Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 6, June -2017

A Review Paper on : Foreground Text Extraction Mechanism Using Document Image Binarization

Samhita K Kulkarni [1], Mrs. S.S.Patil [2]

[1]PG Research Student, Department of Computer Science & Technology, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India.

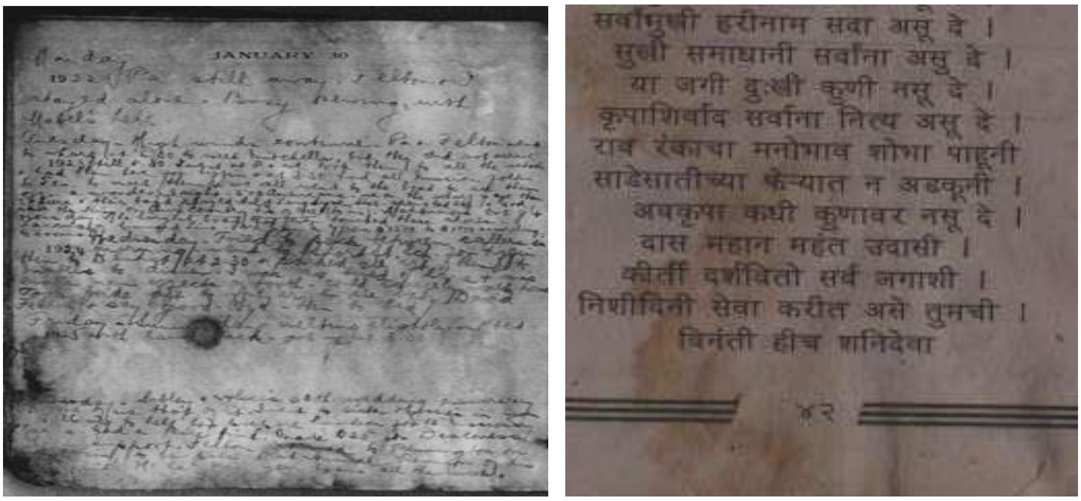
[2]Assistant Professor, Department of Computer Science & Technology, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India.

Abstract — *Now-a-days, there are many activities which depend upon the internet. And there is a great need to shift all the activities which are performed by user towards the digitization of world. Many times it happens that institutes and organizations have to maintain the books or novels for a longer time span and there arises a new challenge for the institutes. Books being a physical object, so it will definitely have the issues of wear and tear. The pages definitely get degraded and so does the text on the pages. The data on the pages can be confidential and sensitive and there should be robust and dynamic mechanism for preserving the data on the same. Due to this degradation many of the document images are not in readable. So, there is a need to separate out text from those degraded images and preserve them for future reference. To make this we have proposed binarized documentation technique. In this method first adaptive contrast map is constructed for input degraded document image then text stroke edges are detected and local threshold is used for text segmentation. Then post-processing is applied to improve document binarization quality.*

Keywords- *Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.*

I. INTRODUCTION

Image processing is a famous and most interested area for researchers. All of our general activities are connected with the image and its processing. Historical documents such as novels or scripts or confidential data and documents are being preserved by storing them into an image format. Separating text and background from poorly degraded document images is a challenging task between the document background as well as the foreground text of various document images due to the higher background variation. Due to low quality papers, documents fail to preserve the text written on it and gradually the text becomes unreadable. Sometimes the documents get degraded due to some natural problems. There should be an efficient technique to recover these degraded document images for future use.



JANUARY 30

सर्वांमुखी हरीनाम सदा असू दे ।
सुखी समाधानी सर्वांना असु दे ।
या जगी दुःखी कुणी नसू दे ।
कृपाशिर्वाद सर्वांना नित्य असू दे ।
राव रंकाचा मनोभाव शोभा पाहूनी
साडेसातीच्या फेऱ्यात न अडकूनी ।
अपकृपा कधी कुणावर नसू दे ।
दास महान महंत उदासी ।
कीर्ती दर्शविती सर्व जगाशी ।
निशीदिनी सेवा करीत असे तुमची ।
विनंती हीच शनिदेवा
४२

Figure 1: Example of Degraded Document Images

This method presents image binarization technique for the better and accurate recovery of such document images. The main function of Binarization of image is performed to separate the foreground text from the document background. A document image binarization method is important for recovering document image with the help of processing tasks such as adaptive Contrast map. Instead Canny’s edge algorithm, grayscale method is used to sharpen the edges found by Otsu thresholding and edge detection method.

II. LITERATURE SURVEY

Sr. No.	Title	Author	Detail Description
1.	Robust Document Image Binarization technique for degraded document images	Bolan Su, Shijian Lu, and Chew Lim Tan	In this system the input image goes through different methods. These methods include contrast inversion, threshold estimation and binarization. Even though it passes through these all techniques, it is not producing efficient output. The edge detection done by the canny's method is not much efficient to detect all the text strokes. The produced output still contains some background pixels.
2.	Document image binarization using background estimation and stroke edges	S. Lu, B. Su, and C. L. Tan	This technique first estimates a document background surface through an iterative polynomial smoothing procedure. The text stroke edge is further detected from the compensated document image. Finally, the document text is segmented by a local threshold that is estimated based on the detected text stroke edges.
3.	Binarization of historical handwritten document images using local maximum and minimum filter	B. Su, S. Lu, and C. L. Tan	This makes use of the image contrast defined by the local image maximum and minimum. Compared with the image gradient, the image contrast evaluated by the local maximum and minimum has a nice property that it is more tolerant to the uneven illumination.
4.	A fast adaptive method for binarization of document images	L. Eikvil, T. Taxt, and K. Moen	This method is evolved by multi resolution approximation; with considerably lower computational complexity, the threshold surface is constructed and is smooth, yielding faster image binarization and better visual performance.
5.	Adaptive document image binarization	J. Sauvola and M. Pietikainen	The contrast value of the text background and the text are focused here. There are two different approaches to find the threshold which are soft decision method (SDM) and text binarization method (TBM). The capabilities of SDM has noise filtering and tracking of signal, To separate text components from background of the image the TBM is used, due to uneven illumination or noise which is in bad conditions format. At last, the output of these two algorithms is combined together.

III. ALGORITHM

A. Algorithm: Edge Width Estimation

Requirements: The Image I is the Input Document Image and Edge is the corresponding Binary Text Stroke Edge Image.

Ensure: EW is the Estimated Text Stroke Edge Width

```
1: Store the width and height of Image I
2: Then for Each Row i in Image I = 1 to height in Edg do
3: to find edge pixels scan the Image from left to right that meet the following criteria:
a) If its label is 0 (background);
b) If the next pixel is labeled as 1(edge).
4: pixels which are selected in Step 3, Check the intensities in I of those pixels, and the pixels that have a minimum concentration than the coming pixel cut out that next within the same row of I.
5: Then the remaining adjacent pixels are matched into pairs in the same row, and then distance between two pixels in pair will find.
6: end for
7: A histogram for those calculated distances is then calculated.
8: Then as the estimated stroke edge width EW use the most frequently occurring distance.
```

B. Algorithm: Post-Processing

Ensure: The Final Binary Result Bf

```
1. Find all the connect components of the stroke edge pixels in Edge
2. Eliminate those pixels that are not connected with other pixels.
3. For Each remaining edge pixels (i, j): do
4. Get its neighborhood pairs: $(i − 1, j)$ and $(i + 1, j)$; $(i, j − 1)$ and $(i, j + 1)$
5. If the pixels in the same pairs belong to the same class (both text and background) then
6. Allot the pixel with lower intensity to foreground class (text), and the other to background class.
7. End if
8. End for
9. Eliminate single-pixel artifacts along the text stroke boundaries after the document thresholding.
10. Store the new binary result to Bf.
```

IV. METHODOLOGY

The system consists of five modules: Contrast image construction, Text stroke edge pixel detection, Local threshold estimation, Binary conversion, Post processing. Given a degraded document, initially the contrast image is constructed which then determines the edge strokes of the text document. Text is segmented based on the local threshold which is estimated from the detected text stroke pixels. It is further converted to binary form. Finally post processing is done in order to improve the efficiency of the resultant image.

A. Contrast Image Construction

Contrast is the difference in luminance and/or color that makes an object clear. In visual approach of the real world, within the same field of view, contrast is the variant in the color and intensities of the object and other objects. The image gradient gives better results for documents that have a uniform background. But it identifies many non-stroke edges from the document background. To extract only the stroke edges properly, the image gradient must be normalized to compensate the image variation among the document background. The local image contrast and local image gradient are useful methods for segmentation of text from document background. To overcome over-normalization problem the local image contrast and the local image gradient have been combined and the equation of the adaptive image contrast given as follows:

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

Where, $C\,(i, j)$ denotes the local contrast, $(Imax\,(i, j)-Imin(i, j))$ refers to the local image gradient. The local windows size is set to 3.$\alpha$is the weight between local contrast and local gradient. We model the mapping from document image intensity variation to α by a power function as follows: $\alpha = (Std/128)$ ^γ where Std denotes the document image intensity standard deviation, and γ is a pre-defined parameter. The local image gradient will play the major role when γ is large and the local image contrast will play the major role when γ is small.

B. Text Stroke Edge Pixel Detection

We obtain the stroke edge pixels of the document text properly from contrast image construction. The constructed contrast image consist a clear bi-modal pattern. For detection of the edges of each pixel we are using otsu edge detection algorithm. The contrasted image which is further processed for edge detection. This will produce the border of the pixel around the foreground text. Pixels are classified into two parts, background pixels and foreground pixels. A foreground pixel is the area included within text stroke. And a background pixel is the degraded pixel. From text stroke image construction we obtain the stroke edge of the predicted text patterns found on the degraded document. For performing clustering based image thresholding the Otsu's method is very useful.

Grayscale Conversion:

The Edge Stroke Image obtained from the second module is then transformed to image that are grayscale so as to sharpen the edges of the text stroke detected and thereby increase the efficiency of the further modules.

The gray scale method is the most convoluted. The most common grayscale conversion routine is "Averaging", and it works like this:

Gray = (R+ G + B) / 3

Where, R is Red, G is Green, B is Blue. Equivalent to grayscale, this formula works nice, and it is very simple to propose and optimize because it is simple. However, this formula works poorly for representing shades of gray. We need something better. The proposed system uses Luminance grayscale method that is more suitable for enhancing the text strokes. Luminance grayscale method is as shown below:

Gray = (Red * 0.21 + Green * 0.71 + Blue * 0.072)

C. Local Threshold Estimation

Once the high contrast stroke edge pixels are detected properly the text can then be extracted from the document background pixels. Characteristics can be observed from different kinds of document images are: 1.It will detect text pixels which are close to the text stroke edge pixels. 2. There is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The neighborhood window should be at least larger than the stroke width in order to contain stroke edge pixels. So the size of the neighborhood 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 as stated in Edge Width Estimation Algorithm.

D. Convert To Binary

The threshold estimated image is then converted into binary format i.e. 1 and 0.The image pixels at background are marked as 0 and image pixels at foreground are marked as highest intensity and then Combining both to form a bimodal clear image.

E. Post-Processing

After deriving the initial binarization result in previous method that binarization result can be further improved by using Post processing procedure algorithm.

V. CONCLUSION

The proposed system is based on recovering the degraded document contents. The binarization technique is more efficient .The proposed system is an adaptive method to recover the contents from any degraded document. This is a very simple and efficient technique with any sort of document. The main advantage is that it is not language specific. It can recover any language contents. The application is useful in many fields like forensics, historical department, government organizations etc. Thus we can conclude that this method can create more efficient output. This can become very useful to retrieve original data from degraded documents.

REFERENCES

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, “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, pp. 1375–1382.

[3] 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.

[4] 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.

[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] 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] 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.

[8] 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.

[9] L. Eikvil, T. Taxt, and K. Moen, “A fast adaptive method for binarization of document images,” in Proc. Int. Conf. Document Anal. Recognit., Sep. 1991, pp. 435–443.

[10] J. Sauvola and M. Pietikainen, “Adaptive document image binarization,” Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.

[11] B. Gatos, I. Pratikakis, and S. Perantonis, “Adaptive degraded document image binarization,” Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[12] Y. Liu and S. Srihari, “Document image binarization based on texture features,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 5,pp. 540–544, May 1997

[13] 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.