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

Efficient RDIB Technique for Degraded Document Images

Namrata Pawar[1], Komal Walke[2], Seema Satpute[3], Pallavi Jagtap[4], Prof. Ajita Mahapadi[5],

UG Students, Department of Information Technology, Dhole Patil College of Engineering,

Pune, Maharashtra, India[1,2,3,4]

Assistant Professor, Department of Information Technology, Dhole Patil College of Engineering, Pune, India[5]

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: Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.

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. Its because of the the high inter/intra-variation between the foreground text stroke and the unnecessary document background across different documents and imag,es. .

The over-normalization problem of the local maximum minimum algorithm [5]. At the same time, the parameters used in the algorithm can be adaptively estimated

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

Fig1. Sample Degraded Images taken from DIBCO Series of images

II. LITERATURE SURVEY

Many thresholding techniques have been reported for document image binarization. As many degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. For example, the early window-based adaptive thresholding techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. The main drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Other approaches have also been reported, including background subtraction texture analysis, recursive method decomposition method, contour completion, Markov Random Field, matched wavelet cross section sequence graph analysis, self-learning, Laplacian energy user assistance and combination of binarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The local image contrast and the local image gradient areVery useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques [5], [6], [7], [8]. In [14], the local 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). Bernsen'smethod is simple, but cannot work properly on degraded document images with a complex document background. We have earlier proposed a novel document image binarization method [5] by using the local image contrast that is evaluated as follows:

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

Where is a positive but infinitely small number that is added in case the local maximum is equal to 0. Compared with Bernsen's contrast in Equation 1, the local image contrast in Equation 2 introduces a normalization factor (the denominator) to compensate the image variation within the document background. Take the text within shaded document areas such as that in the sample document image in Fig. 1(b) as an example. The small image contrast around the text stroke edges in Equation 1 (resulting from the shading) will be compensated by a small normalization factor (due to the dark document background) as defined in Equation 2.

III. PROPOSED SYSTEM

In This section describes the proposed 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.

A. CONTRAST IMAGE CONSTRUCTION:

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many on stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background. In our earlier method [5], the local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient The denominator is a normalization factor that suppresses the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator and accordingly result in a relatively high image contrast. However, the image contrast in Equation 2 has one typical limitation that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small. To overcome this over-normalization problem, we combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

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

where $C(i, j)$ denotes the local 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 windows size is set to 3 empirically. α is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large α) when the document image has significant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. Where Std denotes the document image intensity standard deviation, and γ is a pre-defined parameter. The power function has a nice property in that it monotonically and smoothly increases from 0 to 1 and its shape can be easily controlled by different γ . γ can be selected from $[0,\infty]$, where the power function becomes a linear function when $\gamma = 1$. Therefore, 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. The setting of parameter γ will be discussed in Section IV. Fig. 2 shows the contrast map of the sample document images in Fig. 1 (b) and (d) that are created by using local image gradient, local image contrast [5] and our proposed method in Equation 3, respectively.For the sample document with a complex document background in Fig. 1(b), the use of the local image contras Produces a better result as shown in Fig. 2(b) compared with

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. 2. Contrast Images constructed using (a) local image gradient

variation within the text strokes, the use of the local image contrast removes many light text strokes improperly in the contrast map as shown in Fig. 2(b) whereas the use of local image gradient is capable of preserving those light text strokes as shown in Fig. 2(a). As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3 can produce proper contrast maps for document images with different types of degradation as shown in Fig. 2(c). In particular, the local image contrast in Equation 3 gets a high weight for the document image in Fig. 1(a) with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image in Fig. 1(b). The result by the local image gradient as shown in Fig. 2(a) (Because the normalization factors in Equation 2 helps to Suppress the noise at the upper left area of Fig. 2(a)). But For the sample document in Fig. 1(d) that has small intensity Variation within the document background but large intensity.

B. TEXT STROKE EDGE PIXEL DETECTION:

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bi-modal pattern [5], where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background. We therefore detect the text stroke edge pixel candidate by using Otsu's global thresholding method. For the contrast images in Fig. 2(c), Fig. 3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly.As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be further improved through the combination with the edges by Canny's edge detector, because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3(b) without tuning the parameter manually.In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Fig. 3(c).

C. POST-PROCESSING:

Once the initial binarization result is derived from Equation 5 as described in previous subsections, the binarization result can be further improved by incorporating certain domain Knowledge 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 as described in [4].

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

IV. EXPERIMENTAL RESULT

We have used our system on various on various type of images, like novel, books, and records, historical literature. Some of the results are stated as follow:

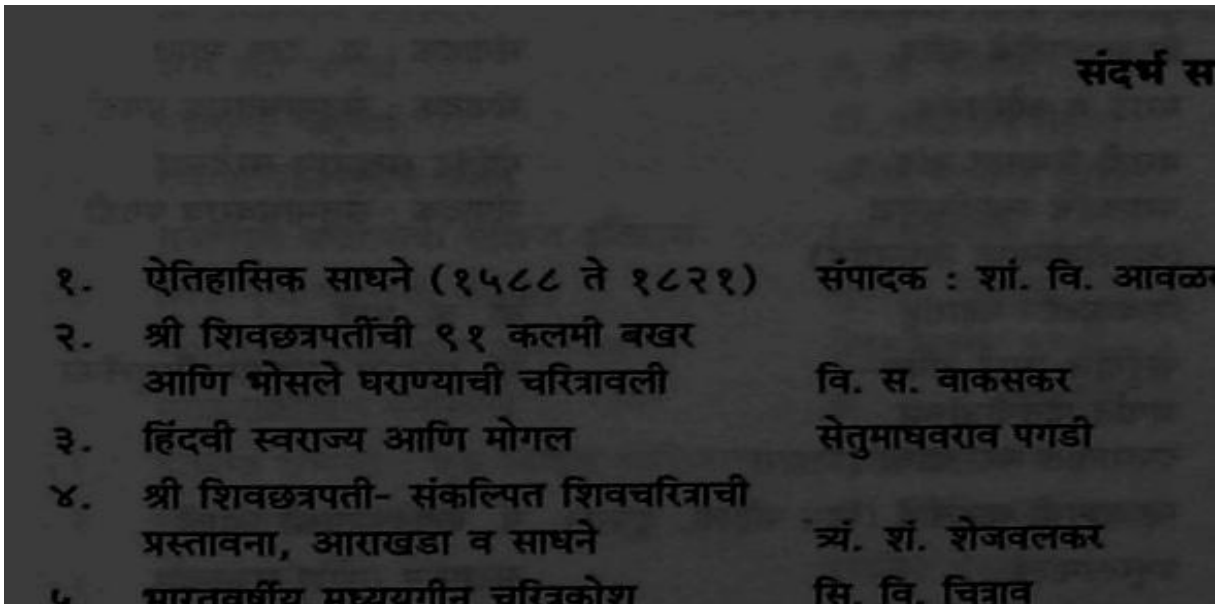

Fig. 3: Input image for proposed system

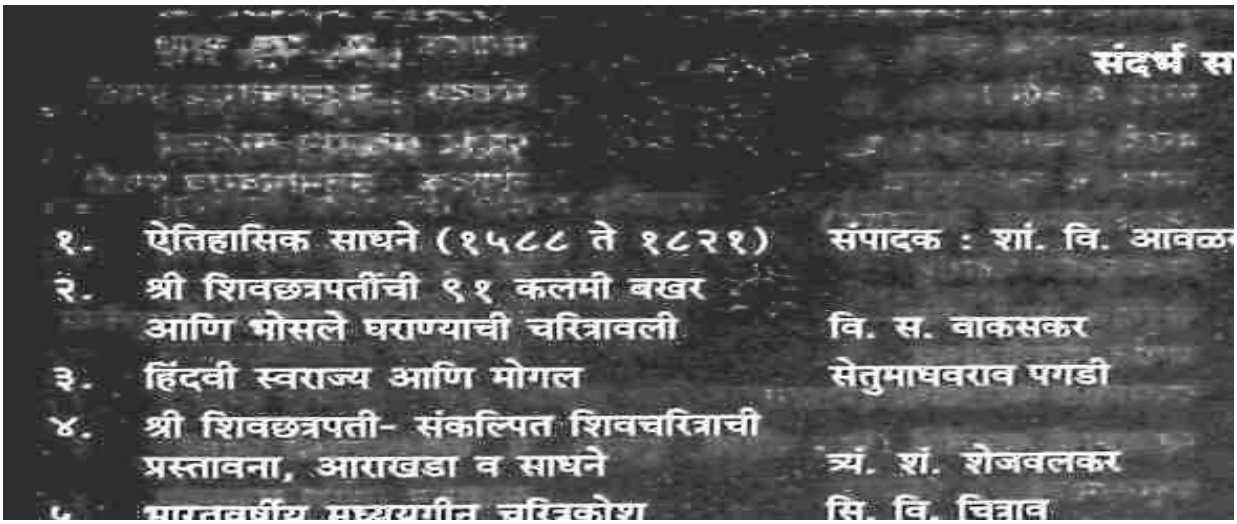

Fig. 4 Final Output Of Project

V. CONCLUSION

This paper 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

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

images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, NRM, MPM and DRD.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. J. Bernsen, "Dynamic thresholding of gray-level images," in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.
7. J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.
8. W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1986.