Image binarization techniques for degraded document images: A review

Binarization techniques

[1]Amoli Panchal, [2]Chintan Panchal, [3]Bhargav Shah
[1]Student, [2]Assistant Professor, [3]Assistant Professor
[1]Electronics & Communication,
[1]Sarvajanik Collage of Engineering and Technology, Surat, India

***Abstract*— Improvement in a degraded document images is one of the salient and challenging research now days. Image get degraded due to unbalance illumination spread over document including smearing of text, bleeding of ink to the other side of page, degradation of paper ink due to aging, manuscript characters from background side appear as noise on the lead side and get blend with the lead side characters etc. Binarization is used to recover text from degraded document images. Recovering text from degraded document images is a very difficult task due to inter/intra variation between background and foreground pixels. Various binarization approaches are available to recover the degraded document image with their own pros and cons. This paper accomplishes a comprising survey of recent ongoing different binarization techniques.**

***Index Terms*— Binarization; inter/intra variation; degraded document** *(key words)*

I. INTRODUCTION

Historical documents are important as they contain some important information sometimes not available in the recent trends. To get this information from these documents are the main motive for binarization. All the historical documents are handwritten. Handwritten text within document images shows different changes in stroke width, stoke connection etc. In addition historical document images are often degraded by bleed-trough, where the ink of the other side seeps through to the front and different types of imaging artifacts like water blobs. Several historical documents suffer from fungus as they begin to age over a period of time. These documents are digitized into images and they can be also degraded by uneven illumination, image contrast variation, background noise and smear.

Binarization is usually performed in preprocessing stage. The aim of the binarization is to separate foreground text from the document background because the text belongs to foreground and the motive is to recover the text from the historical document images. Thresholding is important parameter refers to the conversion of gray-scale image to a binary image. The thresholding of degraded documents is a major challenge. Primarily image binarization approach is a segregation of pixel values in two different groups, white as background and black as foreground.

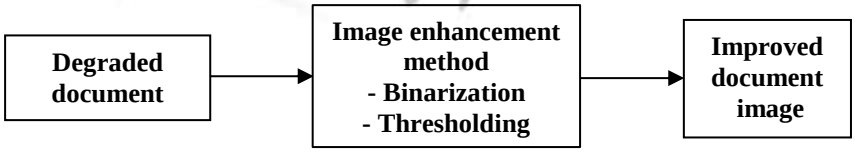
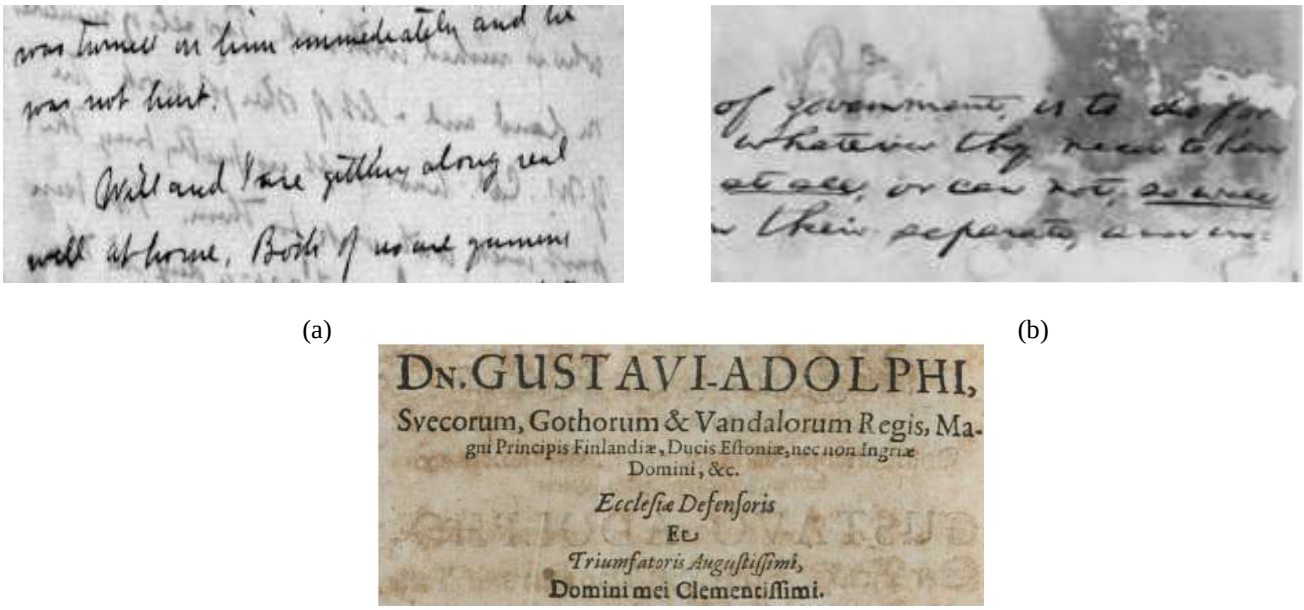


Figure 1 Binarization of degraded document image [9]

Some degraded document images are shown below, in which different kind of degradation are present. Every year the contest for image binarization methods is held by the experts that is known as DIBCO (Document Image Binarization COntest), images are taken from this series dataset only. Images (a) and (b) are taken from DIBCO 2009 & 2011 handwritten dataset respectively; image (c) is taken from H-DIBCO 2010 machine printed dataset. Binarization deals with this kind of degradation shown in the images and recover the images in its normal original form.

Many document image binarization techniques have been reported for binarization of degraded document images. Current state of the art binarization techniques have been reviewed in the paper.



(a) (b)

(c)

Figure 2 Degraded document image examples (a)–(c) are taken from DIBCO series datasets [1, 2, 3]

II. DEGRADED DOCUMENT IMAGE BINARIZATION TECHNIQUES

The key feature of robust document image binarization **[12]** is adaptive image contrast. Inspired by Bernsen binarization **[7]**, Adaptive contrast map is constructed to handle more complex background degradations. The local contrast computed by the local image maximum and minimum is used to minimize the background variation. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) identifies the local image difference that is similar to the image gradient. The denominator is a normalization factor that partially removes the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to balance the numerator and results in a relatively low image contrast.

To overcome the over-normalization problem, Local image contrast and local image gradient is combined to construct the adaptive contrast map, that is tolerant to different image arifacts and text to background variation caused due to different document image degradations. An adaptive contrast map is first constructed then combined with Canny's edge map to properly detect edge pixels of text stroke. Then to detect Text stroke edge pixel candidates, global Otsu thresholding is applied on contrast map. Now binarization map is constructed. To extract text from the document background, high contrast text stroke edge pixels are computed. The document text is again segmented by local threshold that is estimated from the detected text stroke edge pixels using local neighborhood window. At last stage, post-processing is performed to remove remaining artifacts and noise. The isolated foreground pixels that do not connect with other foreground pixels are filtered out in post processing.

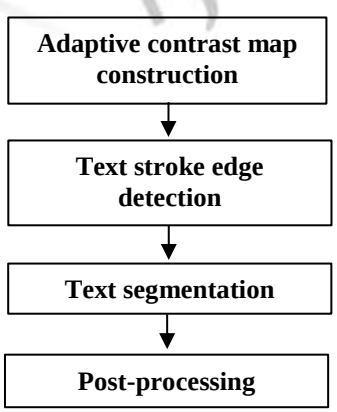


Figure 3 Robust document image binarization [12]

Global (adaptive) form of Otsu method (AdOtsu) **[13]** is used for better binarization. Due to high variations of degradation in document images, adaptive threshold based methods are required. Otsu method can be good for such degradations but it is non-adaptive. So, using this fact Otsu's global threshold based adaptive form (AdOtsu) is introduced by the author. The background can be estimated using a novel multiscale approach. Generalization of AdOtsu method in multiscale form is done using

framework **[14]**. Background estimation is used in the method, which is very successful for high-intensity document images having more degraded background. Otsu's threshold is used for segmentation. Finally, the computational cost is reduced using grid-based modeling. To remove remaining artifacts and non-connected sub-strokes, skeletonization is used in post processing.

In adaptive waterfall model **[15]**, surface of an image is considered as 3D terrain and pour water on it. The water fills the valleys after finding them. The wet regions are considered as blobs. Blob extraction is done. Then a multilayer perceptron is used to classify the blobs as the text or non-text. It preserves stroke-connectivity because the blobs are classified instead of pixels. The method works best on document images having uneven illumination. This method is based on the watershed segmentation **[16]** for the physical significance. It is obvious that that the text (foreground) pixels are darker than their neighboring background pixels, this fact is used in the method.

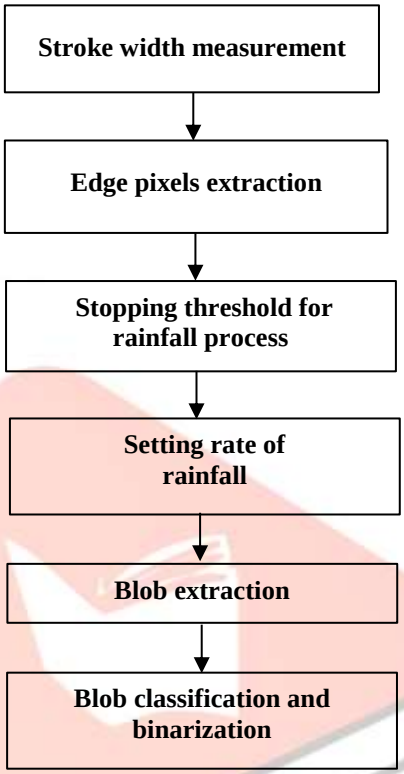


Figure 1 An adaptive water flow model [15]

To solve the problem of low detection rate of faint characters in binarization of handwritten document images, a combination of a global and a local adaptive binarization method **[17]** with connected component is used that improves overall performance too. Initially, for background compensation, background estimation is applied along with image normalization. Afterwards, global binarization is performed on the normalized image. In the binarized image noise components are discarded and the stroke width and the contrast are computed. Furthermore, local adaptive binarization is performed on the normalized image by taking above mentioned characteristics of image into consideration. Finally, the two binarization outputs are combined with connected component left.

The main idea of background estimation is to perform inpainting using Niblack's binarization output as the inpainting mask. Niblack's foreground result is dilated using a 3x3 mask in first stage to remove background pixels near the character edges having intensity closer to the foreground intensity. At this stage, fixed parameter settings for Niblack that can handle effectively most degradations, i.e. window size w = 60 and k = - 0.2. More background noise is erased when global Otsu is applied on the normalized image. Background estimation after image normalization is often used to balance the illumination of a picture taken under inappropriate lighting conditions. This method modifies the image to the direction of a bi-modal distribution that can be globally binarized with improved results.

Post-processing result of Otsu detects the stroke width of the characters and can be used to compute the average value and the standard deviation of the text. Firstly, the post-processing result of Otsu is skeletonized before the final result of binarization. Skeleton is also used to calculate the average value and the standard deviation of the foreground that are required to compute the image contrast. The estimated background image was calculated in accordance with the minimum of the 4 directional inpainting passes to have better normalization and ultimately better binarization results.

The result of the Otsu binarization on the normalized image contains low background noise but fails to recover the faint parts of the characters. Opposite to that, the Niblack result contains much background noise but detects the faint characters efficiently. To avoid the noise at the Niblack result, keep every connected component of Niblack that has common foreground pixels with the Otsu binarization result after post-processing. Therefore, Otsu result gets combined with the Niblack result, because Niblack achieves high separation of text-to-background when applied on the normalized image.

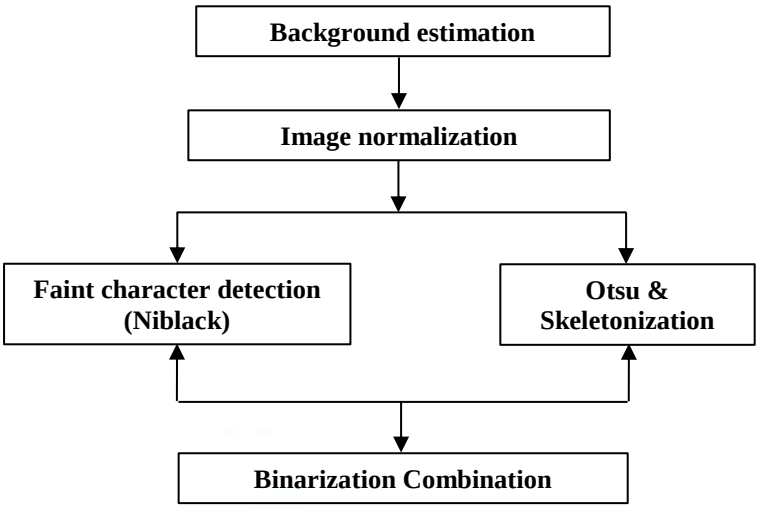


Figure 2 Combined binarization approach [17]

Document image binarization using local features and Gaussian mixture modeling **[18]** is performed in three stages in the method: First, every possible stains and general document background information are removed from the image through a background removal stage. The remaining misclassified background and character pixels are then separated using a Local Co-occurrence Mapping, local contrast and a two-state Gaussian Mixture Model. Finally, some isolated misclassified components are removed by a morphology operator.

In the first stage, the background removal technique is enhanced by selecting the window size automatically for the median filter and improving the threshold selection between the document image and the estimated background. In the second stage, the local neighborhood representation is redesigned to also include local contrast information to enhance the presence of character outlines. Binarization is then performed by separating two clusters of document characters and background artifacts that were not removed during the first stage of background removal. Clustering is performed using Mixtures of Gaussians (MoG). The Gaussian having lowest mean value belongs to the character cluster. In the final post processing stage, small-size 8-connected clusters are removed to exclude possible binarization noise.

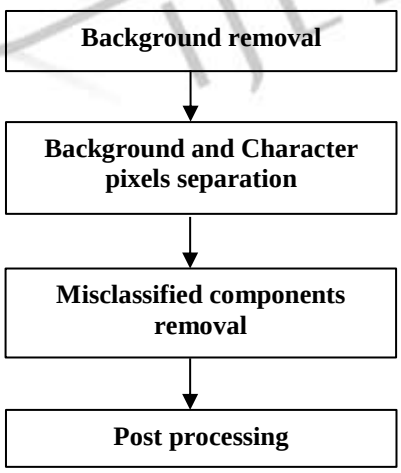


Figure 3 Binarization using local features and Gaussian mixture modeling [18]

III. SUMMARY

In this paper, different binarization techniques are discussed that are recently in trend. There are so many binarization methods available that can work well with particular type of degradation, but binarization method that can handle any type of degradation is still left for future work. A comparative study made on various binarization methods which will help to an individual to select the suitable binarization technique as per need of problem. The binarization is important aspect for document image analysis and recognition system. The main goal of binarization is to separate text from the degraded document background. there is need to propose a fast and accurate binarization technique which suitable for all types of degraded document images or a system which will automatically select and apply the suitable binarization technique depending on the document image for better performance.

IV. 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. Handwrit. Recognit., Nov. 2010, pp. 727–732.
[4] N. Otsu, "A threshold selection method from gray level histogram," IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.
[5] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.
[6] W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[7] J. Bernsen, "Dynamic thresholding of gray-level images," in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.
[8] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.
[9] 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.
[10] S. Rachmawati Yahya, S. N. H. Sheikh Abdullah, "Review On Image Enhancement Methods of Old Manuscript with the Damaged Background", International Conference on Electrical Engineering and Informatics 5-7, August 2009, Selangor, Malaysia.
[11] Chaki N., Shaikh S.H., Saeed K., "A Comprehensive Survey on Image Binarization Techniques. In: Exploring Image Binarization Techniques", Vol. 560, 2014.
[12] Bolan Su, Shijian Lu, and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE transactions on image processing, vol. 22, no. 4, pp. 1408-1417, april 2013.
[13] Reza Farrahi Moghaddam, MohamedCheriet, "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization", International Journal on Pattern Recognition, Vol. 45, No. 6, Pages 2419-2431, June 2012.
[14] R. Farrahi Moghaddam, M. Cheriet, "A multi-scale framework for adaptive binarization of degraded document images", Pattern Recognition, Vol. 43, No. 6, Pages 2186–2198, June 2010.
[15] Morteza Valizadeh, Ehsanollah Kabir, "An adaptive water flow model for binarization of degraded document images", International Journal on Document Analysis and Recognition, Vol. 16, No. 2, Pages 165-176, June 2013.
[16] Kim, I.K., Jung, D.W., Park, R.H., "Document image binarization based on topographic analysis using a water flow model", Pattern Recognition, Vol. 35, No. 1, Pages 265–277, January 2002.
[17] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis, "A combined approach for the binarization of handwritten document images." Pattern recognition letters, Vol. 35, Pages 3-15, January 2014.
[18] Mitianoudis N, Papamarkos N, "Document image binarization using local features and Gaussian mixture modeling." Image and Vision Computing, Vol. 38, Pages 33-51, June 2015.