Binarization Technique on Historical Documents using Edge Width Detection

Vrushali S. Nikam.
Department of Computer Engineering
MET-BKC, Savitribai Phule Pune University
Nashik, Maharashtra
E-mail: Vrushali.nik@gmail.com

Prashant M.Yawalkar.
Department of Computer Engineering
MET-BKC, Savitribai Phule Pune University
Nashik, Maharashtra
E-mail: Prashant25yawalkar@gmail.com

***Abstract*-**Document images often suffer from different types of degradation that renders the document image binarization is a challenging task. Document image binarization is of great significance in the document image analysis and recognition process because it affects additional steps of the recognition development. The Comparison of image gradient and image contrast is estimated by the local maximum and minimum which has high quality. So, it is more tolerant to the rough lighting and other types of document degradation such as low contrast images and partially visible images. The distinction between the foreground text and the background text of different document images is a difficult task. This paper presents a new document image binarization technique that focus on these issues using adaptive image contrast. The grouping of the local image contrast and the local image slope is the adaptive image contrast so as to tolerate the text and surroundings distinction caused by dissimilar types of text degradations. In image binarization technique, the construction of adaptive contrast map is done for an input degraded document image which is then adaptively binarized and combined with Canny's edge detector to recognize the text stroke edge pixels. The document text is advance segmented by means of a local threshold. We try to apply the self-training adaptive binarization approach on existing binarization methods, which improves not only the performance of existing binarization methods, but also the, toughness on different kinds of degraded document images.

Keywords *- Binarization, thresholding, image contrast, image gradient.*

*****

I. INTRODUCTION

The conversion of a scanned grey scale image into a binary image is called binarization which not only selects textual components but also non-textual components. Original documents are often unclean due to smearing and smudging of text and aging. On the other hand document image binarization has been deliberated for many years; the thresholding of chronological document images is at a standstill an uncertain trouble due to the elevated variation within both the document background and foreground as illustrated in fig.1.The most recent Document Image Binarization Contest (DIBCO) [1] held in the construction of the International Conference on Document Analysis and Recognition (ICDAR) 2009 also shows current firm work on this issue. In this paper, we focus on documents where the foreground mainly encompasses handwritten text. There is seeping of ink from the other side of the page and general degradation of the paper and ink. We cannot differentiate between foreground and background text.

Image binarization [2] is the process of partition of pixel values into double collections, black as foreground and white as background. Thresholding has produced to be a well-known technique used for binarization of document images. Thresholding is further divide into the global and local thresholding [2] technique. In document with consistent contrast delivery of background and foreground, global thresholding is has found to best technique. In degraded documents, where extensive background noise or difference in contrast and intensity exists i.e. there exists many pixels that cannot be fluently categorized as foreground or background. In such cases, local thresholding has significant over other techniques. The main objective of this chapter is to evaluate the different image binarization techniques [5] to find the gaps in existing techniques.

The global thresholding technique computes most advantageous threshold for the entire image; these techniques require few computations and can work well in straight- forward cases but fails in complex backgrounds, such as non-uniform color and disadvantaged illuminated backgrounds. These methods are usually not suitable for degraded document images, because they do not have a clear pattern that separates foreground text and background. The local binarization techniques set dissimilar thresholds for different target pixels depending on their neighborhood/local information. In general, these techniques are susceptible to background noises due to large difference in case of a poor illuminated document or low contrast degradation.

The problems commonly seen in the degraded document images are poor contrast, non-uniform image background intensity, immoderate amount of and noises. For a given document image, different binarization methods may create

 ISSN: 2321-8169
3793 - 3798

different corresponding binary image. Some binarization methods perform superior on certain kinds of document image, while others create better results for other kinds of document images. By combining different binarization techniques, better performance can be achieved with carefully analysis. A binary image [3] is a digital image that has just two practicable values meant for every pixel. Usually, two colors are used for a binary image i.e. black and white on the other hand any two colors can be used. The color used for the objects in the image is the foreground color while the rest of the image is the background color. Binary images [4] commonly occur in image processing as masks or as the result of some operations as segmentation and thresholding. A small number of input/output devices, for example, laser printers, bi-level computer displays, are able to straight away bi-level images. Binary images are formed from color images by segmentation.

All the reported thresholding methods [8] have been established to be effective in constrained preprocessing environments with conventional images. None has demonstrated effectual in all cases of universal document image processing. In this paper we examine some preprocessing algorithms and evaluate their effectiveness on 'difficult' document image processing troubles. The objective is to identify specify algorithms or combinations of algorithms which can be functional to any category of document image to construct effective thresholding.

Fig. 1. Historical and degraded document image

II. LITERATURE SURVEY

Earlier, many attempts had been taken to look up the image of such documents by using different kinds of methods of binarization involving use of moreover global or local threshold. These methods do acquire an initiative towards civilizing the visibility of text on the degraded document, but not enough efficiently to be easily recognizable by the OCR. Though document image binarization [1] has been considered for many times, the thresholding of corrupted text images is at a standstill an uncertain problem owing to the high inter/intra-variation stuck between the text stroke and the document background of the diverse document images. The handwritten text within the corrupted documents often shows a convinced amount of variation in terms of the stroke intensity, stroke association, stroke width, and document background. These different types of document degradations [2] tend to produce the document thresholding error and make degraded document image binarization a huge challenge to most state of-the-art techniques.

Document image binarization [3] plays a key role in document processing since its performance affects quite seriously the degree of success in subsequent character segmentation and recognition. In general, approaches that deal with document image binarization are categorized in two main classes: (i) global and (ii) local. The Existing system gives local contrast [1] by Bernsen's method follows:

$$C(i,j) = I_{max}(i,j) - I_{min}(i,j \quad (1)$$

Where C (i, j) denotes the contrast of an image pixel (i, j) and $fmax(i,j)$ and $fmin(i,j)$ are the maximum and minimum intensities within a local neighborhood windows of (i,j) respectively. Bernsen's method is simple, but cannot effort appropriately on degraded document images with complex document background. The second local contrast is given by:

$$c(i,j) = \frac{fmax(i,j) - fmin(i,j)}{fmax(i,j) + fmin(i,j) + \varepsilon} \quad (2)$$

Where Ɛ is a positive and equal to 0. The local image contrast in Equation 2 introduces a normalization factor. However, the image contrast in Equation 2 has one typical restriction that it may not control document images with the bright text properly. This is because a fragile contrast will be considered for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small. To conquer this over-normalization problem, the local image contrast is combined with the local image gradient and derives an adaptive local image contrast [1].

Shapiro et al. [10] introduce a global thresholding method, where the independency is stressed in the object/background areas ratio, intensity transition slope, object/background shape and noise-insensitivity.

Pavlidis et al. [11] presents a technique based on the observation that after blurring a bi-level image, the intensity of original pixels.

 ISSN: 2321-8169
3793 - 3798

Yang et al. [12] propose an algorithm to perform thresholding for scenes containing distinct objects. The sequence of graphs is constructed using the size of connected objects in pixels as a classifier.

The enhanced speed entropic threshold selection algorithm is proposed in Ref. [13] by Chen et al. They reduce the image grey-scale levels by quantization and produce a global threshold candidate vector.

Yanowitz and Bruckstein et al. [14] proposed an image segmentation algorithm based on adaptive binarization, where different image quality problems were taken into consideration.

Bolan Su et al. (2012) [9] have proposed that document image binarization is a important pre-processing technique for document image analysis that segments the text from the document image backgrounds. They have recommended a learning framework that makes utilize of the Markov Random Field to move forward the performance of the existing document image binarization methods for those degraded document images. Wide-ranging experiments on the recent Document Image Binarization Contest datasets express that significant enhancement of the accessible binarization methods when applying our proposed framework.

III . PROPOSED SYSTEM

A various degraded document image, the construction of an adaptive contrast map is done and the text stroke edges are then perceived through the combination of the binarized adaptive contrast map and the canny edge map and local & global thresholding is apply to whole image It generally gives combination of local image gradient and local image contrast . The text is subsequently segmented based on the local threshold that is predictable from the detected text stroke edge pixels. Finally, some post-processing is further applied to recover the document binarization excellence.

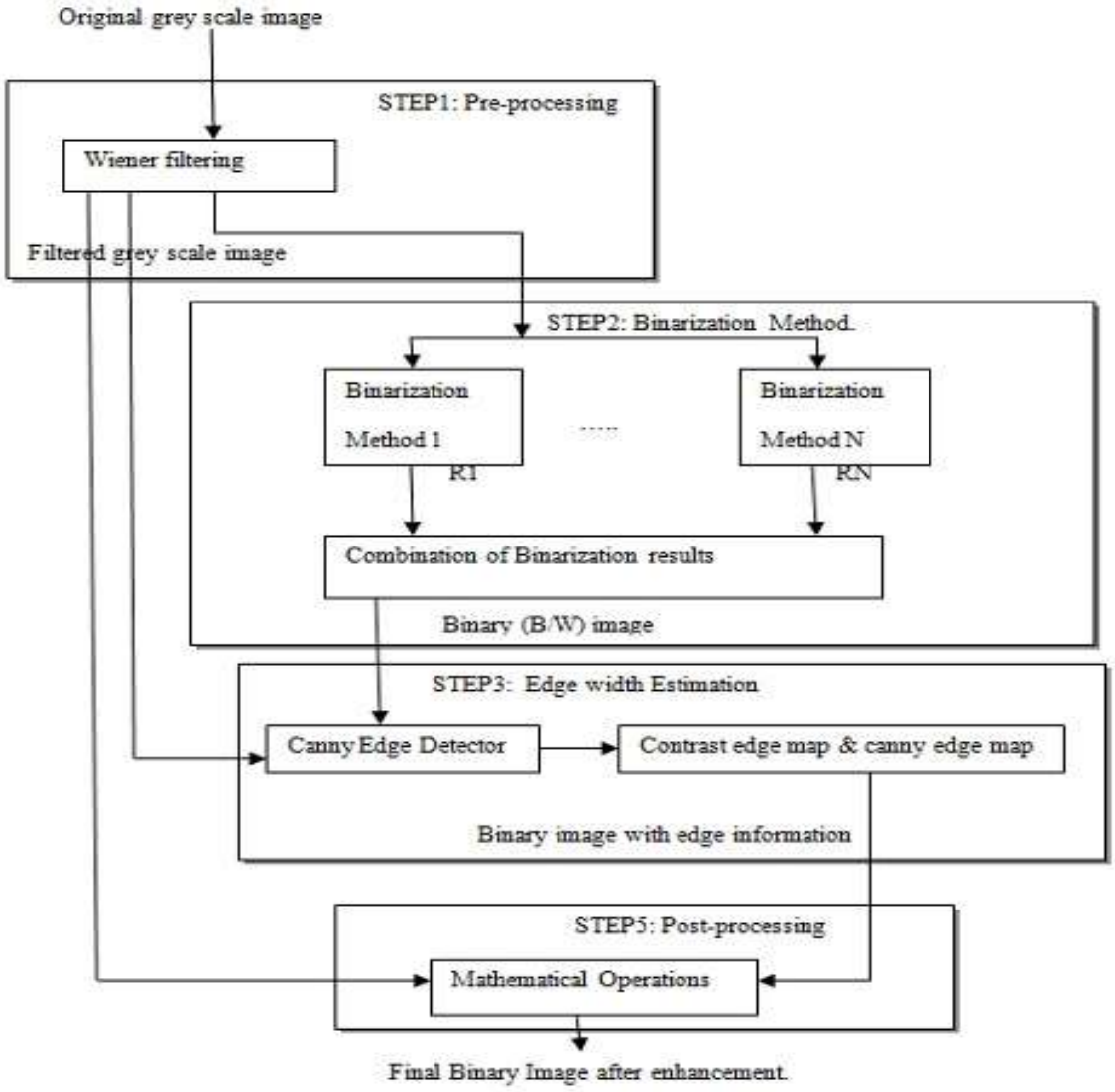

Fig.2: System Architecture

 ISSN: 2321-8169
3793 - 3798

A. Contrast Map Generation(Preprocessing):

It gives construction of contrast map from gray scale image. In the proposed method the degraded input image is taken and it is converted into gray scale image. Then contrast map is generated. Then the local contrast C (i, j) is calculated using the Equation 3.Wiener filtering is also used to give good localization property.

$$Ca(i,j) = \alpha Ca(i,j) + (1-\alpha)(fmax(i,j) - fmin(i,j)) \quad (3)$$

Where C (i, j) is local contrast and $(fmax(i,j) - fmini,j$ refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 0r less than 3 and α is the weight between local contrast and local gradient that is restricted based on the document image numerical information. A is given by:

$$\alpha = (\frac{Std}{128})^{\gamma} \quad (4)$$

B. Combination Of Binarization:

A pre-processing stage of the grey scale source image is essential for historical and degraded documents for the elimination of noisy areas, smoothing of background texture as well as contrast enhancement between background and text areas. The use of Wiener filtering has been proved efficient for the aforementioned goals. The Wiener filter is commonly used for image restoration. At this step, we calculate a binary image which combines the N binarization results. Since we aim to mark as foreground pixels only those pixels that the majority of the binarization methodologies classify as foreground text.

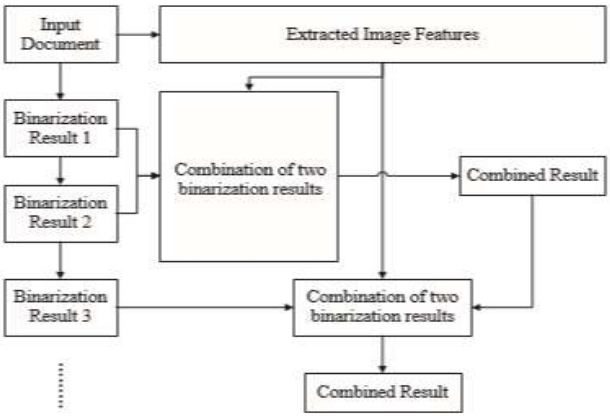

Fig.3: Binarization Method

C. Edge Width Evaluation

The edge evaluation is given by:

- First the edge image is scanned horizontally row by row and edge pixels candidates are selected.
- If the edge pixels, which are labeled 0(background) and pixels next to them are labeled to 1(edge) in edge map are correctly detected.
- Remove those pixels that have lower intensity which is selected in step 2.
- Finally matching of adjacent distance in same row into the pairs and calculate distance between two pixels.
- Histogram is constructed and most frequently occurring distance as estimated stroke edge width EW.

At the same time as the local image contrast and the local image gradient are evaluated by the differentiation between the maximum and minimum intensity in a local window(W), 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 [1], because Canny's edge detector has good localization possessions that it can spot the edges close to real edge locations in the detecting image.

D. Post-processing:

The final improvement in binary image is given by post-processing procedure:

- The isolated foreground pixels that do not connect with other foreground pixels are filtered out.
- The neighborhood pixel that lies on symmetric sides of a text stroke edge pixel should belong to different classes.
- If pixels belong to same class then assign them with lower intensity to foreground class (text) and other to background class.
- Finally, some single pixel artifacts along the text stroke boundaries are filtered out using several logical operators.

IV. Result Analysis and Experiments

In this experiment, we compared proposed system with the other techniques on DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. The three datasets are self-possessed of the equivalent series of document images that undergo from numerous common document degradations such as spread smudge, bleed-through and low down difference. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. The F-measure and PSNR values are evaluated. The values are compared with other methods and shown in table. There exist several experimental measures to evaluate and compare the binarization methods.

F-measure: F-measure (FM) is the harmonic mean of precision and recall. F-measure is calculated at the pixel level.

ISSN: 2321-8169
3793 - 3798

$$F - measure = \frac{2*recall*precision}{recall+precision} \quad (5)$$

Where recall=TP/TP+FN, precision= TP/TP+FP and TP, FP, FN denote the True positive, False positive and False Negative values, respectively. It consists of total of 26 images from the datasets given above. Our datasets consist of four degraded handwritten documents and five degraded printed documents from DIBCO 2009 dataset, nine degraded handwritten documents from H-DIBCO 2010 dataset and eight degraded handwritten documents from DIBCO 2011 dataset. There are two parameters that need to be set properly and the size of the neighborhood window and the edge width of the high contrast image pixels. The size of the local neighborhood window based on the double edge construction of the character strokes is determined.

Commonly, a more local window size will help to diminish the classification error that is habitually induced by the deficient in of edge pixels surrounded by the local neighborhood window. In addition, the performance of the proposed method becomes steady when the local window size is larger than 2EW constantly on the three datasets.

The PSNR (Peak Signal to Noise Ratio) is defined as follows:

$$PSNR = 10log\frac{C^2}{MSE} \quad (6)$$

Where MSE denotes the mean square error and C is a constant and can be set at 1. This metric measures how close the result image to the ground truth image.

V.EVALUATION RESULT OF DIBCO DATASET

As shown Graph, the proposed method produces improved results on DIBCO dataset when the value is much smaller than 1 and the local image contrast dominates. On the other hand, the F-Measure performance of H-DIBCO dataset improves significantly when increases to 1. Consequently the proposed method can assign more suitable to different images when is nearer to 1. The value of is set be approximately 1 when the adaptableness of the proposed technique is maximized and enhanced and extra robust binarization results can be consequential from diverse kinds of degraded document images. Another parameter, i.e., the local window size W, is experienced in the second experiment on the DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. W is intimately related to the stroke width EW. Graph shows the thresholding outcomes when W varies as of EW to 4EW.

In this experiment, we compared proposed system with the other techniques on DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. The three datasets are self-possessed of the equivalent series of document images that undergo from numerous common document degradations such as spread smudge, bleed-through and low down difference. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. The F-measure and PSNR values are evaluated. The values are compared with other methods and shown in table 1.
The table 2 shows all combined images from DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. The proposed framework can produce better results than other methods, which are the best performance method in the DIBCO 2009 contest. And the combined results can perform better in terms of F-Measure PSNR than the two origin methods separately. This means a higher precision and better text stroke contour can be obtained after combination of binarized image.

EW(Estimated Width)	DIBCO2009 (series1)	HDIBCO2010 (series2)	DIBCO2011 (series3)
1.5EW	0.84	0.66	0.91
2EW	0.66	0.74	0.8
2.5EW	0.72	0.91	0.92
3EW	0.71	0.79	0.91
3.5EW	0.66	0.78	0.91

Table1: Values of F-Measure

Table2: Values of PSNR

DIBCO 2009(Series 1)	PSNR values
HW01	70
HW02	64
HW03	63
HW04	70
HW05	65

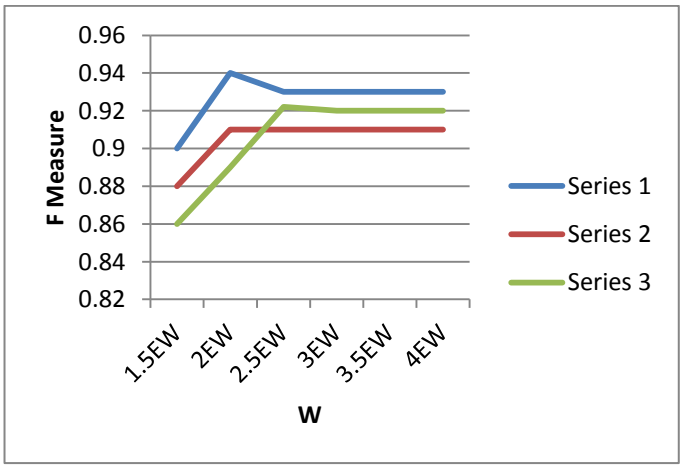

Fig.4: Graphical representation of F-measure using DIBCO

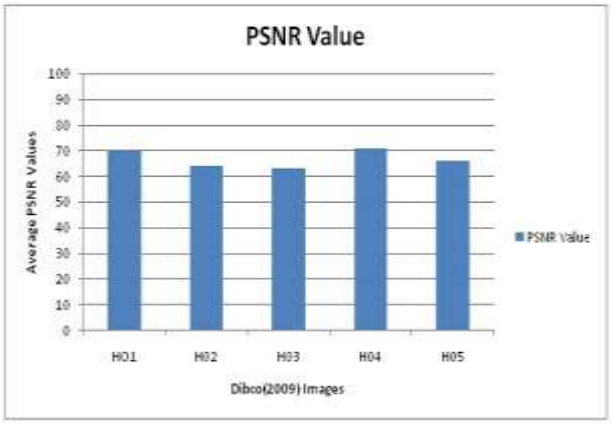

Fig.4: Graphical representation of PSNR using DIBCO

Conclusion

It gives document image binarization technique which is based on an adaptive image contrast. The binarization technique is efficient as it consists of less parameter. It is tolerant to different types of degraded images. It makes use of the local image contrast that is evaluated by local maximum and minimum. It also makes use of the document background surface and the text stroke edge information. The proposed method has been tested on the various datasets. This method aims at humanizing the textual image by removing noise and removes the unwanted gaps within the text to be filled and hence provide a complete understandable image which is detected easily by the OCR. The F-measure and PSNR values give considerable improvement in the proposed method than other techniques. The time require for this method is more than other techniques but the quality of result is considerable.

Acknowledgment

The support from the MET COE institute of engineering and research center is gratefully acknowledged. We also thank Dr. M.U. Kharat and Prof. P.M. Yawalkar for their contributions

References

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, *Senior Member, IEEE*, "Robust image Binarization technique for degraded document images," *Int. J.Document Anal. Recognit.* VOL. 22, NO. 4, APRIL 2013)

[2] 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.

[3] 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.*

[4] 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.*

[5] 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.*

[6] 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.

[7] J.-D. Yang, Y.-S. Chen and W.-H. Hsu. Adaptive thresholding algorithm and its hardware implementation. Pattern Recognition Letters, 15(2):141–150, 1994.

[8] B. Gatos, I. Pratikakis, and S. Perantonis. Adaptive degraded document image binarization. Pattern Recognition, 39(3):317–327, 2006.

[9] Jeena Joy, Jincy Kuricokose ,Adaptive Contrast Binarization using Standard deviation for degraede document image , 39(3):317–327, 2006.

[10] V.A. Shapiro, P.K. Veleva, V.S. Sgurev, An adaptive method for image thresholding, Proceedings of the 11th KPR, 1992, pp. 696 699.

[11] J. Yang, Y. Chen, W. Hsu, Adaptive thresholding algorithm and its hardware implementation, Pattern Recognition Lett. 15 (2) (1994) 141}150.

[12] T. Pavlidis, Threshold selection using second derivatives of the gray scale image, International Conference on Document Analysis and Recognition, ICDAR '93, Japan, 1993, pp. 274}277.

[13] W. Chen, C. Wen, C. Yang, A fast two-dimensional entropic thresholding algorithm, Pattern Recognition 27 (7) (1994) 885}893.

[14] S.D. Yanowitz, A.M. Bruckstein, A new method for image segmentation, CVGIP 46 (1989) 82}95.