = \left\{\begin{matrix} 1 & I(x,y) \leq Emean + \frac{Estd}{2} \\ 0 & otherwise \end{matrix}\right\} \qquad (4)$$

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 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

IV. CONCLUSION AND FUTURE WORK

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 which work 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.

REFERENCES

1. Bolan Su, Shijian Lu, And Chew Lim, “Robust Document Image Binarization Technique for Degraded Document Images ” [Online] Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6373726, (2013).
2. Yibing Yang, Hong Yan, “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 (2000).
3. Ostu N, “A Threshold Selection Method from Gray Level Histograms” [Online] Available: http://www.citeulike.org/group/1650/article/1116982. (1979).
4. Bernsen J, ”Dynamic Threshold of Gray Level Images” [Online] Available: http://academic.research.microsoft.com/Paper/2042781.aspx. (1986).
5. Niblack W, “An Introduction to Digital Image Processing” [Online].Available: http://dl.acm.org/citation.cfm?id=4901. (1986).
6. Sauvola J, Pietikainen, “Adaptive Document Image Binarization” [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320399000552, (2000)
7. Couasnon B, Camillerapp J, Leplumey I, “Making handwritten Archieves 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, (2004).
8. Marinai S, Marino E,Cesarini F, Soda G,”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. (2004).
9. B Su,S Lu and C L Tan,”Binarization of Historical Handwritten Document Images using Local Maximum And Minimum Filter” Available: http://dl.acm.org/citation.cfm?id=1815351, (2010).
10. D Ziou and A Tabbone,”Age Detection Techniques” .[Online]. Available: http://wwwmath.tau.ac.il/~turkel/notes/Maini.pdf, (1998).
11. Canny," A computational approach to edge detection"[Online]. Available: http://dl.acm.org/citation.cfm?id=11275 , (1986).
12. V.R.Vijaykumar, P.T.Vanathi, P.Kanagasabapathy "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,(2010).
13. R. Garnett, T. Huegerich and C. Chui "Noise Removal Algorithm With An Impulse Detector"[Online] Available: http://www.cis.rit.edu/~cnspci/references/dip/filtering/garnett2005.pdf,(2005).

BIOGRAPHY

Savita Borole is a Assistant Professor in the Computer Engineering Department, AISSMS Institute of Information Technology Pune, India. She received Master of Engineering (ME) degree in 2013 from Government College of Engineering, Aurangabad, India. Her research interests are Image Processing, Algorithms, Data Mining etc.