OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

ROBUST DOCUMENT IMAGE BINARIZATION TECHNIQUE FOR DEGRADED DOCUMENT IMAGES

SHAIKH SAIQUA IRAM

PG Student EES Computer Science and Engineering, Aurangabad

Abstract: Segmentation of text from badly degraded document image is a very challenging task due to the high inter/intra- variation between the document background and the foreground text of different document images. In this project, we propose a novel document image binarization technique that addresses these issues by using adaptive image contrast. The adaptive image contrast is a combination of the local image contrast and the local image gradient that is tolerant to text and background variation Caused by different types of document degradations. In the proposed technique, an adaptive contrast map is first constructed for an input degraded document image. The contrast map is then binarized and combined with Cannys edge map to identify the text stroke edge pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum parameter tuning. Experiments on the Bickley diary dataset that consists of several challenging bad quality document images also show the superior performance of our proposed method, compared with other techniques.

Keywords: Binarization, Text Stroke Detection, Local Thresholding,Adaptive Contarst, Local Gradient, Post Processing

----------∴∴∴----------

I INTRODUCTION

The proposed system presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved.

Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. Preserve objective is to induce application and sensitive data that is on verge to get degraded.

- To preserve ancient document.
- Many of the Documents related to government deeds, or official documents, Ancient novels, or sensitive information containing documents, are to be preserved over a longer period of time and so there is dense possibility that these documents may get degraded.
- We need to preserve these documents and thereby development of the proposed system is very important.

II LITERATURE SURVEY

There Cluster based image threshold used for gray level image to binary image reduction. These methods combine different types of image information and domain knowledge and are often complex. These algorithms try to extract combined text with help of assuming two combined images having two different pixel. It assumes that an image follows a bimodal histogram i.e. it contains foreground and background pixels. Then it will be calculated threshold to extract two images to ensure that two spreading images are minimal. This method gives acceptable results when the pixels in each class are close to each other. Limitation of this system there is images not clear accurately by bimodal pattern. Second limitation is minimization of intra class variance between class scatter [3] Niblack's algorithm [4] calculates a pixel wise threshold by sliding a rectangular window over the gray level image. The threshold T is computed by using the mean m and standard deviation s, for all the pixels within the window, and this threshold is denoted as:

$T = m + k \times s$

Where k is a constant, which determines how much of the total print object edge is retained, and has a value between 0 and 1. The value of k and the size SW of the sliding window defines the quality of binarization [9].The limitation of Niblack's method is that the resulting binary

image suffers from a great amount of background noise especially in areas without text.[10].Another approach for document images binarization has been adopted by Savona [5] .In this method the page is considered as a collection of subcomponents such as text, background and picture. To define a threshold for each pixel of the background and pictures a soft decision method is used. The neighborhood window should be at least larger than the stroke width in order to contain stroke edge pixels. Pixel of the both sides of the text stroke will be selected as the high contrast pixels. To define a threshold for each pixel of textual and line drawing areas a text binarization method is used. Finally the Results of these algorithms are combined.[5].Although this method solves the problem posed by Niblack's approach but in many cases the characters become extremely thinned and broken.[10] In Bunsen's method [6] the local image contrast is defined as follows

: $C(i, j) = Imax(i, j) - Imin(i, j)$

Where $C(i, j)$ denotes the contrast of an image pixel (i, j), $Imax(i, j)$ and $Imin(i, j)$ denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively. If the local contrast $C(i, j)$ is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of Imax (i, j) and I min(i, j). Where is a positive but infinitely small number that is added in case the local maximum is equal to zero. Local images differences has been detected by some numerical like local minimum and local maximum which is similar to image gradient. Denominator behave as normalization which is lower the image factor contrast and brightness variation. In dark region around the text boundary for the image pixel, denominator is small and accordingly results in a relatively low image contrast. which compensates the small numerator and accordingly results in a relatively high image contrast. [7]. The limitation of this method is that, it cannot handle document images with bright text having bright background properly. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background.

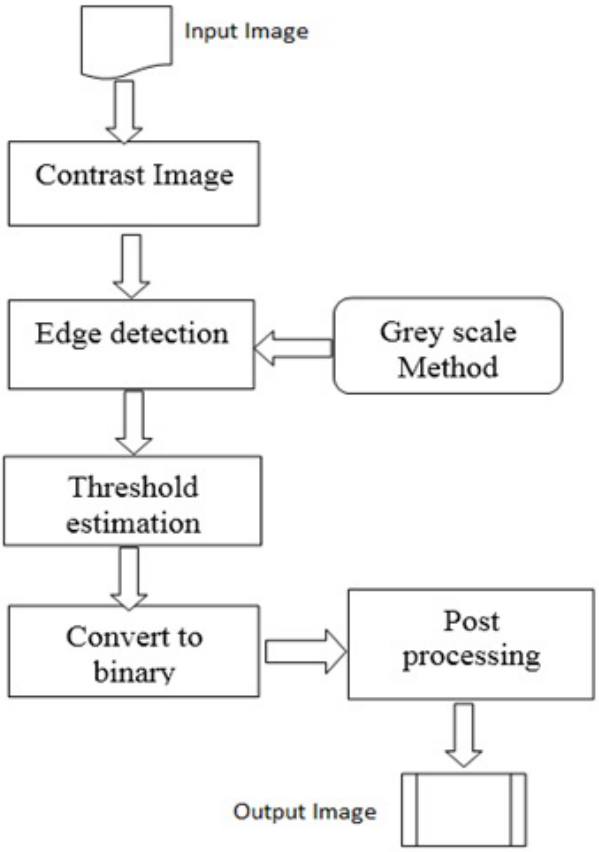
A weak contrast is calculated for image pixels having bright text stroke edges and which lie within bright regions. A large denominator and as small numerator are produced for documents having bright text stroke edges and which lie within bright regions. Where the power function becomes a linear function therefore, the local image gradient will play the major role. When is large and the local image contrast will play the major role when is denominator small. The setting of parameter will be discussed in shows the contrast map of the sample document images in Fig 2 This problem is known as over normalization problem. The proposed technique overcomes the over normalization problem by assigning weights to image contrast and image gradient.

III PROBLEM STATEMENT

The main problem with document image binarization is that the text recovery of the existing system is not enough efficient. Document binarization is an active research area for many years. The choice of the most appropriate binarization algorithm for each case proved to be a very difficult procedure itself.

In this paper, we propose a new technique for validation of document binarization algorithm. Our method is simple in its implementation and can be performed on any binarization algorithm since it doesnt require anything more than the binarization stage.

Document binarization processing task, very useful to document analysis system. Its automatically converts the document image in bi-level from such way that the foreground information represented black pixel and background by white ones.



The proposed system proposes document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is

estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

The proposed techniques have some limitations. To overcome these limitations our system uses new binarization technique along with grey scale method. There are four modules in our system. To detect the exact text stroke it is very necessary to adjust the level of contrast in the image. In this module we are keeping the image contrast at min or max level. It is depend upon how much the foreground text is mixed with background noise in the image. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post processing is further applied to improve the document binarization quality. Above figure shows the diagrammatic representation about architecture of proposed system, it has five different stages from input given up to output.

Here we invert the current level of image contrast i.e. here we are reversing the color of the image. The contrasted image is further match with grey scale method output image, which further will produce the outline of the pixel around the foreground text. These pixels then divided into two categories. First category is related pixels and second is non-related pixels. A connected pixel occupies the area around text stroke. And a non-related pixel shows the other noisy area present in the image. The edge detected image is then converted into binary format of 0's and 1's. 0 indicates that the image pixels are non-connected pixels and 1 indicates that image pixels are connected pixels and the represents the text strokes. The pixel 0s are eliminated from the processing image because they are part of background image. Output of the binarization method creates separation in the image. In post processing 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 neighborhood 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 labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators. So post processing eliminates the non-strokes image from binary image. And it returns clear images which consist of only text actual strokes. Now this produced image is when compare with input image, then we can easily figure out the significance of our system. Output image contain clean and readable text.

IV CONCLUSION

The proposed method is simple binarization method, which produces more clear output. It can be work on many degraded images. This technique uses contrast enhancement along with threshold estimation. We introduced new module post processing

Which will remove the background degradations found in the binarized image. In this technique we are going to used grey scale method to create outlined map around the text. The output of this system produces separated foreground text from collided background degradation. For that we have maintain the contrast level at min and max level. This will help to make more clear and readable output.

REFERENCES

[1] I.-K.Kim, D.-W. Jung, and R.-H. Par, "Document image binarization based on topographic analysis using a water ow model," in Proc. 7th Pattern Recognit., vol. 35, no. 1, pp. 265-277, 2002.

[2] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mia, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images"

[3] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Josh, "Iterative multimodel subimage binarization for handwritten character segmentation," IEEE Trans. Image Process., vol. 13, no. 9,pp. 1223-1230, Sep. 2004.

[4] M. Sezgin and B. Sanku, "urvey over image thresholding techniques and quantitative performance evaluation," J. Electron. Image., vol. 13, no. 1, pp. 146165, Jan. 2004.

[5] B. Gatos, K. Ntirogiannis, and I. Pratikaki, "ICDAR 2009 document image binariza- tion contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 13751382.

[6] B. Su, S. Lu, and C. L. Ta, "Binarization of historical handwritten document images using local maximum and minimum filter," in Proceedings of Advances in Cryptology EUROCRYPT 03, ser. LNCS, vol. 2656. Springer, 2003, pp. 416432.

[7] S. Lu, B. Su, and C. L. Ta, "Document image binarization using background estimation and stroke edges," Int. J. Document Anal. Recognit., vol. 13, no. 4, pp. 303314, Dec. 2010.

[8] Pratikakis, B. Gatos, and K. Ntirogianni, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. 2010, pp. 727732.

[9] Pratikakis, B. Gatos, and K. Ntirogianni, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 15061510.

[10] Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Membe, "Robust Document Image Binarization Technique for Degraded Document Images," IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.