ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Binarization of Multi Background Text Images for OCR

Saranya Kanagaraj, Manoj Ravindra Phirke

Imaging Tech Lab, HCL Technologies, Bangalore, India

ABSTRACT: This paper proposes a new binarization technique for scanned Images containing text in different backgrounds. Document images with different background, during binarization losses some data in it because of performing adaptive binarization. The proposed method aims to retrieve text information present in varying background document images without any data loss. This method will be applicable for lossless retrieval of text from document images (OCR). The advantages and disadvantages of proposed method are demonstrated. Experiments have been conducted and results are presented to show the effectiveness of proposed method.

KEYWORDS: Adaptive binarization, Document images, OCR (Optical Character Recognition)

I. INTRODUCTION

Text Image binarization could be a crucial step in document image process. For the purpose of optical character recognitions, when we go for RGB images, we will have a big range of colors. In RGB images, it may contain various backgrounds with various foregrounds and hence when we try performing OCR on an RGB image some foreground data may be missed. Similarly when we go for grayscale converted images, they will also have a range of pixels. All the required foreground data will not be of the same color and hence grayscale images will also have some losses during OCR. To overcome the difficulties in OCR, we go for binarization. Text Image binarization involves a task that automatically converts the document images from a grayscale or color image into a binary image in a manner that foreground info is represented by black ones and background info by white pixels. This process of thresholding applies to allow document to be recognized and retrieved more efficiently.

II. RELATED WORK

Binarization has been a subject of intense research interest during the last ten years. Most of the developed algorithms rely on statistical methods, not considering the special nature of document images. However, recent developments on document types, for example documents with mixed text and graphics, call for more specialized binarization techniques.

Several Binarization techniques have been proposed in the past decades. Those techniques can be classified into two different types, one using Global thresholding and the other using local thresholding. Global thresholding will apply only one threshold which is chosen based on some strategic features of whole image. The main drawback of this method is that it cannot adapt to uneven illumination in images, noise produced, lower resolution, improper structure of images (mixture of graphics and text) and for images with multiple background. In local thresholding, the thresholds are locally chosen based on the varying contents of the image. In local thresholding the threshold values are determined locally, e.g. pixel by pixel, or region by region. Then, a specified region can have 'single threshold' that is changed from region to region according to threshold candidate selection for a given area. Local thresholding overcomes the disadvantages of global thresholding by its quality of adaptability.

In [6] & [7], the authors have proposed one such binarization technique. The article [6] the author obtains local information of images using connectivity and image intensity. This method contains both global and locally adaptive approaches. In [7] the author uses three stages of image fusion, FastICA and K-Harmonic classifiers for binarization. Though these binarization techniques are accurate and work well for document images, information on its performance for providing unique output representation of foregrounds (Texts) and backgrounds are not present.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

In comparison with the global thresholding, local thresholding performs better. One such local thresholding method is Adaptive binarization. This method splits the document image into various sub components based on luminance and calculates local thresholds based on the gray scale value of the sub components. When compared to global thresholding method, adaptive binarization bought a good revolution in binarization of document images with rapid luminance changes. Such binary outputs are further fed to character recognition. Compared to global thresholding, adaptive binarization yielded better character recognition results for low quality document images.

Though the technique of adaptive binarization works well for degraded or low quality document images, it fails to yield good results for most of multi background document images and for images with different structure like a combination of graphics and texts. Several solutions have been provided for document images binarization in terms of illumination ([1] [2] & [4]) but the factor of varying background color were not taken into picture.

For binarization of multi background images, [5] has proposed a novel method using edge detection and connected components. The method employs an edge-based connected component approach and automatically determines a threshold for each component. It has several advantages over existing binarization methods. The algorithm was designed for camera captured images. Screen captured images were not focused. In situations like Optical character recognition of screens of various smart machines which interface with user like ATM, vending machines, etc. output of screen captured images will be more accurate than camera captured images.

This paper proposes one such binarization method to overcome them. In many practical applications, we need to recognize or improve mainly the text content of the documents. In such cases, it is preferable to convert the documents into a binary form for good character retrieval. The goal of this approach is to convert the given input grayscale or color document into a bi-level representation. This representation is particularly convenient because most of the documents that occur in practice have one color for text (e.g. black), and a different color (e.g. white) for background. The method followed in the proposed method becomes a simple but effective tool to separate objects from the background. The output of this operation is a binary image whose one state will indicate the foreground objects, that is, printed text, a legend, a target, defective part of a material, etc., while the complementary state will correspond to the background. Depending on the application, the foreground can be represented by gray-level 0(Black), that is, black for text, and the background by the highest luminance for document paper, that is, 1(White) in 8-bit images as white, or conversely the foreground by white and the background by black. The proposed method is explained in details in further sections.

III. THE PROPOSED ALGORITHM

The proposed method enhances the techniques of adaptive binarization by performing some post processing steps to adaptive binarization. When adaptive binarization is performed for scanned text images with multiple backgrounds, the output binary image will contain some background in black and some background in white. Also the text present in black background will be white in color and for white backgrounds; it will be black in color. When such images are fed for text/character retrieval, the text extraction would not be complete because of variation in background colors. Steps of proposed algorithm are discussed below.

A. Algorithm

i. Convert Input image to gray scale
ii. Perform OSTU binarization
iii. Perform Connected Components
iv. Perform thresholding to eliminate components below the threshold fixed.
v. Obtaining the final mask.
vi. Logically Ex-or between the Mask image obtained and the OSTU output

B. Description

The proposed approach solves the problem of adaptive binarization by performing three new steps to the adaptive binarization output. For the adaptive binary output obtained, connected components, thresholding followed by masking are performed. The adaptive binarization algorithm used in the proposed method is OTSU [3]. The OTSU algorithm assumes that the image contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), it then calculates the optimum threshold separating the two classes so that the output will be binary. From [3], the following equations were used to formulate adaptive threshold.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

For the two classes separated by a threshold *th*, the Weights of the classes is Wt_i and variances of the classes is S_i. Otsu in [3] shows that minimizing the intra-class variance is the same as maximizing inter-class variance. Hence

$$S_b^2(t) = Wt_1(t) * Wt_2(t) [M_1(t) - M_2(t)] \quad (1)$$

where Mi(t) is the class mean. Class mean and weight (class probability) can be obtained from the following equations,

$$Wt_i(t) = \sum Pixel(i) \quad (2)$$

$$M_i(t) = [\sum Pixel(i)* Cv(i)] / Wt_i \quad (3)$$

Pixel is the class array, *Cv(i)* is the value at centre of i^{th} histogram bin and t is the adaptive threshold. The *th* value corresponding to the maximum Si is the final threshold. The final threshold obtained will be in the range of 0-1 and hence it is mapped to scale 0-255. The obtained threshold is applied on the grayscale image to get binary output.

OTSU binary output will contain a varying background color and text color as stated earlier. For the OTSU binary output obtained, connected components algorithm is performed in order to locate the part of images with varying background color.

For example, let the OTSU binary output contain three different parts with first part having black background and white text, second part having white background with black text and the third part similar to first part. So now when connected components algorithm is performed to identify black components, those varying background components with black background and white text and each black text is identified as a component.

In our example, along with several black texts, the first and third part will also be identified as a component. Among them the text components will be of smaller areas and the black background components will be of bigger areas. For the obtained output, thresholding is performed using the following equation such that the components with area above our threshold will be present and the rest (small black texts) will be eliminated.

$$\text{Selected Components} = C(CA > \text{threshold}) \quad (4)$$

Where, C is the set of output components of step connected component, CA is the area of every component in pixels, and threshold is the desired threshold value for performing thresholding.

Threshold values for thresholding will be dependent on the fonts in images. The threshold output image will contain only the varying background components along with white text inside. In order to eliminate the text inside them, the holes are filled with the background color (background color- black) itself. The final output after the process of thresholding will be used as a mask. The masked output (MaskImage) will contain only the black background components completely in black and rest in white. Now logical operation exclusive-or (Ex-Or) is performed between the masked output and OTSU output (AdaptiveBinaryImage).

IV. EXPERIMENTATION

The proposed algorithm was tested on some sample image. In the example, fig.1 is the original sample from the dataset. The input RGB image is converted into gray scale and OTSU binarization is performed. Fig.2 is the OTSU output of fig.1. In the OTSU output, as the first step of masking, a thresholding is performed by removing regions below the threshold fixed using connected components approach.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

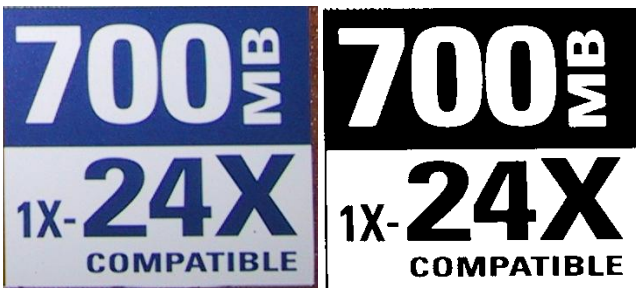

Fig. 1 Original sample image, Fig. 2 Adaptive Binary output

Fig. 3 Removal of White foreground present in black background of adaptively binarized image. Fig. 4 Final mask after removal of black foreground present in white background of fig. 3.

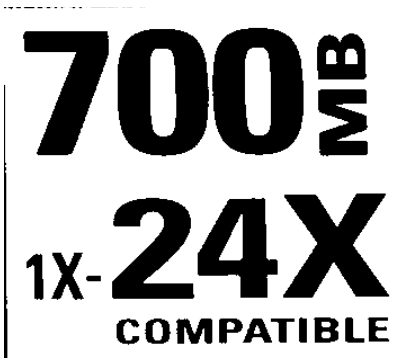

Fig. 5 Final complete binarized output

As an outcome of this step, all the white text in black region will be removed and the same can be seen in fig.3. In the same manner, for fig.3, second step of masking (i.e.) the removal of black text in white region is performed and the output can be seen in fig.4. Fig.4 is the final mask image. This image when logically Ex-or with OTSU output, a complete binarized image shown in fig.5 can be obtained.

The output of proposed binarization algorithm in fig.11 is compared against various binarization algorithms like OTSU in fig.10, Mean Adaptive binarization [9] fig.6, Gaussian adaptive binarization [9] in fig.7, Niblack binarization [8] in fig.8 and Sauvola [10] in fig.9. All the algorithms were executed with default values. The proposed output is better than those binarization algorithms.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

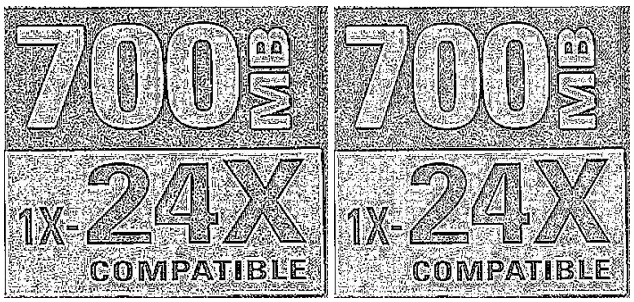

Fig. 6 Mean Adaptive output Fig. 7 Gaussian Adaptive

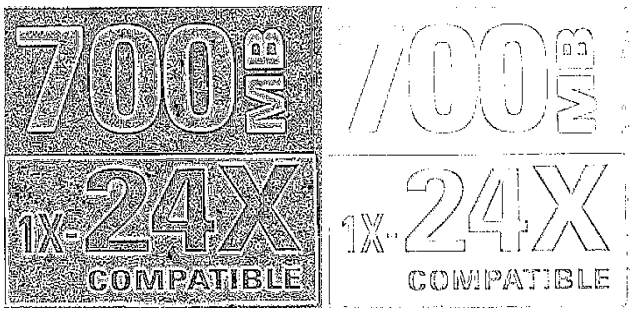

Fig. 8 NiBlack's Output Fig. 9 Sauvola Output

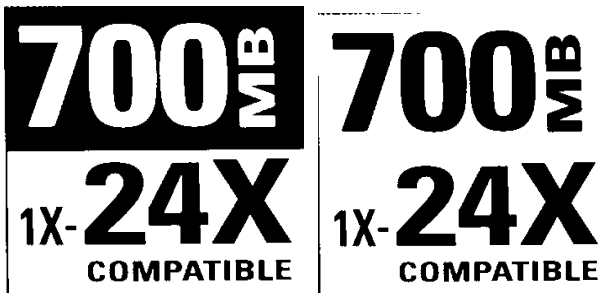

Fig. 10 OTSU output Fig. 11 Proposed Binarization output

V. CONCLUSION

The proposed algorithm enhances OTSU output in terms of unique representation of text and background colors such that, these enhancements make character recognition of scanned images perfect. This algorithm is helpful in getting consistent background (white) and foreground (black); in spite of variations present in input images. This approach performs well for scanned document images than camera captured image. After performing our binarization, final output can be subjected for further processing of Optical character recognition. Though the proposed algorithm was focused on unique representation of texts and background for multi background images, focus can be extended to suit for illumination changes also. Also algorithm can be enhanced to work for images other than scanned images.

ACKNOWLEDGMENT

The authors would like to express their gratefulness to Mr. Vasudeva yelluru rao, Operations Director, Imaging Tech Lab, HCL Technologies, Bangalore for providing his support for the article. Also we would like to thank Dr.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Punitha Puttu Swamy, Imaging Tech Lab, HCL Technologies, Bangalore who had reviewed our paper and guided in selection of conference.

REFERENCES

[1] Meng-Ling-Feng, Yap-Peng-Tan – "Contrast adaptive binarization of low quality document images" IEICE Electronics Express Vol.1, No 16,501-506.
[2] PeetaBasaPati 2, A G Ramakrishnan – "Binarization and Localization of Text Images Captured on a Mobile Phone Camera" ICISIP 2006, DOI: 10.1109/ICISIP.2006.4286101-Bhavna Antony 1.
[3] Nobuyuki Otsu – "A threshold selection method from gray-level histograms" IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66. doi:10.1109/TSMC.1979.4310076.
[4] B Gatos,I. Pratikakis, S.J. Perantonis - "Adaptive degraded document image binarization" Elsevier-Pattern Recognition 39 (2006) 317 – 327.
[5] T Kasar, J Kumar and A G Ramakrishnan – "Font and Background Color Independent Text Binarization" International Workshop on Camera Based Document Analysis and Recognition 2007.
[6] Lawrence O'Gorman – "Binarization and Multithresholding of Document Images Using Connectivity" – Elsevier Volume 56, Issue 6, November 1994.
[7] Nikolaos Mitianoudis and Nikolaos Papamarkos – "Multi-Spectral document image binarization using image fusion and background subtraction techniques"- IEEE - Image Processing (ICIP), 2014.
[8] Niblack Adaptive thresholding Matlab code of Jan Motl – W.Niblack, "An Introduction to Digital Image Processing". Prentice Hall, Englewood Cliffs, (1986).
[9] "Mean adaptive thresholding and Gaussian adaptive thresholding" – http://docs.opencv.org/master/d7/d4d/tutorial_py_thresholding.html#gsc.tab=0.
[10] J. Sauvola and M. Pietikainen, "Adaptive document image binarization" Pattern Recognition 33, 2000