

Image Segmentation and Binarization Technique for Manuscript

Mayur Sonar

Department of MCA, K K Wagh Institute of Engineering, Education & Research, Nashik, Maharashtra

Abstract: To do segmentation from badly degraded document images is very tough and challenging tasks. It is due to the high inter/intravariation between the document background and the foreground text of different document images. we propose document image binarization technique that focuses on these issues by using adaptive image contrast. It Combine local image contrast and the local image gradient for construct adaptive contrast map that is tolerant to text and background variation caused by different types of document degradations. In the proposed technique, we first constructed adaptive contrast map for an input degraded document image. And then image segmentation algorithm is used 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. This system was tested on three public datasets that were used in the recent. Those datasets are Document Image Binarization Contest (DIBCO) 2009 & 2011 and Handwritten Document Image Binarization Contest (H-DIBCO) 2010 and thus come up with an accuracies of 93.5%, 87.8% and 92.03%, respectively that are significantly higher than or close to that of the best-performing methods reported in the three contests.

Keywords: Adaptive Image Contrast, Document Analysis, Document Image Processing, Degraded Document Image Binarization, Pixel Classification.

1. INTRODUCTION

Image Binarization is a common first step to document image analysis, converts the gray values of document images into two level representations for text and non-stroke regions. The handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. Historical documents are often degraded by the bleed-through where the ink of the other side seeps through to the front.

For a given document image, different binarization methods may create 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. Document binarization is technique for removing noise from document background and extracts the foreground text. Using Document image Binarization technique, improves degraded document which contains uneven lighting bleed. This method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations.

2. LITERATURE REVIEW

There are many thresholding techniques developed for document image binarization. Many degraded documents

do not have a clear bimodal pattern, so global thresholding is not a suitable approach for the degraded document binarization. Therefore adaptive thresholding which estimates a local threshold for each document image pixel is better approach to deal with degraded document images. For example, the early window-based adaptive thresholding techniques uses mean and the standard deviation to estimate the local threshold of image pixels which is better approach to deal with different variations within degraded document images.

The main drawback of these window-based thresholding techniques is that the thresholding performance totally depends on the window size and the character stroke width. Other approaches for window-based thresholding are: 1. Background subtraction [6],[7] this approach presents a document binarization technique that makes use of the document background surface and the text stroke edge information. 2. Texture analysis [8], a locally adaptive approach for the binarization and enhancement of degraded documents. 3. Recursive method [9], [10], 4.

Decomposition method [11], this approach is a local adaptive analysis method, which uses local feature vectors to find the best approach for thresholding a local area. 5. Contour completion [12]–[13], 6. Markov Random Field [14][15], a new approach to the binarization of document images based on the energy minimization. For energy minimization, formulate the energy using MAP-MRF approach and perform the optimization via graph cut. 7. Matched wavelet [16], a technique for locating the text part based on textural attributes using GMWs. 8. Cross Section Sequence Graph Analysis [17], ICSSG, an

algorithm for handwritten character segmentation that tracks the characters' growth at equally spaced thresholds. The iterative thresholding reduces the effect of information loss associated with image binarization. 9. Self-learning [18], 10. Laplacian energy [19], 11. User assistance [20][21], a comprehensive approach for converting scanned documents to black and white. 12. Combination of binarization techniques [22], [23]. These methods combine different types of image information and domain knowledge and are often complex.

3. METHODOLOGY

To demonstrate the system given degraded document images are taken from dataset (Document Image Binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010).

To get the resultant binary images we have to apply following methodology:

1. Contrast Image Construction:

To extract only stroke edges from the degraded document images, the image gradient needs to be normalized to compensate the image variation within the document background. In this method, we combine local image contrast and local image gradient to construct adaptive local image contrast as follow:

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

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. α 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:

$$\alpha = (Std/128)^\gamma$$

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.

2. Image Segmentation Algorithm

In this algorithm, each pixel in an image has its own threshold, which is estimated by calculating the mean of the grayscale values of its neighbor pixels, and the square variance of the grayscale values of the neighbor pixels are also calculated as an additional judge condition, so that the result of the proposed algorithm is the edge of the image. Results of this algorithm show that it is apparent to obtain better results by the proposed algorithm than by Canny operator. The proposed algorithm has an obvious advantage in noise restraining, which is a good edge detecting and image segmentation algorithm with wide applicability.

3. Edge Width Estimation Algorithm:

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 :

- It will detect text pixels which are close to the text stroke edge pixels.
- 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 Algorithm.

Algorithm: Edge Width Estimation

Require: The Input Document Image I and Corresponding Binary Text Stroke Edge Image Edg

Ensure: The Estimated Text Stroke Edge Width EW

- 1: Get the width and height of I
- 2: for Each Row $i = 1$ to height in Edg do
- 3: Scan from left to right to find edge pixels that meet the following criteria:
 - Its label is 0(background)
 - The next pixel is labeled as 1(edge).
- 4: Examine the intensities in I of those pixels selected in Step 3, and remove those pixels that have a lower intensity than the following pixel next to it in the same row of I .
- 5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
- 6: end for
- 7: Construct a histogram of those calculated distances.
- 8: Use the most frequently occurring distance as the estimated stroke edge width EW .

Since we do not need a precise stroke width, we just calculate the most frequently distance between two adjacent edge pixels in horizontal direction and use it as the estimated stroke width.

4. Post Processing Algorithm

Once we get the initial binarization result from local threshold estimation, the previous result can be further improved by Post-Processing Procedure algorithm. First, the filtered out foreground pixels those are not connect with other foreground pixels to make the edge pixel set precisely. Second, if the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel then they belong to different classes. If both of two pixels belong to the same classes then the one pixel of the pixel pair is labeled to the other category. Finally, by using several logical operators filtered out single-pixel artifacts along the text stroke boundaries.

Algorithm Post-Processing Procedure

Require: The Input Document Image I , Initial Binary

Result: B and Corresponding Binary Text Stroke Edge Image Edg

Ensure: The Final Binary Result Bf

- 1: Find out all the connect components of the stroke edge pixels in Edg .

- 2: Remove those pixels that do not connect 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 then
- 6: Assign the pixel with lower intensity to foreground and the other to background.
- 7: end if
- 8: end for
- 9: Remove single-pixel artifacts along the text stroke boundaries after the document thresholding
- 10: Store the new binary result to B_f .

4. EXPERIMENTAL RESULTS

The proposed method has been tested over the handwritten images of the dataset that is used in the recent Document Image Binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010.

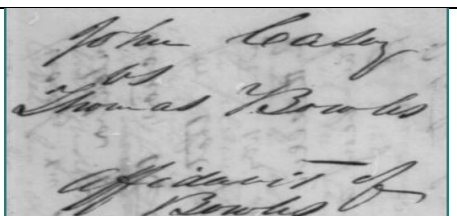


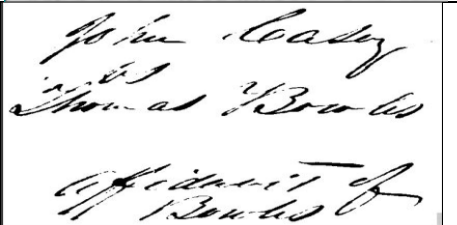
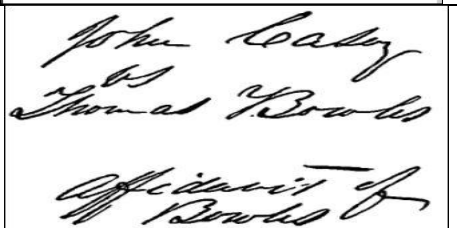
Input Image	
Contrast Image Construction	
Text Stroke Edge pixel Detection	
Edge Width Estimation	
Binarized Result	

TABLE: shows Binarization results of the degraded document image

The DIBCO 2009 dataset contains ten testing images that consist of five degraded handwritten documents and five degraded printed documents. The H-DIBCO 2010 dataset consists of ten degraded handwritten documents. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. In total, we have 36 degraded document images with ground truth.

5. 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 images.

The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. A new Image segmentation algorithm is proposed that each pixel in the image has its own threshold by calculating the statistical information of the grayscale values of its neighborhood pixels. An additional judge condition is given that it is possible to get the edge of the image as the result of the algorithm. The Image Segmentation algorithm also has an obvious advantage in noise restraining. The proposed method has been tested on the various datasets.

REFERENCES

- [1]. S. Zhu, X. Xia, Q. Zhang, K. Belloulata, "An Image Segmentation Algorithm in Image Processing Based on Threshold Segmentation," in Proc. Int. IEEE Conf. on Signal- Image technologies and Internet-Based Sys.
- [2]. 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. yst., Jun. 2010, pp. 159-166
- [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]. 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.
- [7]. B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317-327, 2006.
- [8]. 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.
- [9]. M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," in Proc. IEEE Trans. Image Process. Jun. 1998, pp. 918-921.
- [10]. S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Iterative multi model subimage binarization for handwritten character segmentation," IEEE Trans. Image Process., vol. 13, no. 9, pp. 1223-1230, Sep. 2004.
- [11]. 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
- [12]. Q. Chen, Q. Sun, H. Pheng Ann, and D. Xia, "A double-threshold image binarization method based on edge detector," Pattern Recognit., vol. 41, no. 4, pp. 1254-1267, 2008.

- [13] I. Blayvas, A. Bruckstein, and R. Kimmel, "Efficient computation of adaptive threshold surface for image binarization," *Pattern Recognit.*, vol. 39, no. 1, pp. 89–101, 2006.
- [14] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, "Document image segmentation using a 2D conditional random field model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Sep. 2007, pp. 407–411.
- [15] J. G. Kuk, N. I. Cho, and K. M. Lee, "Map-MRF approach for binarization of degraded document image," in *Proc. Int. Conf. Image Process.*, 2008, pp. 2612–2615.
- [16] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Text extraction and document image segmentation using matched wavelets and MRF model," *IEEE Trans Image Process.*, vol. 16, no. 8, pp. 2117–2128, Aug. 2007.
- [17] A. Dawoud, "Iterative cross section sequence graph for handwritten character segmentation," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2150–2154, Aug. 2007.
- [18] B. Su, S. Lu, and C. L. Tan, "A self-training learning document binarization framework," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3187–3190.
- [19] N. Howe, "A Laplacian energy for document binarization," in *Proc. Int. Conf. Doc Anal. Recognit.*, Sep. 2011, pp. 6–10.
- [20] F. Deng, Z. Wu, Z. Lu, and M. S. Brown, "Binarizationshop: A user-assisted software suit for converting old documents to black-and-white," in *Proc. Annu. Joint Conf. Digit. Libraries*, 2010, pp. 255–258.
- [21] H. Yi, M. S. Brown, and X. Dong, "User-assisted ink-bleed reduction," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2646–2658, Oct. 2010.
- [22] E. Badeskas and N. Papamarkos, "Optimal combination of document binarization techniques using a selforganizing map neural network," *Eng. Appl. Artif. Intell.*, vol. 20, no. 1, pp. 11–24, Feb. 2007.
- [23] B. Gatos, I. Pratikakis, and S. Perantonis, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.