International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey on Document Image Binarization for Degraded Document Images

Yogita Kakad[1], Savita Bhosale[2]

[1]Mumbai University, MGM College of Engineering, Kamothe, Navi Mumbai, India

[2]Professor, Mumbai University, MGM College of Engineering, Kamothe, Navi Mumbai, India

Abstract: ***The Technology is connecting the whole world together by the medium of internet. Each segment of our data is present in the form of digital document. We can able to store, duplicate, and backup our data in digital form. But what we think about old data which is available in the form of traditional document. Sometimes the old documents plays important role in a major challenge. Many of the paper data is being degraded due to lack of attention. Many of these degraded documents have their front data mix up with rear data. To make this front data separate from backend data we have proposed binarized documentation technique. In this we firstly applying the invert contrast mechanism on degraded document. Then we are going to compare that with grey scald edge detection method and then we are applying the binarization method on that degraded image. This binarized image is further undergoes to the post processing module. The output of this all technique will produce a clear and binarized image.***

Keywords: Adaptive image contrast, document analysis, grey scale method, post processing, document image processing, degraded document image binarization, pixel classification.

1. Introduction

In today's era, image processing techniques has a wide scope. The image can be useful in many areas like astronomy, remote sensing, microscopy or tomography, cryptography. There are many images of the old novels which are very helpful for us but, due to non-maintenance these images are being unreadable. Due to imperfections of measuring devices and instability of observed scene, captured images are blurred, noisy and of insufficient spatial or temporal resolution. These images has becomes degraded images and we can't use them though they are very useful to us. Sometimes some images are degraded manually due to which the quality of the image is being lowered and that useful image becomes one waste image for us.

The degraded images are in the form of mixed foreground and background format. We can separate this background from the foreground text. The proposed technique gives the efficient way to separate these text from background noisy pixels. The image is passed through the several methods which will produce the output image which is in readable format. In this system, image Binarization is performed in the four stage of document analysis and it aims to separate the foreground text from the document background. The accurate document image binarization technique is important for the recovering document image by the help of processing tasks such as Contrast Enhancement. Though document image binarization the thresholding of degraded document images is solved now. It was due to the high inter/intra-variation between the text stroke and the document background across different document images. After contrasting the image next we apply the grey scale method to detect the text strokes present inside the image. As our method is purely based on the novel documents. It is less effective or non-effective on the images other than literature or novel images. After applying the grey scale method we are going to estimate threshold value for each pixel. Depending upon the threshold value binarization method will perform on the image. This binarized image still contents some background degradations. So we are going to apply post processing method on it. The processing will remove the all background degradations and it will produce the clear and readable image.

Figure 1: Example of binarization

2. Literature Survey

Many techniques have been developed for document image binarization. As we know that many degraded documents do not have a clear pattern and it may be in a bad condition. Thresholding alone is not a good approach for the degraded document binarization. Adaptive thresholding, which estimates a local threshold for each document image pixel, is generally a best approach to deal with different variations within degraded document images.

A. The Global Thresholding Technique

It estimates a resultant threshold for the overall image; these techniques requires few calculations and can effectively work in simple cases. But this technique fails if image contains complex backgrounds, such as non-uniform colour and poor illuminated backgrounds. These techniques are not efficient for degraded document images, because they do not have a clear pattern that separates foreground text and background [1].

(a)Input Image

(b)Output Image

Figure 2: Input and Output Image of global thresholding

B. Image Binarization Using Texture Features

This method is for binarization of historical and degraded document images, which is based on texture features of images. This technique is a versatile edge-based. This recent is processed by utilizing a scripter focused around a co-event framework of image. The proposed technique is tested on the basis of objects, utilizing DIBCO dataset debased documents and it is utilizing a set of old corrupted document gave by a national library [2].

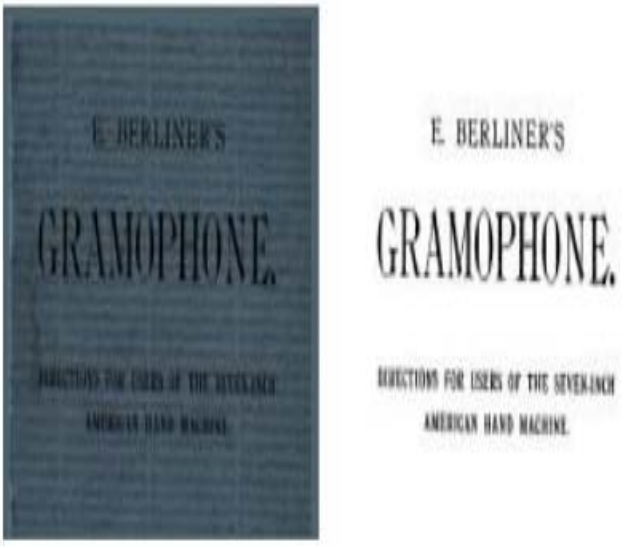


Figure 3: Input and output of the texture feature

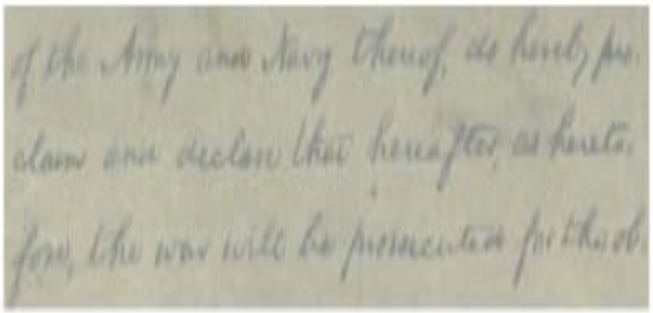
C. Adaptive binarization for degraded images:

This method utilize the dilation and erosion within light black-scale picture preparing; therefore get another picture in which the shadow levels and noise densities will be reduced. This method is combination of contrast and variation in contrast. The binarization method joined the system which enhanced Niblack and the neighbourhood thresholding utilizing the little neighbourhood which affected the mean value of the areas [3].

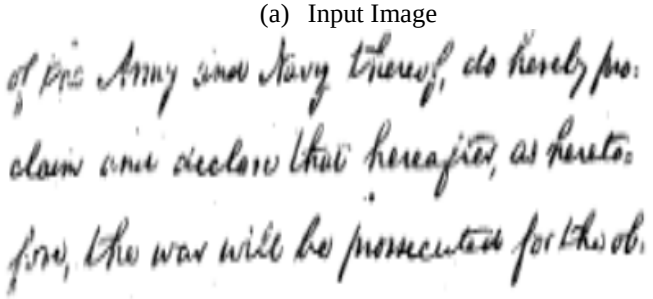
Figure 4: Input and Output image for adaptive thresholding.

D. Combination of Document Image Binarization Techniques.

In this technique hierarchical structure to relates different thresholding methods and generate improved performance for document image binarization is described. The given binarization output of a number of comparative methods, this system methodologies divides the document image pixels into three sets, which are named as, first foreground pixels, background pixels and uncertain pixels [4].



(a) Input Image



(b) Output Image

Figure 5: Results of the combined binarization techniques.

E. Dynamic Threshold Binarization

The binarization methods such as iteration method defines the threshold of a pixel with the grey level values of its own and neighboring pixels and the coordinate of each pixel. This image binarization method is commonly used for the bad quality images, especially the images with single – peak constructed histogram. However, owing to the dynamic threshold calculation, the method has high computation complexity and slow speed [5].

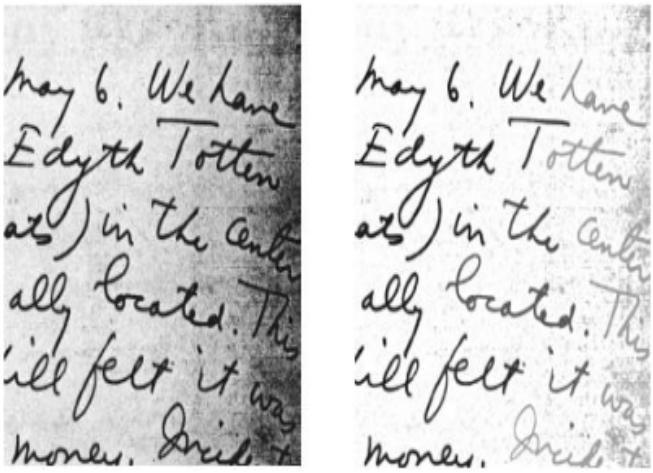


Figure 6: Dynamic threshold binarization

F. Otsu Method

Otsu method is solitary of the well - known global methods. This technique discovered the threshold T which split the gray level histogram into two segments. The computation of inter classes or intra classes variances is based on the normalized histogram of the image H = [h0.....h255] where {hi=1}.Otsu method is apply to routinely execute clustering based image thresholding method. In Otsu's we thoroughly look for the threshold that minimizes the intra - class variance distinct as a weighted sum of variances of the two classes.

$$\sigma^2 \mathrm{prb}(t) = \mathrm{prb}_1(t)\sigma_1^2 + \mathrm{prb}2(t)\sigma^2 2(t)$$

Here prb are the probabilities of the two classes divided by threshold and variances of the classes. The class probability and class means can be computed iteratively [6].

G. Brensen Method

It is an adaptive local technique of which the threshold is designed for each pixel of the image. For each pixel of coordinates(x, y) in image, the threshold is given by two researcher named as: Zlow and Zhigh are the minimum and the maximum gray level in a squared window r*r centered more than the pixel (x, y).

$$T(x,y) = \frac{Z_{low} + Z_{\mathrm{high}}}{2}$$

If the pixel having distinction quantity which is lesser than a threshold 1, then the neighborhood consists of a single class: background or text [7].

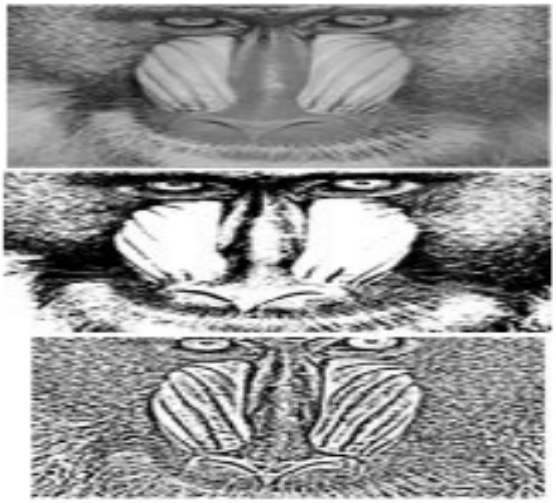


Figure 7: Applying brensen

H. Niblack Method

Niblack algorithm analyze a confined threshold for every pixel by descending a rectangular window above entire picture .the calculation of the threshold is based on confined mean m and the standard deviations of all pixels in the window and is given by: The threshold T is deliberate by using mean m and standard deviation σ of all pixels in the window.

$$Tniblack = m + k * s$$

$$Tniblack = \mathrm{m} + k\sqrt{\frac{1}{NP}\sum(\mathrm{pi} - \mathrm{m})^2}$$

$$m + k\sqrt{\sum\frac{pi^2}{NP} - m} = \mathrm{m} + \mathrm{k}\sqrt{\mathrm{B}}$$

Thus the threshold T is given by: T= m+ k * σ. Such s k is a parameter used for find out the number of edge pixels measured as object pixels and takes a negative values. Advantage of niblack is that it always recognize the text regions properly as foreground but it tend to generate a huge quantity of binarization noise in non-text region [8].

3. Proposed System

As we saw above, the proposed techniques have some limitations. To overcome these limitations our system uses new binarization technique along with grey scale method. There four modules in our system.

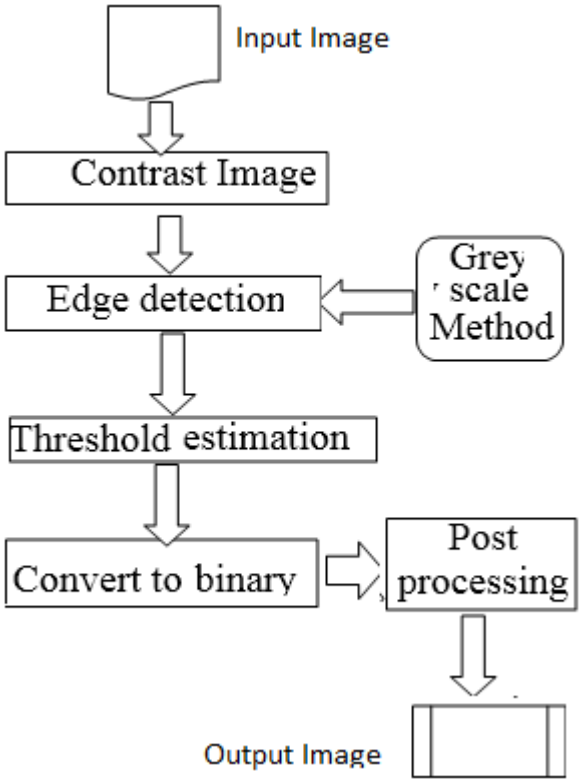


Figure 7: Proposed system Architecture

To detect the exact text stroke it is very necessary to adjust the level of contrast in the image. In this module we are keeping the image contrast at min or max level. It is depend upon how much the foreground text is mixed with background noise in the image. Here we invert the current level of image contrast i.e. here we are reversing the color of the image. The contrasted image is further match with grey scale method output image. Which further will produce the outline of the pixel around the foreground text. These pixels then divided into two categories. First category is related pixels and non-related pixels. Connected pixels occupies the area around text stroke. And non-related pixels shows the other noisy area present in the image.

The edge detected image is then converted into binary format of 0's and 1's. 0 indicates that the image pixels are non-connected pixels and 1 indicate that image pixels are connected pixels and the represents the text strokes. The pixel 0's are eliminated from the processing image because they are part of background image. Output of the binarization method creates separation in the image. So post processing eliminates the non-strokes image from binary image. And it returns a clear image which consist of only text actual strokes. Now this produced image is when compare with input image, then we can easily figure out the significance of our system. Output image contain clean and readable text.

4. Result and Analysis

We have used our system on various on various type of images, like novel, books, and records, historical literature. Some of the results are stated as follow.

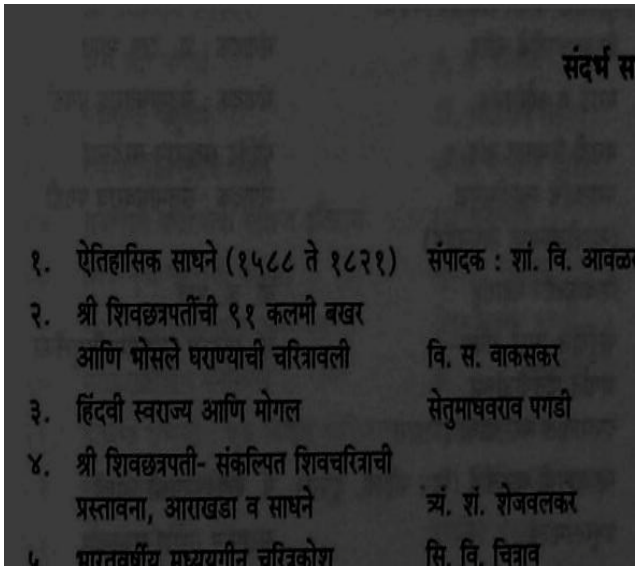
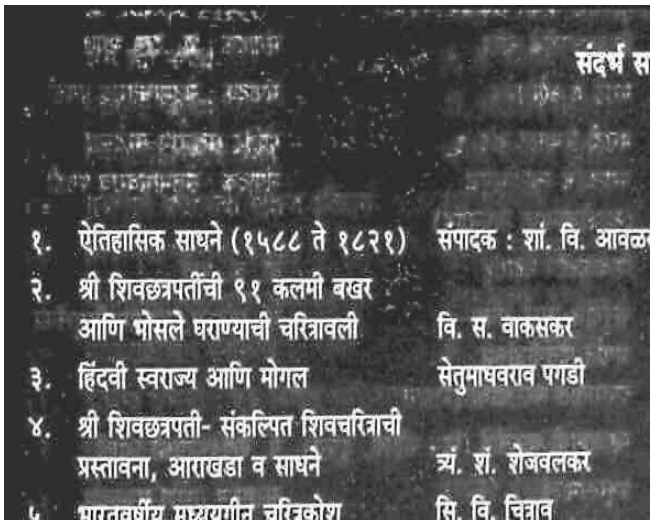
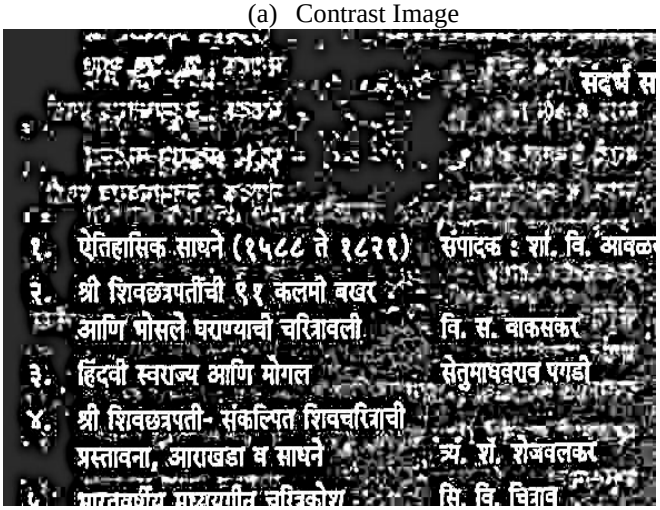


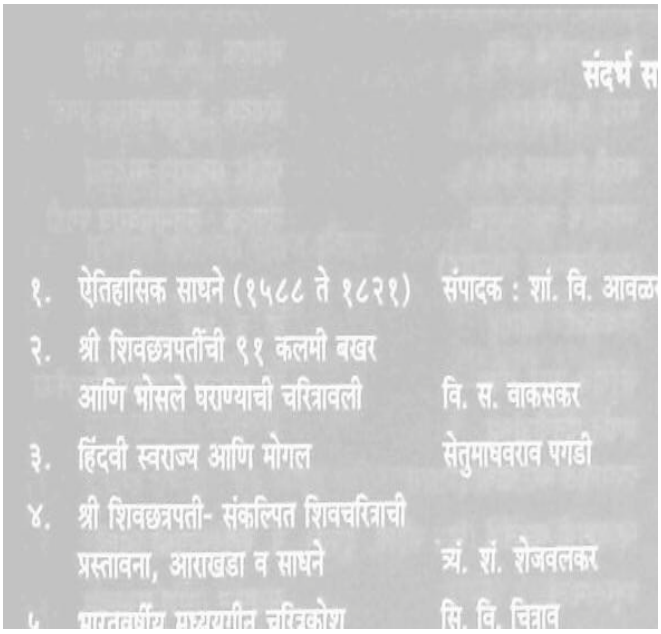
Figure 8: Input image for proposed system



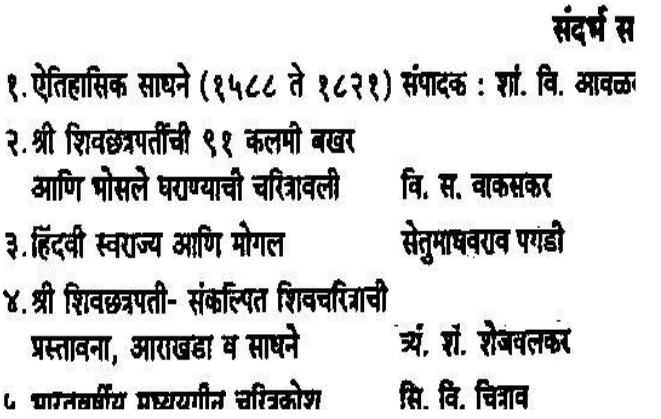
(a) Contrast Image



(b) Grey scale method edge detection.



(c) Binary Image.



(d) Final Output image.

5. Conclusion

Here we come to conclude that the proposed method is simple binarization method, which produces more clear output. It can be work on many degraded images. This technique uses contrast enhancement along with threshold estimation. We introduced new module post processing which will remove the background degradations found in the binarized image. In this technique we are going to used grey scale method to create outlined map around the text. The output of this system produces separated foreground text from collided background degradation. For that we have maintain the contrast level at min and max level. Which will help to make more clear and readable output.

References

[1] Wagdy, M., Ibrahima Faye, and DayangRohaya. "Fast and efficient document image clean up and binarization based on retinex theory."Signal processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on.IEEE, 2013.
[2] Sehad, Abdenour, et al. "Ancient degraded document image binarization based on texture features." Image and Signal Processing and Analysis (ISPA), 2013 8th International symposium on.IEEE, 2013.
[3] Su, Bolan, S hijian Lu, and Chew Lim Tan. "Robust document image binarization technique for degraded document images."Image Processing, IEEE Transactions on 22.4 (2013): 1408-1417.
[4] Su, Bolan, Shijian Lu, and Chew Lim Tan. "Combination of document image binarization techniques."Document Analysis and Recognition (ICDAR), 2011 International Conference on.IEEE, 2011.
[5] Gaceb, Djamel, Frank Lebourgeois, and Jean Duong. "Adaptative Smart-Binarization Method: For Images of Business Documents." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on .IEEE, 2013
[6] N. Otsu, “A threshold selection method from gray level histogram,” IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.
[7] Brensen, 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.
[8] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

Author Profile

Yogita Kakad received the B.E. degree in Computer Engineering from HVPM College of Engineering & Technology, Amravati in 2009. Pursuing M. E. in Computer Engineering from MGM College of Engineering and Technology, Kamothe, Navi Mumbai.

Savita Bhosale received M.E degree and Pursuing her PhD in Electronics Engineering from Dr. Babasaheb Ambedkar Technical University, Raigad and working as Assistant Professor at MGM College of engineering and Technology, Kamothe Navi Mumbai.