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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

RDIB Technique to Recover Degraded Document Image

Ghodekar Neha, Bodake Shraddha, Kakade Pradnya

UG Student, Department of Computer Engineering, Savitribai Phule University of Pune, Pune, India

UG Student, Department of Computer Engineering, Savitribai Phule University of Pune, Pune, India

UG Student, Department of Computer Engineering, Savitribai Phule University of Pune, Pune, India

ABSTRACT: In this digital world of technology, we are interconnected to each other via a soft and strong internet medium. Our Entire data, being in a digital world, is available in the form of soft copies of documents. With this, we can update, store, backup and preserve the soft copies of our documents. This is the case with the latest data, but going towards our old traditional data, which is available only on hard copies of the paper, we come across a lot of problems while preserving such rad copies of data. Many a times the old and ancient traditional documents play a vital role in our day to day life. Most of the papers containing our data get degraded due to lack of attention and improper handling and preservation. Most commonly seen degradation of such papers is interference of the text written on the front and back of the papers. In order to make this interfered front end data separate from rear page data many researchers have been proposed binarized documentation methodology. Here we study and analyze various binarization techniques proposed previously and then propose the new and innovative technique for the same. We create the binarized image of the degraded image through some intermediate steps. Ultimately, the binarized image will be next processed by the post processing module. The final output of entire process will generate a clear and binarized image with foreground text clearly seen without interference.

KEYWORDS*:* Binarization, DIBCO, Ground-truth image, OCR- optical character recognition, PNSR.

I. INTRODUCTION

In this digital world, various image and document processing techniques emerged in a wider scope for data extraction or text extraction. The images are widely used in various domains of the researches such as geography, tomography, etc. Most of the novels written few of the years ago on the papers are of utmost use in our day to day life, but due to improper maintenance of such novels, the data is degraded and becomes unreadable for users and thus leads to loss of useful data. Such images becomes degraded after a particular span of time, and we can't use them in spite of them being very useful for us. Sometimes some documents get degraded due to low quality papers or inks used to type or write on the papers, thus making such useful image of no use for further use.

The degraded document images either scanned or captured are in the form unreadable text in foreground format. We need to differentiate between the foreground and background text. The techniques for image binarization are therefore emerged as useful ways for obtaining text from degraded documents. The degraded images are then passed through various intermediate methods which will produce the output image in a foreground text readable format. This survey will first analyze various techniques and then make compare the existing techniques to the proposed one. Although document or image binarization issue still prevails, threshold consideration of degraded and interfered document images have been resolved.Fig.1 shows how historical documents suffer from different problems like ink seeping through front side and smudge of ink on the document

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Fig. 1.Sample degraded document images taken from DIBCO dataset

II. EXISTING SYSTEM

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:

$$T = m + k \times s \qquad (1)$$

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

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

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:

$$C(i,j) = Imax(i,j) - Imin(i,j) \quad (2)$$

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).

$$C(i,j) = \frac{Imax(i,j) - Imin(i,j)}{Imax(i,j) + Imin(i,j) + \epsilon} \quad (3)$$

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. PROPOSED SYSTEM

Here the proposed system presenting an document image binarization technique that is based on adaptive image contrast which is tolerant to different types of degradation of documents such as uneven illumination and document smear. This is the simple and robust technique that involves only few parameters. It also works for different kinds of degraded document images. The local image contrast is used in this technique that is evaluated based on the local maximum and minimum values.

A. Module of Contrast Image

In this module, Local image gradient and local image contrast are combined to get the adaptive image contrast so that the adaptive combination of the local image contrast and the gradient can produce proper contrast maps for document images.

B. Module to Find the Edges

In this module contrasted image is matched with gray scale edge detection graph. It produces the outline of the pixels around the foreground text. These pixels are divided into two groups : connected pixels and non-connected pixels.The area around text stroke is occupied by connected pixel. And the other noisy area present in the image is occupied by non-connected pixel. From the contrast image constructionwe can get the stroke edge pixels of the document text properly.Constructed contrast image consist of a clear bi-modal pattern.The difference between the maximum and minimum intensity in a local window is used to evaluate the local image gradient. Pixels that are present

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

at both the sides of the text stroke are selected as the high contrast pixels. After that the binary map is constructed. The pixels that are present in both the high contrast image pixel map and gray scale method, only those areplacedin the combined map. Accurate extraction of the text stroke edge pixels is done with the help of this combination.

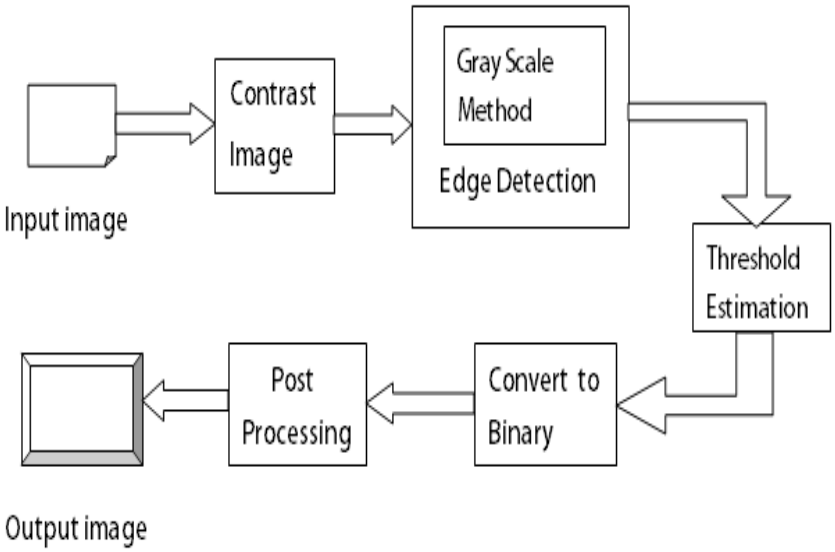

Fig.3. Block diagram of proposed system

C. Local Threshold Estimation

As soon as the detection of high contrast stroke edge pixels is done properly, the extraction of the text from the document background is carried out. From the different kinds of document images two characteristics can be noticed. First one is the text pixels are close to the detected text stroke edge pixels and second is the divergent intensity difference between the surrounding background pixels and the high contrast stroke edge pixels. Redrawing of the actual text can be done using edge detection method. In this module the mean value is calculated.

D. Module to Convert into Binary

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 indicate that image pixels are connected pixels and the represents the text strokes. The 0's are removed from the image because they are part of background image. From the contrast image construction we can get the stroke edge pixels of the document text properly. Constructed contrast image consist of a clear bi-modal pattern.The difference between the maximum and minimum intensity in a local window is used to evaluate the local image gradient.Pixels that are present at both the sides of the text stroke are selected as the high contrast pixels. After that the binary map is constructed. The pixels that are present in both the high contrast image pixel map and gray scale method, only those areplacedin the combined map. Accurate extraction of the text stroke edge pixels is done with the help of this combination.

E. Post Processing Module

Separation in the image is created by binarization method. So post processing is done to eliminate the non-strokes image from binary image. And it gives a clear image that consists of only text strokes. The comparison of output image with input image shows the significance of our system. Output image consists of clean and readable text. The binarization result from binarization method can further improved usingpost processing procedure algorithm. Itrequires:

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Input Document Image I, Binarization Result B and Corresponding Binary Text Stroke Edge Image Edge.First, the pixels that do not connect with other foreground pixels i.e. isolated foreground pixels are filtered out to precisely make the edge pixel set. 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). A single pixel out of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. At last, certain numbers of single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators.

IV. EXPECTED OUTPUT

This paper introduces novel document image binarization technique which will helps to recover the badly degraded documents which are not exactly bimodal. The expected output as per the base paper is as follow:

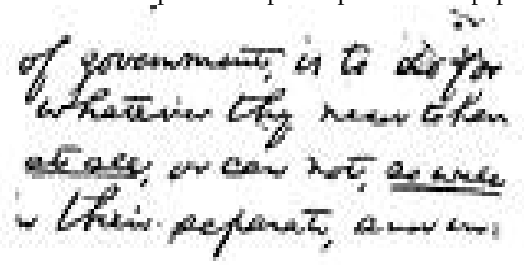

Fig. 4. Expected output image for Fig. 2(a)

V. CONCLUSION

We introduced new post processing module which will remove the background degradations found in the binarized image. This technique uses contrast enhancement along with threshold estimation. This will help into create more clear and readable output. For that we have maintain the contrast level at minimum and maximum level. The output of this system produces separated foreground text from collided background degradation. It can live work on many degraded images.The proposed method is simple binarization method, which produces more clear output. In this technique we are going into used grey scale method through create outlined map around the text.

REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "*ICDAR 2009 document image binarization contest (DIBCO 2009)*,", Jul. 2009.
[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "*H-DIBCO 2010 handwritten document image binarization competition*," , Nov. 2010.
[3] S. Lu, B. Su, and C. L. Tan, "*Document image binarization using background estimation and stroke edges*,", Dec. 2010.
[4]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*," , 2003.
[5] M. Sezgin and B. Sankur, "*Survey over image thresholding techniques and quantitative performance evaluation*," , Jan. 2004.
[6]W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 19806.
[7] 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.
[8] 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.
[9]O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 12, pp. 1191–1201, Dec. 1995.
[10] 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.