International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Comprehensive Survey on Binarization of Degraded Document Images

Aparna Patil[1], Deepak Gupta[2]

[1]Pune University, Department of Computer Engineering, Siddhant College of Engineering, Pune, Maharashtra, India

[2]Professor, Pune University, Department of Computer Engineering, Siddhant College of Engineering, Pune, Maharashtra, India

Abstract: ***In this digital world of technology, we are interconnected to each other via a soft and strong internet medium. Our Entire data, being in a digital world, is available in the form of soft copies of documents. With this, we can update, store, backup and preserve the soft copies of our documents. This is the case with the latest data, but going towards our old traditional data, which is available only on hard copies of the paper, we come across a lot of problems while preserving such degraded copies of data. Many times the old and ancient traditional documents play a vital role in our day to day life. Most of the papers containing our data get degraded due to lack of attention and improper handling and preservation. Most commonly seen degradation of such papers is interference of the text written on the front and back of the papers. In order to make this interfered front end data separate from rear page data many researchers have been studying and proposing binarization of the degraded document methodology. Here we study and analyze various binarization techniques proposed previously and then propose the new and innovative technique for the same. We create the binarized image of the degraded image through some intermediate steps. Ultimately, the binarized image will be next processed by the post processing module. The final output of entire process will generate a clear and binarized image with foreground text clearly seen without interference.***

Keywords: Image contrast, document degradation, adaptive binarization, document image processing, degraded document binarization, pixel classification.

1. Introduction

In this digital world, various image and document processing techniques emerged in a wider scope for data extraction or text extraction. The images are widely used in various domains of the researches such as geography, tomography, etc. Most of the novels written few of the years ago on the papers are of utmost use in our day to day life, but due to improper maintenance of such novels, the data is degraded and becomes unreadable for users and thus leads to loss of useful data. Such images become degraded after a particular span of time, and we can‟t use them in spite of them being very useful for us. Sometimes some documents get degraded due to low quality papers or ink used to type or write on the papers, thus making such useful image of no use for further use.

The degraded document images either scanned or captured are in the form unreadable text in foreground format. We need to differentiate between the foreground and background text. The techniques for image binarization are therefore emerged as useful ways for obtaining text from degraded documents. The degraded images are then passed through various intermediate methods which will produce the output image in a foreground text readable format. This survey will first analyze various techniques and then make compare the existing techniques to the proposed one. Although document or image binarization issue still prevails, threshold considerations of degraded and interfered document images have been resolved. It‟s because of the the high inter/intra-variation between the foreground text stroke and the unnecessary document background across different documents and images.

2. Literature Survey

Image binarization being a very interesting domain for most of the researchers. As observed by previous studies, lots of documents that are degraded are missing a clear distribution pattern and thus are classified as badly degraded. Considering the threshold alone is not at all sufficient method for the degraded document image binarization. Adaptive local threshold calculation, which computes a threshold for each local document pixel, is analytically the best mechanism to handle the variations in degraded images of the documents.

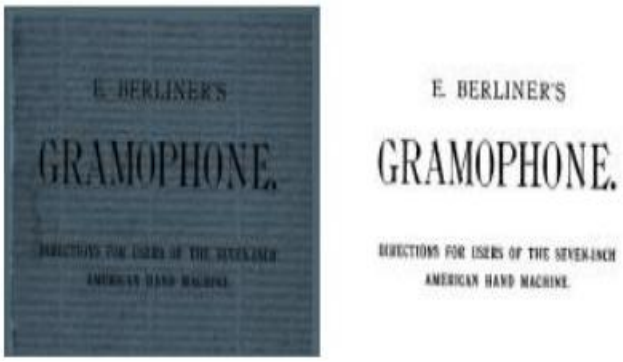
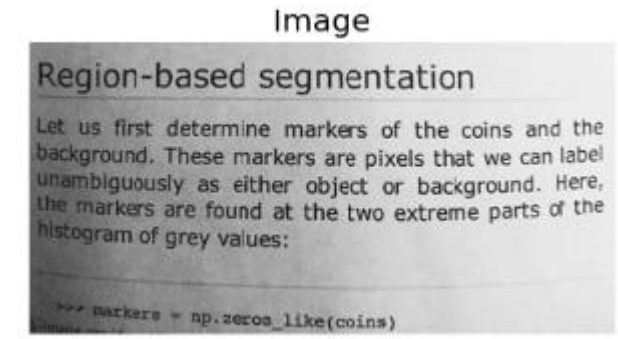


Figure 1: Sample binarization

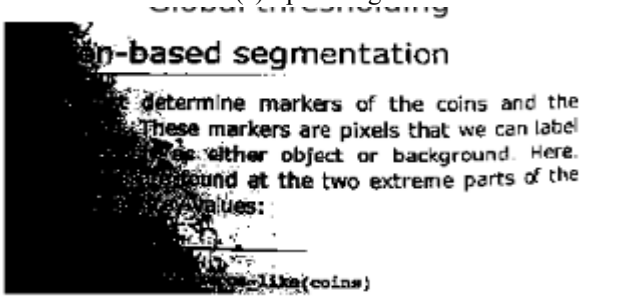
A. The Global Thresholding Mechanism

Global threshold technique estimates and computes overall value of a threshold for complete image; this method essentially needs some mathematical computations and can efficiently work in normal cases. But this method does not work properly if the image comprises of some differentially illuminated backgrounds, such as improperly distributed colour and unstructured illuminated backgrounds. Thus this

global or overall threshold methodology is inefficient for degraded images of the documents, as they fail to have a clear bimodal distributed pattern that distinguishes the foreground text and the background [4].



(a)Input Image



(b)Output Image

Figure 2: Input image and Output Image of global thresholding.

B. Image binarization using texture features:

The texture based image binarization is used for majorly historical documents that have been degraded over the time, which is based on texture features of images. This texture based methodology is a unique and dynamic one. The recent pixel is processed focused around the framework of co-event by the scriptor. The texture based technique is tested on the basis of few objects, by making use of DIBCO dataset degraded and badly illuminated color documents and it used a set of old degraded document gave by a library [1].

Figure 3: Input image and output of the texture feature

C. Adaptive binarization for degraded images:

The adaptive binarization technique uses erosion and dilation process by preparing light and dark scale picture; thus producing the image as output that reduces the shadow intensity and noise levels in the output. Ultimately, this approach is the uniform combination of contrast variation and contrast technique. This adaptive technique of binarization combined the workings of two systems that highly improved Niblack and also thresholding (neighbourhood thresholding) by making use of small amount of neighbouring pixels which changed and made effect on the average value of the areas [2].

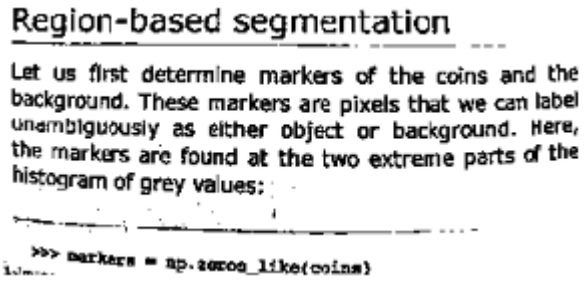
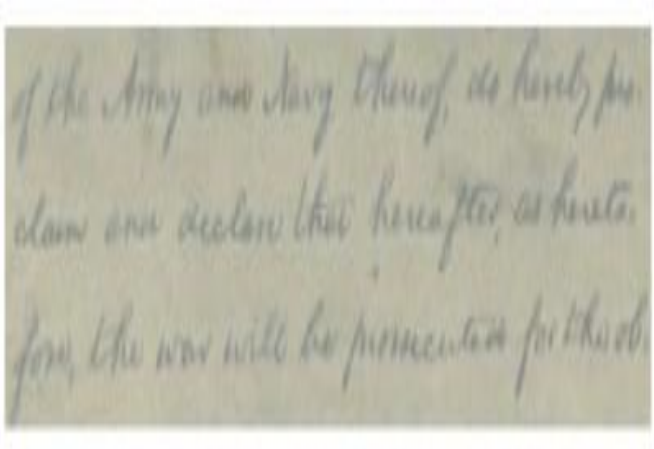


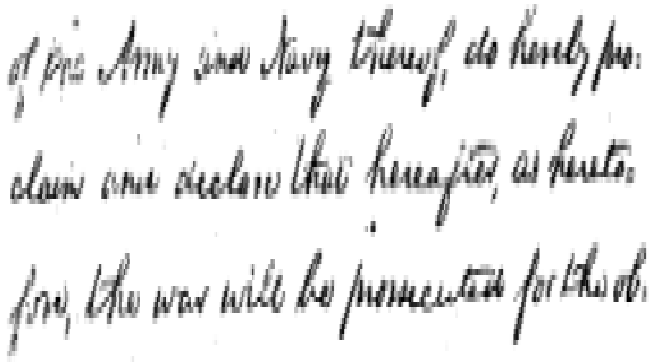
Figure 4: Input fig.2 (a) and Output image for adaptive thresholding.

D. Combinatory Document Image Binarization Techniques

A hierarchical newly generated structure is formed by thresholding techniques and obtains the optimized working performance for binarization of document image is described. The mentioned outputs of this technique divides all the image pixels into three different sets of pixels, namely, uncertain pixels, background pixels and first foreground pixels [3].



(a) Input Image



(b) Output Image

Figure 5: Results of the combinatory binarization techniques.

E. Dynamic Threshold Binarization

The techniques proposed for binarization like iteration technique, compute the threshold of current pixel by comparing with its own grey level values and neighboring pixels, and then the pixel coordinates. Such binarization

methodology is widely used for images with very bad and degraded quality, especially the images having histogram constructed from single peak values. But still, because of calculation of the threshold dynamically, this technique comprises of a very high calculation and thresholding complexity and immensely slow execution speed [5].

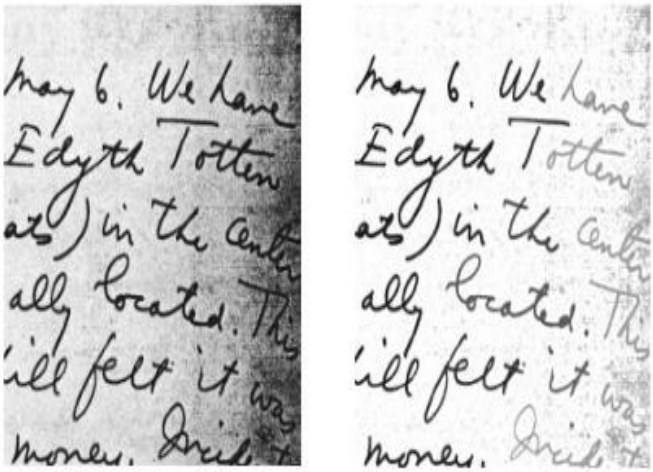


Figure 6: Dynamic binarization using dynamic threshold.

F. Otsu Method

A very well-known technique of binarization was proposed by OTSU using global thresholding. This binarization method invented the „T" threshold which splits the histogram used in gray level pixels into two different segments. The calculation and estimation of classes (inter or intra) variances is based on the final histogram of the image that is normalized H = [h0,h1,h2,h3.....h254,h255] where {hi=1}.This technique proposed by OTSU is applicable to generally execute all clustering based thresholding for images. In Otsu technique we check for the threshold that reduces the intra-class variation as a weighed sum of variances of the two different classes.

$$\sigma^2 prb(t) = prb_1(t)\sigma_1^2 + prb2(t)\sigma^2 2(t)$$

Where prb, are the possible probabilities of the two different classes which are divided by the threshold value and the variances of both classes. The class existence (possibility) probability and class mean value can be computed in an iterative manner [5].

G. Brensen Method

This is an adaptive method which decides different threshold for single pixel, considering every pixel at coordinate(x, y) in the image, this adaptive threshold is computed by two research parameters namely, Z_low and Z_high, which are the min and the max gray levels in the squared window w*w which is centered more than a at pixel (x, y).

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

If the pixels has the distinct quantity that is too less as compared to the threshold 1, then the neighboring pixels comprise of single class: which is either background or text [6].

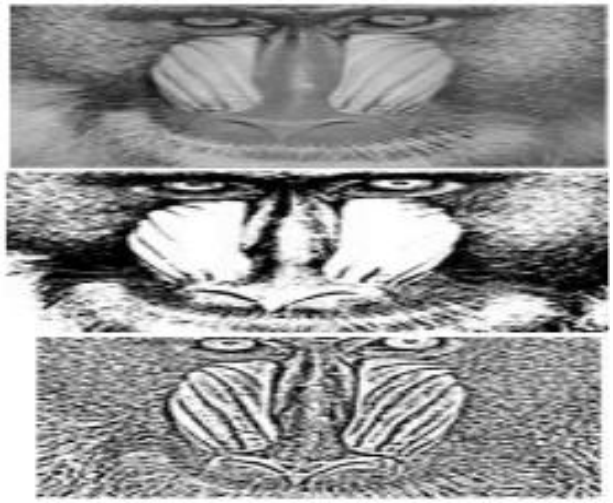


Figure 7: Applying bresnen.

H. Niblack method

The Niblack method defines a confined and suitable threshold for each and every pixel by decreasing the rectangular window over the complete image. The estimation of the precise threshold depends on confined mean value „m" of all pixels in the rectangular window which is represented as: The threshold „T", which is deliberate by making use of mean value „m" and the standard deviation „σ" of all the pixels in the rectangular window.

$$Tniblack = m + k * s$$

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

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

Hence the mean threshold „T" is obtained by: „T= m+ k * σ". Where the k is a parameter used in finding out the count of the edge pixels which are measured as object pixels and thereby takes a negative value. Major benefit of the niblack method is it properly recognizes the text regions but creates a lot of noisy binarization data for non-textual regions of the background. [7].

3. Proposed System

After studying and analyzing of the existing methods above, it can be inferred that the previously proposed techniques have some limitations which we need to overcome. To overcome these limitations the new proposed system uses new binarization technique along with grey scale method. The working four modules in the proposed system are:

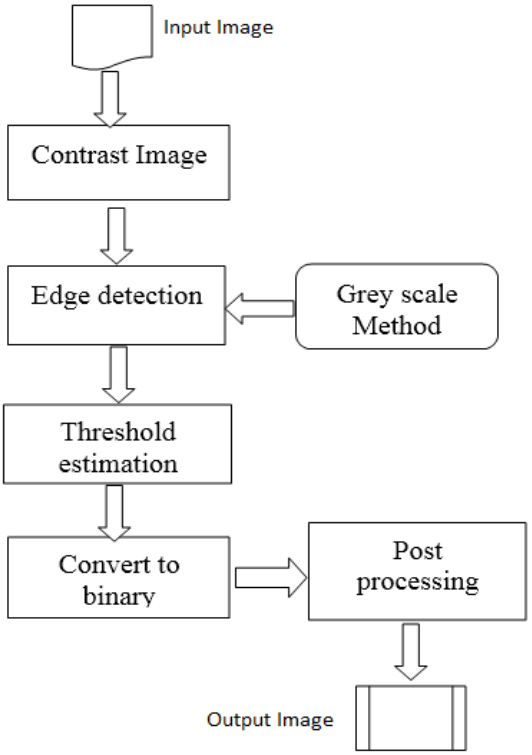


Figure 7: Proposed system Architecture

To identify the precise foreground text strokes, it is very essential to properly adjust the levels of the contrast for the image. Within this module, the proposed system prefers to keep the image contrast considering the min or max level. Which eventually depends on what is the amount of foreground text getting mixed up with background text i.e. noise. In Contrast construction module, the proposed system inverts the current values of the pixels between 0 to 255 for each component of RGB i.e. here the color values of pixels get reversed. The contrast converted image is further processed for test stroke edge detection procedure. This further, will produce the outline of the foreground text pixels. These bordered pixels and pixels outside the border are then divided into two categories. Out of which, the first category is connected pixels and the second one is non-connected pixels. Connected pixels determine the area to be taken under text stroke for foreground text.

This image, after edge detection, is then processed under binarization method, i.e. converted to binary format of 0‟s and 1‟s where the pixels with „0‟ implies that the image pixels are background pixels and „1‟ implies that the image pixels are foreground pixel. The pixels with „0‟s are removed from the processing image, being the part of the background pixels. The binarization method output image creates bifurcation of image into two layers. The post processing further aids in eradicating the unwanted noisy pixels still seen on the binarized image.

4. Analysis

The output generated by the previously existing systems does not gives the expected efficiency and thus need to be worked on for achieving the required accuracy and efficiency of upto 95%. The existing systems provide efficiency from 84% to 92% as mentioned in [3]. The proposed system is expected to provide more efficiency than the existing systems. The proposed system comprises of an additional gray scaling technique instead of canny‟s edge detection that aids in increasing the efficiency and decreasing the system complexity.

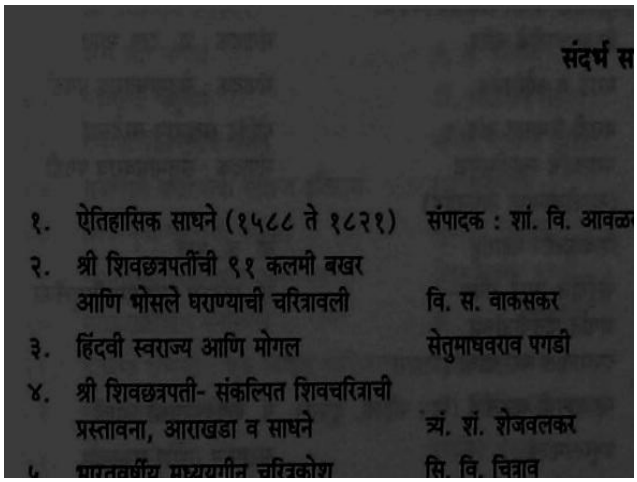


Figure 8: Input image for the proposed system

5. Conclusion

Thus the study and analysis of all the major existing systems for binarization techniques conclude that the previous techniques fail to achieve the efficiency due to detection of background or non-textual regions on the binarized output image. The proposed method is therefore is simple, robust, as accepts any type of image as input, binarization method that process and binarize the image properly and obtain the clear text from the foreground text. The output of the proposed system produces great difference between foreground recovered texts from the interfered background text or watermarks. The parameters such as MSE, SNR and PSNR that will be computed for the proposed system will determine the efficiency of the proposed system.

References

[1] 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.
[2] 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 (2014): 1408-1417.
[3] 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.
[4] 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.

[5] N. Otsu, "A threshold selection method from gray level histogram," IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.
[6] 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.
[7] W. Niblack, *an Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

Author Profile

Aparna Patil received the B.E. and Pursuing M.E. degrees in Computer Engineering from Siddhant College of Engineering, Pune. In academic year 2015-16.

Prof. Deepak Gupta received the B.E and M.Tech degrees in Information Technology from S.A.T.I., Vidisha, Madhya Pradesh, in 2002 and 2007, respectively. Now he is working as assistant Professor in Siddhant College of Engineering, Pune.