A New Image Binarization Technique by Classifying Document Images

Soumik Datta*, Pawan Kumar Singh, Ram Sarkar, and MitaNasipuri

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
{soumik.datta86,pawansingh.ju,raamsarkar,
mitanasipuri}@gmail.com

Abstract. The present work proposes a binarization algorithm based on classification of document images. The method first classifies the images into two categories namely, simple and complex images. The global threshold value is used for binarizing the simple document images whereas complex document images are binarized by applying local threshold values. A background checking method is introduced in this method to detect the blocks which can be marked as purely background blocks. Finally, a post-processing mechanism has been applied to improve the quality of the binarized image.

Keywords: Binarization, Document image analysis, Optical Character Recognition, Global Thresholding, Local Thresholding, Simple Document Images, Complex Document Images.

1 Introduction

We are standing in the fifth generation of computer - the age of automation and artificial intelligences. During the last four decades, attention was paid in document analysis through computer. Document analysis involves handwriting recognition, writer identification, signature verification, etc. Most of these algorithms take binarized image as input to reduce the complexity and computational cost of the algorithms. Thus, binarization is the forerunner of many image processing techniques used in document image analysis. It is the process of converting 256 levels of grayscale information into two levels (black and white) image information. To binarize a grayscale image at first, threshold value(s) has to be determined. If a pixel value is less than the threshold value, then the pixel value of the corresponding output image is set as 1 (black) otherwise it is set as 0 (white).

Binarization is a challenging task [1-4] when noise is contaminated with the document image due to various reasons. In case of degraded documents, often the background and data pixels are misclassified if optimal threshold(s) has not been determined by the binarization process. It becomes more challenging when the following two cases occur: (i) the image contains text and graphic, (ii) the image contains text objects having a wide variety of gray-levels.

* Corresponding author.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 539–544, 2013.

2 Previous Work

In the literature, a lot of research work has been found on binarization algorithms. The algorithms can be broadly classified into three categories depending on the estimation of threshold values, namely: global thresholding algorithms, local thresholding algorithms and hybrid thresholding algorithms. In case of global thresholding approach, a single threshold value is determined to binarize the whole document. Image Binarization algorithms developed by Otsu [5], Kittler and Illngworth [6] belong to this category.

On the other hand, in case of local thresholding approach, the document image is divided into sub-images and threshold values are determined for each sub-image. Popular local threshold based binarization algorithms are Bernsen [7], Niblack [8], Sauvola [9], etc. Many recent research works are going on to combine the results of both global and local thresholding schemes to produce better results by removing the drawbacks of both types of algorithms. These algorithms are referred to as hybrid binarization algorithms [10-11].

3 Present Work

Textual documents often contain noise, shadow and other types of degradations. On the other hand, some document images are also found which are almost noise free. Therefore, the proposed binarization algorithm works into three modules (steps) to deal with both categories of the images:

- **Module 1:** Categorization of the document image, whether it is simple or complex.
- **Module 2:** Binarize the document image depending on its type.
- **Module 3:**Perform the post processing operations to improve the quality of the binarized images.

3.1 Categorization of Document Images

The proposed algorithm, first classifies the document images into two classes depending on the histogram analysis of the images namely— 1) Simple Document Images (*SDI*) and 2) Complex Document Images (*CDI*). The document images with very less or no noisy pixels are classified as *SDI*, whereas, the document images with a large amount of noisy pixels are classified as *CDI*.

Histogram of an image, which is a useful tool for numerous spatial domain image processing techniques, shows the frequency distribution of its pixels. To determine the type of the document, a technique proposed by S. H. Shaikh et al. [11] is used. At first, the number of peaks in the histogram of the input image is determined. A gray value 'P' will be determined as peak, if its frequency is greater than that of its two previous and two next gray values. In the next step, the average of all the peak values is determined which is called average peak. The peak, whose frequency is greater than the average peak value, is called sharp peak. If the image has at most two sharp peaks then the image is marked as *SDI* otherwise, it is *CDI*. Fig. 1 shows all types of peaks.

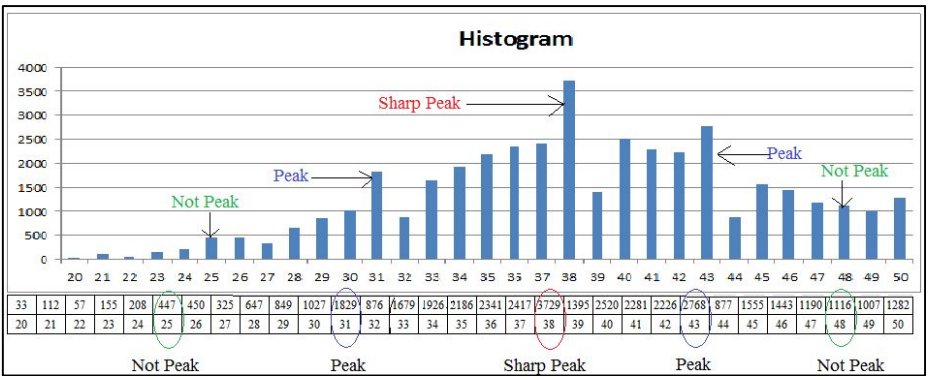


Fig. 1. Showing the peaks and sharp peaks in a part of an image histogram

3.2 Binarization of document images

- **Simple Image:** In case of *SDI*, the existence of at most two sharp peaks signifies (implies) that the intensities of the pixels possibly dispersed into two different categories. Among these, the first category implies the foreground whereas the second category implies the background. In this case, Otsu's methodology [5] is directly applied to binarize the document image.
- **Complex Image:** For *CDI*, the image has more than two sharp peaks which imply that the pixel intensities may be dispersed into more than two different categories. Thus, in this case Otsu's methodology may not be directly applied to binarize the document image. In the current work, the following technique is used to binarize such types of document images.

At first, the global threshold value as GTh for the whole document image is determined by using the following formula:

$$GTh = \frac{\text{Maximum } PixelIntensity\ Value + Minimum\ PixelIntensity\ Value}{2} \quad (1)$$

Next, the image is divided into blocks of NxN pixels. Then, maximum and minimum gray level values for each block are determined. After that *background blocks* are detected and merged. The process of detecting these blocks is known as background thinning. If all the pixels of a block belong to the background class then the block is referred to as *background block*. These blocks will not be considered for further processing. An image block is determined as a *background block* if the following two conditions are satisfied:

Condition 1: Maximum pixel intensity value must be greater than the global threshold for white background i.e., Maximum Gray Level Value > GTh

Condition2: The difference of the maximum and minimum pixel value must be less than or equal to B_t i.e., Maximum Gray Level Value – Minimum Gray Level Value ≤ B_t, where B_t is a constant.

The blocks other than *background blocks* are referred to as *data blocks* which are analyzed in the next step of the binarization process. Finally, the data blocks are binarized using Otsu's algorithm.

3.3 Post-processing

- **Post-processing of *SDI*:** To improve the quality of the binarized document image, a post-processing method has been introduced. For this, 8-connected neighbors of a particular data pixel in the original grayscale image are considered. If any of them is found to be a background pixel, then the grayscale value at that pixel position in the original image is checked. That pixel position is set as data pixel in the output binarized image if the said gray scale value is within 120% of the GTh. An extension of the threshold value is used to improve the image quality because some object pixels, become faded (whitish) due to degradation of the image, are misclassified as background pixels in the output image as the Otsu's Algorithm considers only the pixel intensity values and not the relation between the pixels.
- **Post Processing of *CDI*:** As *CDI* contain more noise, improving such images is more challenging. Another post processing operation, discussed below, is performed to remove those noisy pixels.

Step 1: Count the connected data pixel-clusters for each image block using flood fill algorithm.

Step 2: Repeat the steps for all the image blocks

2 (a): Read the value of the Threshold-cluster

2 (b): If data pixel-clusters ≥ Threshold-cluster

Make the image block as background block

End If

Step 3: End

More number of data pixel-clusters in a particular image block signifies that the block contains more noisy pixels. To remove these, a user defined parameter called Threshold-cluster is introduced. The performance of the post-processing method is heavily dependent on this parameter. The value of this parameter is tuned based on the nature of the document to yield better results.

4 Experimental Results of the Proposed Algorithm:

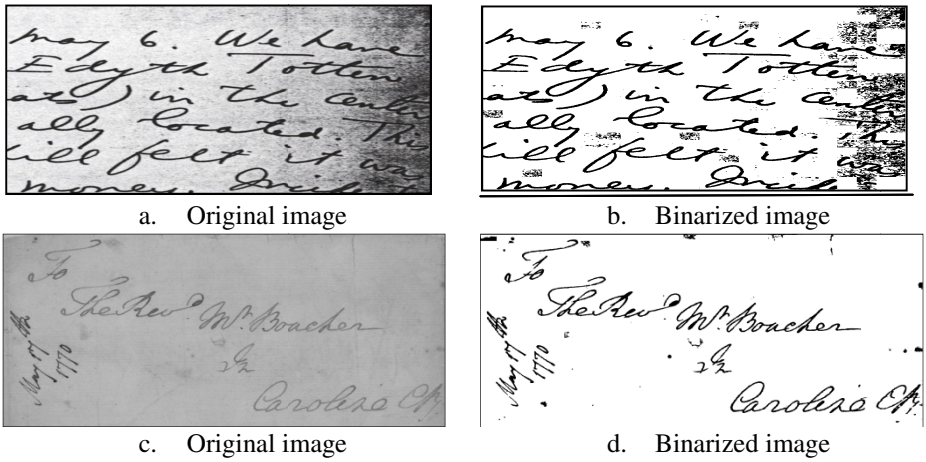
To evaluate the proposed binarization technique, document images are collected from different sources [12-14]. Both the good quality and degraded document images have been used to measure the strength of the algorithm. Different block sizes have been used for the *CDI* and the optimal result was found for the block size 30X30. Both the values of B_t and Threshold-cluster are determined experimentally. B_t is set to 30 and Threshold-cluster is set to 10 which produce optimum results for most of the cases.

The proposed technique produces satisfactory results on most of the document images. Some of the sample outputs of *SDI* and *CDI* are shown in Fig. 3 and Fig. 4 respectively.

a. Original image b. Binarized image

c. Original image d. Binarized image

Fig. 2. Binarized output images (b and d) of the original images (a and c) which are classified as *SDI*



a. Original image b. Binarized image

c. Original image d. Binarized image

Fig. 3. Binarized output images (b and d) of the original images (a and c) which are classified as *CDI*

5 Conclusion

Binarization is the process of converting a color or grayscale image into a bi-level image. The challenge lies in the task of determining the optimal threshold value(s) so that no background or data pixels are misclassified. In the present work, input document images are first categorized into 1) *SDI* and 2) *CDI*. Otsu's methodology has been directly applied to the SDI. To improve the resultant image quality, a post-processing is performed on the output image to retrieve the missing data pixels. Whereas for *CDI*, the image is divided into NxN blocks and local threshold values are determined using Otsu's algorithms. Post processing of document image is performed by detecting and deleting the noisy data blocks. The efficiency of the proposed algo-

rithm can be improved if the block size of the image and the value of the Threshold-cluster can be set dynamically. However, the present binarization algorithm is producing satisfactory results for a reasonable number of degraded documents.

References

1. Valizadeh, M., Armanfard, N., Komeili, M., Kabir, E.: A novel hybrid algorithm for binarization of badly illuminated document images. In: 14th International CSI Computer Conference (CSICC), pp. 121–126 (2009)
2. Kawano, H., Oohama, K., Maeda, H., Okada, Y., Ikoma: Degraded document image binarization combining local statistics. In: ICROS-SICE International Joint Conference, August 18-21 (2009)
3. Chang, Y.F., Pai, Y.T., Ruan, S.J.: An efficient thresholding algorithm for degraded document images based on intelligent block detection. In: IEEE Int. Conf. Syst. Man Cybern. SMC (2008)
4. Gatos, B., Pratikakis, I., Perantonis, S.J.: Efficient binarization of historical and degraded document images. In: The Eighth IAPR Workshop on Document Analysis Systems (2008)
5. Ostu, N.: A thresholding selection method from gray-level histogram. IEEE Trans. Systems Man Cybernet. SMC 8, 62–66 (1978)
6. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recognition 19(1), 41–47 (1986)
7. Bernsen, J.: Dynamic thresholding of gray-level images. In: Proc. Eighth International Conference on Pattern Recognition, Paris, pp. 1251–1125 (1986)
8. Niblack, W., Prentice, N.J., Cliffs, E.: An Introduction to Digital Image Processing (1986)
9. Sauvola, J., Pietikainen, M.: Adaptive Document Image Binarization. Pattern Recognition 33, 225–236 (2000)
10. Su, B., Lu, S., Tan, C.L.: Combination of Document Image Binarization Techniques. In: International Conference on Document Analysis and Recognition (2011)
11. Shaikh, S.H., Maiti, A.K., Chaki, N.: A new image binarization method using iterative partitioning. Springer (2012) (published online: January 6, 2012)
12. http://users.iit.demokritos.gr/~bgat/DIBCO2009/benchmark/
13. http://users.iit.demokritos.gr/~bgat/H-DIBCO2010/resources.html
14. https://www.google.co.in/imghp?hl=en&tab=wi