International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

# Highly Optimized and Robust Binarization Technique for Degraded Document Image

**Shobhika T. Ingle, S. P. Bhosale**

Electronics & Telecom Department, AISSMS COE Pune, Maharashtra, India

**Abstract**: *Binarization technique for degraded document is widely used technology all over the world. Various binarization techniques are used in practices but no technique used till now is full proofs which work for all kind of degraded document. Due to the high inter intra variation between the document background and the foreground segmentation of the text from degraded document is very challenging task. In the proposed method, first of all we convert the text that is input image to black and white. And then further apply thresholdinding algorithm to the resulted document. Post processing of the resulted document is done using sobel method for edge detection and ostus algorithm for every window to achieve proper weight so that high intravarationion of the document can be detected. And then, morphological median filtering is introduced to eliminate the salt pepper noise. One more addition to the proposed method is that the OCR is added so that text is easily recognized. The experiment results show that the proposed method runs quickly, accurately and fits for all kinds of degraded document.*

**Keywords:** Window thresholding, Classification of pixel, Sobel Binarization of document, Processing of degraded document.

## 1. Introduction

Various thresholding techniques are used for separating text from the background ,but there is not any fixed technique which will work for all the documents which are severely degraded reasearch is still going on in which ink at back side of the document reflect on front side. Binarization of image has been studied for many years but it is stilled an unsolved issue due to lot of variation between strokes and background.

Degradation of historical document results in imaging artifacts. This degradation induce thresholding error and makes the binarization a challenging task. In the present paper, first the separation of R, G, B from the image is done. And the grayscale image is obtained to reduce the complexity of the algorithm. Also the size of the algorithm is reduced. Secondly the obtained image is divided into number if window having some fixed size(rows and column of pixel) After that for each window Ostus is applied and fixed weight is obtained for each window. This is done because image can consist of large variation and applying Ostus alone to the whole image can missed some area having large variation andText can get missed due to single weight or threshold is obtained for the whole image if ostus is used.

## 2. Motivation

The text localization due to document digitization process is the core motivation for the proposed project evaluation. The location of the correct boundaries between two different regions is a subjective and been research by various authors. Also ground truth data contained boundary localization errors. Due to this a historical document fade with character and thus we motivate to choose the area of document binarization mainly for historical document.

The proposed method is simple and required minimum parameter tuning.

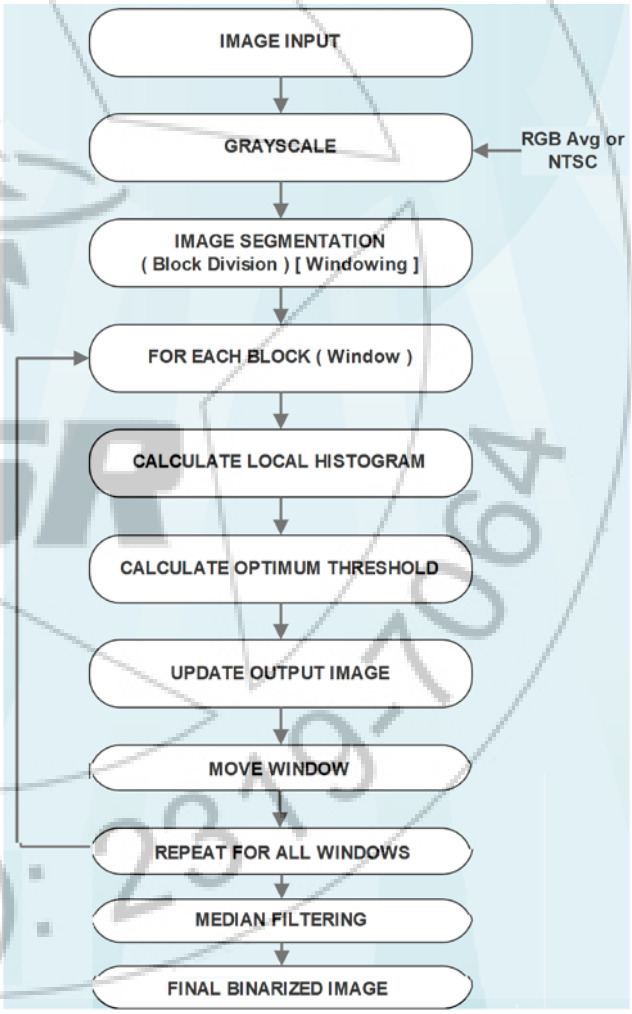


**Figure 1:** Project flow graph

For detecting strokes Sobel algorithm is applied. And lastly median filter is used to remove the salt paper noise (those pixel which are not the text but still present in the resulted document)

## 3. Related Work

To convert the image into its binary format, many thresholding techniques are used in practices. As many degraded documents, lot of variation in image pattern, global thresholding cannot be a better approach for the degraded document binarization and thus adaptive thresholding is better approach. Generally Otsu's thresholding is used as global thresholding. But cannot be used for image having large number of variation, but if windowing is used and then for each window Ostus is applied then the thresholding can be enhanced and new array can be evaluated for further processing to get a clear segmentation of text from the background. Other approaches have also been reported, including background subtraction texture analysis, recursive method decomposition method, contour completion Markov Random Field, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance and combination of binarization techniques But these method are often complex for analyses.

## 4. Proposed Method

This section describes the proposed document image binarization techniques. For a given a degraded document image,R,G,B colours are separated from a given coloured image and then each color is ANDed with 0xff to obtained the 8 bit binary value of each colour(R,G,B).After separating each colour gray scaling which is 8 bit binary value is obtained. Thresholding is applied on the gray scale image. In this paper basically window based thresholding is applied and then Ostu's is applied over which window to obtain. The threshold value for each window. This is done in the preprocessing stage. Edges are then detected using sobel algorithm to the threshold image. And finally filtering is done using median filter. Proposed method is simple and requires minimum parameter tuning. To enhance the quality performance of the technique OCR will be added at the last stage.

**Figure 2:** Degraded input image

### A. Separation of R, G, B and Gray scaling

After giving the degraded input image, separation of Colour R, G, B is done to make the 24 bit image into 8bit image this is done in order to reduce the complexity of Algorithm.

B= current pixel at with location ANDed with 0XFF
G= current pixel at with location shifted right by 8bit Position and then ANDed with 0XFF
R= current pixel at with location shifted right by 16bit position and then ANDed with 0XFF Now every bit changes to 8 bit and thus grey scaling will is:

Formula for Grey Scaling is (GS) = (R+G+B)/3.

Thresholding will be applied to gray scale image value i.e. only two values will be generated either Black Or white I.e. gray scale value >threshold then the pixel will turned to white & If gray scale value <threshold then the pixel will turned black

### B. Window function & Ostu''s

After getting the gray scale image window function is applied to the grayscale image. For image blur Window can be of size 3 by 3, 5 by 5or 9 by 9 Less will be the window size less blur vice versa Windows width and height (means how many pixels in X and how many pixels in Y)

Eg- (100*100)*(window size) 100 * 100 * (3*3)
W * H * (size of the window)

While traversing through each window Ostu's will be applied to each window so that threshold is obtained for each window. This is done because the image has large variation. Ostu's alone cannot be used for the complete image as it gives the single global threshold value, so if the image has variation information will be lost.

### C. Edge Detection

To detect the edges of the image sobel algorithm is used. It has standard matrix for horizontal and vertical edges can be named Hx and Hy, where Hx is used to find horizontal edges and Hy is used to find vertical edges of the text.

Matrix Hx

| | | |
|---|---|---|
| -1 | 0 | 1 |
| -2 | 0 | 2 |
| -1 | 0 | 1 |

Matrix Hy

| | | |
|---|---|---|
| -1 | -2 | -1 |
| 0 | 0 | 0 |
| 1 | 2 | 1 |

### D. Filtering

After creating a foreground pixel map, some morphological post processing operations such as erosion, dilation and closing are performed to reduce the effects of noise and. enhance the detected regions. The noise is also called salt Paper noise and is removed by using median filter. While traversing through the text in the window if neighboring pixel does not have any overlap edge of the text then it will be treated as noise and converted to white(i.e. 1) So after applying the filter we will get the text in pure form.

## 5. Application

1) Finger print recognition
2) MRI images
3) Licence plate detection
4) Arabic and historical document recognition
5) Palm recognition

## 6. Conclusion

The proposed approach will recognised the scanned image or web image. The system will recognise the scanner image that contain degraded document image with different shade font and strokes. The system will successfully recognise the degraded document image. We will try to recognised highly degraded font, shaded, and strokes.

## References

[1] Robust Document Image Binarization Technique for Degraded Document Images Bolan Su, Shijian Lu, and Chew Lim Tan,Senior Member, IEEE IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013
[2] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.
[3] Adaptive document image binarization J. Sauvola*, M. Pietika Kinen Machine Vision and Media Processing Group, Infotech Oulu, University of Oulu, P.O. BOX 4500, FIN-90401 Oulu, Finland Received 29 April 1998; accepted 21 January 1999
[4] Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral ImagesFaisal Shafait a Daniel Keysers, Thomas M. Breuel Image Understanding and Pattern Recognition (IUPR) Research Group German Research Center for Artificial Intelligence (DFKI) GmbH
[5] A Comparison of Binarization Methods for Historical Archive Documents J. He, Q. D. M. Do*, A. C. Downton and J. H. Kim* Department of Electronic Systems Engineering, University of Essex, UK *Division of Computer Science, KAIST, Kusung-Dong, Yusung-Gu, Daejon, Korea. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)
[6] 1520-5263/05 $20.00 © 2005 IEEE
[7] Automatic Thresholding of Gray-level Using Multi-stage Approach Sue Wu, Adnan Amin School of Computer Science and Engineering University of New South Wales Sydney, 2052, Australia
[8] A New Mixed binarization Method used in real time Application of Automatic Business Document and Postal mail sorting , The International Arab Journal of Information Technology vol.10.No.2