www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 3 March 2015, Page No. 11023-11025

Robust Document Image Binarization Technique for Degraded Images

Gaurav Divate[1],Aniket Vaidya[2],Vipul Wankhede[3]

[1]SSPU University,GN.Sapkal College of Enggineering ,
Anjaneri, Nashik,Maharashtra.
gauravsdivate@gmail.com

[2]SSPU University,GN.Sapkal College of Enggineering ,
Anjaneri, Nashik,Maharashtra.
Aniketvaidya13@gmail.com

[3]SSPU University,GN.Sapkal College of Enggineering ,
Anjaneri, Nashik,Maharashtra.
Vipulwankhede@gmail.com

Abstract: ***Segmentation of text from badly degraded document images is a very challenging task due to the high inter/intravariation between the document background and the foreground text of different document images. We propose a novel document image binarization technique that addresses these issues by using adaptive image contrast. The combination of the local image contrast and the local image gradient that is tolerant to text and background variation caused by different types of document degradations is called as The adaptive image contrast. In the system, an adaptive contrast map is first constructed for an input degraded document image. The contrast map is then converted into binary and combined with Canny's edge map to identify the text stroke edge pixels. Then document text is 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 robust, simple and involves parameter tuning at its minimum. It has been tested on datasets that are used in the recent document image binarizationcontest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010 and achieves accuracies of 93.5%, 87.8%, and 92.03%, respectively, that are higher than or close to that of the bestperforming methods reported in the three contests. The Bickley diary dataset experiments that consists of several challenging bad quality document images also show the superior performance of the system over the others.***

Keywords: Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.

1. Introduction

Image Binarization, 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. Binarization is a process where each pixel in an image is converted into one bit and you assign the value as '1' or '0' depending upon the mean value of all the pixel.if pixel value greater then mean value then its'1' otherwise its'0'. A binary image is a digital image that has only two possible values for each pixel. Two colors used for a binary image are black and white though any two colors can be used. The used colorin the object(s) in the image is the fore-ground color while the rest of the image is the background color. In the document scanningindustry this is often referred to as bi-tonal. For the ensuing document image processing tasks such as optical character recognition(OCR) a fast and accurate document image binarization technique is importantThe handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke brightness, stroke width, document background and stroke connection. Historical documents are often degraded by the bleed-through where the ink of the other side seeps through to the front. Also, historical documents are often degraded by different types of imaging artifacts. These types of degradations of documents tends to induce the document thresholding error. 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 different corresponding binary imageSome 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.

2. History

During Current Market survey, we found that there is no such system which will give a better accuracy than our proposed system. So, we can say that there is only one way for degraded images which will give 90 to 95% accuracy than other system. Many thresholding techniques have been available for document image binarization. As various degraded documents which do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Existing system consist Global thresholding method which is based on window based adaptive thresholding which consist character stroke width window size which is very complicated process. Many thresholding techniques have been

reported for document image binarization. Various degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization.Adaptive thresholding, which estimates a local threshold for each document image pixel, it is one of the a better approach to deal with different variations within degraded document images. As the early window-based adaptive thresholding techniques, estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. Severe drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Rest of the various approaches have also been reported, which includes subtractionof background, analysis of texture, recursion method, method of decomposition, completion of contour, Markov Random Field, matched wavelet, cross section sequence graph analysis of graph,learning by self andLaplacian energy user assistance and combination of binarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The contrast of local image and the gradient of local image are very useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighboring document background. They are much effective and used in many document image binarization techniques. In Bernse paper, the local contrast is defined as follows: C(i, j)= Imax(i, j)−Imin(i, j) where C(i, j) denotes the contrast of an image pixel (i, j), Imax(i, j) and Imin(i, j) denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively. If the local contrast C(i, j) is smaller than a value of threshold, the pixel is stored as background automatically.else it is classified into text or background by comparing with the Imax(i, j) of mean and Imin(i, j). Bernsen'sit is a simple method, but not working proper on degraded document images with a complex document background. We have proposed a novel document image binarization method by using the local image contrast that is evaluated as follows: C(i, j)= Imax(i, j)−Imin(i, j) Imax(i, j)+Imin(i, j)+ where is a positive but infinitely small number that is added in case the local maximum value is equal to 0. Comparison with Bernsen's contrast in Equation 1, the local image contrast in Equation2 introduces a normalization factor (the denominator) to compensate the image variation within the background of document. Put the text within shaded document areas such as that in the sample document image as an example.

3.Proposed System

Segmentation of text from badly degraded document is very challenging task due to high inter/intravariation between document background and foreground text of different document images. 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 usesadaptive 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. This technique addresses the over-normalization problem of the local maximum minimum algorithm.

3.1 System Architecture

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. Apart from that, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to uneven lighting, noise and bleed-through. To extract the stroke edges accurately, the image gradient have to be normalized to compensate the image variation within the document background. In our earlier method, local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient.

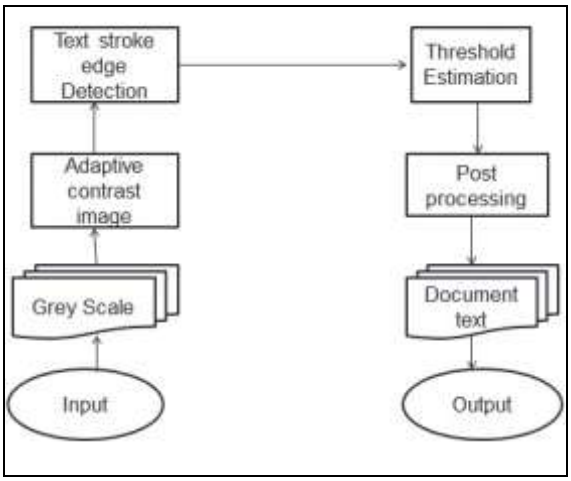

Figure1. System Architecture.

The purpose of the contrast image construction is to detectthe stroke edge pixels of the document text properly. Thediscovered contrast image has a accurate bi-modal pattern,In which adaptive image contrast calculated at text strokeedges is obviously larger than that computed within the documentbackground. The text can then be extracted from the document backgroundpixels once the high contrast stroke edge pixels aredetected accurately. Some characteristics can be observed fromdifferent kinds of document images [5]: Firstly, the text pixelsare close to the detected text stroke edge pixels. Another, thereis a distinct intensity difference between the high contraststroke edge pixels and the surrounding background pixels.Once the initial binarization result is derived from Equation5 as described in above sections, the binarization result can be further improved by incorporating certain domainknowledge as described in Algorithm. Firstly, the isolatedforeground pixels that do not connect with other foregroundpixels are filtered out to make the edge pixel set precisely.Another the neighborhood pixel pair that lies on symmetricsides of a text stroke edge pixel should belong to differentclasses.

4.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.

5. References

[1] Robust Document Image Binarization Technique for Degraded Document Images Bolan Su,
Shijian Lu, and Chew Lim Tan, Senior Member, IEEE

[2] 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.

[3]An Image Segmentation Algorithm in Image Processing Based on Threshold
SegmentationShiping Zhu, Xi Xia, Qingrong Zhang KamelBelloulata

[4] 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,