

# Use of Adaptive Methods to Improve Degraded Document Images

**Rakesh Kumar Sethi**

*Roll No. 110CS0151*

*8th Semester, B.TECH( COMPUTER SCIENCE & ENGG.)*

**Department of Computer Science & Engg**

**NIT Rourkela,769008**

*UNDER THE SUPERVISION OF*

**Professor Ramesh Kumar Mohapatra**

**Department of Computer Science & Engg**

**NIT Rourkela,769008**

**India**

*For the completion of B.Tech Final Year Project*

*During the Period*

*July,2013-April,2014*



**Department of Computer Science and Engineering**  
**National Institute of Technology Rourkela**  
**Rourkela – 769 008, India**

*Professor Ramesh Kumar Mohapatra*  
*Department of Computer Science & Engineering*  
*NIT Rourkela, India*

## **CERTIFICATE**

This is to certify that thesis entitled **Use of Adaptive Methods to Improve Degraded Document Images**” has been completed by *Rakesh Kumar Sethi, Roll No. 110CS0151, National Institute of Technology, Rourkela, India* ,during the period July,2013-April,2014 for the Final Year Project 2013-14 under the supervision of Prof. Ramesh Kumar Mahapatra.

*( Prof. Ramesh Kumar Mohapatra )*

## **ACKNOWLEDGEMENT**

I am grateful to Professor Ramesh Kumar Mohapatra, Department of Computer Science and Engineering, National Institute of Technology, Rourkela who has accepted me to guide me during my Final year project at NIT Rourkela. I would also like to thank Ph.D Scholar Ravi Barpanda for his guidance during my project and thesis work. I am also thankful to my Head of the Department and other faculty members for their help and support in the project.

*( Rakesh Kumar Sethi)*

## **DECLARATION**

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources without citations. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

*Rakesh Kumar Sethi*

## **Abstract**

An adaptive method for enhancing and binarizing the images of degraded documents has been presented here. This method does not need any feature handling by the user and it handles all kinds of degradations, removes noise, ensures connectivity of stroke and improves low-contrast. The project briefly includes following steps: a pre-processing procedure using a low-pass Wiener filter to produce a smoothed image, an approximate estimation of foreground regions, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image followed by a post-processing step which is carried out to enhance the quality of foreground regions and preserve line connectivity of texts.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background and Motivation . . . . .	2
1.2	Literature Survey . . . . .	3
1.2.1	Image Segmentation . . . . .	3
1.2.2	Thresholding . . . . .	4
1.2.3	Otsu's Method . . . . .	5
1.2.4	Niblack's Method . . . . .	6
1.2.5	Sauvola's Method . . . . .	7
<b>2</b>	<b>Theory of the adaptive method to read degraded document</b>	<b>8</b>
2.1	Introduction . . . . .	8
2.2	Pre-Processing . . . . .	9
2.3	Approximate Estimation of Foreground Regions . . . . .	10
2.4	Background Surface Estimation . . . . .	11
2.5	Final Thresholding . . . . .	12
2.6	Post-Processing . . . . .	13
2.6.1	Shrink Filter . . . . .	13
2.6.2	Swell Filter . . . . .	14
2.7	Proposed Algorithm . . . . .	15

<b>3</b>	<b>Experimental Evaluation</b>	<b>16</b>
3.1	Implementation . . . . .	16
3.2	Experimental Setup . . . . .	16
3.3	Results . . . . .	17
<b>4</b>	<b>Conclusion</b>	<b>18</b>

# Chapter 1

## Introduction

### 1.1 Background and Motivation

In order to extract information from the degraded documents which have changed over generations, it is important that the pictures of such documents need to be processed carefully so that the unclear content of the degraded document becomes clearer and hence easily recognizable by the optical character recognition system. Historical manuscripts, calendars, old newspapers and damaged books carry text which are partially clear and blurringly visible and hence it is necessary that methods should be implemented to make the text readable. Previously, many attempts had been taken to improve the image of such documents by using different kinds of methods of binarization involving use of either global or local threshold. These methods do take an initiative towards improving the visibility of text on the degraded document, but not enough effectively to be easily recognizable by the OCR. Hence, this adaptive method is introduced which not only binarizes the document image but also fills up the broken spaces in characters and removes noise from the image. The image undergoes pre-processing, intermediate improvement steps and is followed by post processing.



This thesis concentrates on the steps involved in the adaptive method. It describes in details how each step is carried out and the significance of each step. And it concludes by making comparison between different methods of document binarization and this current method.

## 1.2 Literature Survey

Different methods of binarization are studied and applied to the images of degraded documents to obtain the binary image. These binary images are then applied to OCR to study the efficiency of each of the binarization techniques. In an OCR, one of the main processing stage is binarization of document images, i.e. separation of foreground from background. Binarization of a text image should give us, in an ideal case, the foreground text in black and noisy background in white. Though different thresholding methods already exist in literature, they don't give perfect results for all types of documents. Some algorithms might work better for one type of documents where there are marks of strain while they might give poor results for other types where there are extremely low intensity variations. Different binarization methods have been evaluated in for different types of documents presents an evaluation of eleven locally adaptive binarization methods for gray-scale images with low contrast, variable background intensity and noise. In that evaluation, Niblack's method was found to be the better of them all. Different improvements have since been made to the original Niblack method to improve the results. One of these include Sauvola's algorithm. The binarization algorithms that are applied here for comparing the results with that of the proposed method are Otsu's method, Niblack's Method and Sauvola's Method.

### 1.2.1 Image Segmentation

Segmentation is taken as the first step in image analysis. The purpose is to subdivide an image into meaningful regions which do not overlap, which can be used for further anal-

ysis. In general, autonomous segmentation is one of the most hardest problem in digital image processing. All the image segmentation methods assume that:

1. the intensity values are different in different regions, and,
2. within each region, which represents the corresponding object in a scene, the intensity values are similar.

For example, the following figure represents the concept as:

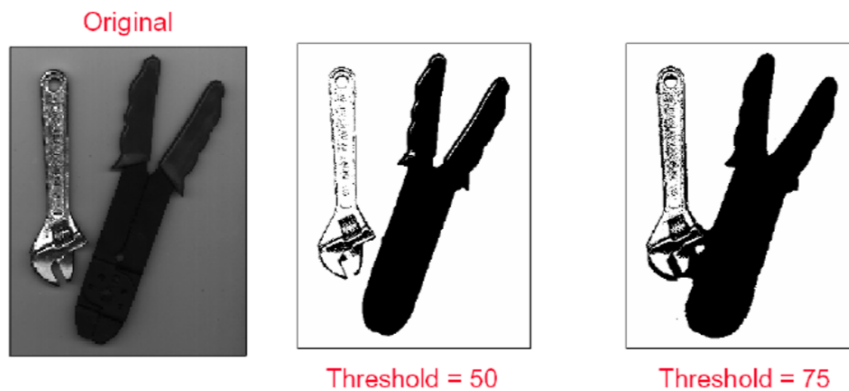


Figure 1.1: Images for different threshold values

## 1.2.2 Thresholding

Thresholding is used to extract an object from its background by assigning an intensity value  $T$ (threshold) for each pixel such that each pixel is either classified as an object point or a background point. In general,  $T = T[x, y, f(x, y), P(x, y)]$

If  $T$  is a function of  $f(x, y)$  only then it is called global thresholding

If  $T$  is a function of both  $f(x, y)$  and local properties  $p(x, y)$ , then it is called local thresholding

If  $T$  depends on the coordinates  $(x, y)$ , it is said to have dynamic/adaptive thresholding

Thresholds can be chosen in one of the following two ways:

### Fixed Threshold

In fixed (or global) thresholding, the threshold value is held constant throughout the image. A single threshold value is determined and each pixel is treated individually. The binarization is done using following formula:

$$b(x, y) = 0 \text{ if } f(x, y) < T$$

$$b(x, y) = 1 \text{ if } f(x, y) > T$$

It assumes that high-intensity pixels are of interest, and low intensity pixels are not of that much interest.

### Iterative Threshold

---

**Algorithm 1** :Adaptive Method

---

- 1: Select an initial estimate of the threshold  $T$ . A good initial value is the average intensity of the image.
- 2: Partition the image into two groups,  $R_1, R_2$ , using the threshold  $T$ .
- 3: Calculate the mean grey values  $R\mu_1$  and  $\mu_2$  of the partitions,  $R_1, R_2$ .
- 4: Select a new threshold:

$$T = \frac{\mu_1 + \mu_2}{2}$$

- 5: Repeat steps 2-4 until the mean values and in successive iterations do not change.
- 

### 1.2.3 Otsu's Method

This technique is based on a very simple idea that involves evaluating the threshold which minimizes the weighted variance within class.[2] This results in maximization of

the variance of between-class. It works directly on the grey level plot of histogram, so the method evaluates faster once the histogram is computed. Otsu's method is used to automatically perform image thresholding based on clustering. It reduces the gray-level picture to a binary picture. The algorithm assumes that the image to be thresholded contains two types of pixels or bi-modal histogram (i.e. foreground and background) then evaluates the optimal threshold distinguishing those two classes so that their combined intra-class variance is minimal.

### Algorithm

---

#### Algorithm 2 : Adaptive Method

---

- 1: Compute histogram and probabilities of each intensity level
  - 2: Set up initial  $\omega_i(0)$  and  $\mu_i(0)$
  - 3: Step through all possible thresholds  $t = 1 \dots$  maximum intensity
  - 4: Update  $\omega_i$  and  $\mu_i$
  - 5: Compute  $\sigma_b^2(t)$
  - 6: Desired threshold corresponds to the maximum  $\sigma_b^2(t)$
  - 7: Compute two maxima (and two corresponding thresholds).  $\sigma_{b1}^2(t)$  is the greater max and  $\sigma_{b2}^2(t)$  is the greater or equal maximum
  - 8: *Desired threshold* =  $\frac{\text{threshold}_1 + \text{threshold}_2}{2}$
- 

### 1.2.4 Niblack's Method

#### Algorithm

Niblack's method implements a sliding rectangular window over a grayscale image to calculate threshold for each pixel.[6] The threshold calculation involves evaluation of the local mean  $\mu$  and the standard deviation  $\delta$  of each of the pixel present inside the window and is determined by the following equation:

$$t_{niblack} = m + k * s \quad (1.1)$$

$$t_{niblack} = m + k \sqrt{\frac{1}{NP} \sum (p_i - m)^2} = m + k \sqrt{B} \quad (1.2)$$

where  $NP$  denotes the number of pixels present in the grayscale image,  $m$  is the average value of the value of the pixels  $p_i$ ,  $k = -0.2$ , by assumption. The benefit of Niblack is that it accurately identifies the text regions as foreground always but produces a huge amount of binarization noise in background regions which is the non-text region.

## 1.2.5 Sauvola's Method

### Algorithm

[7]Niblack's method has been claimed to be improved by the Sauvola's method by computing the threshold making use of the range of dynamism of image gray-value standard deviation,  $R$ :

$$T_{Sauvola} = m * (1 - k * (1 - \frac{s}{R})) \quad (1.3)$$

where  $k$  is set to 0.5 and  $R$  to 128. This method performs better than Niblack's algorithm in images where the text pixels have near 0 gray-value and the background pixels have near 255 gray-values. However, in images where the non-text pixels and gray values of text are almost equal to each other, the efficiency results decrease significantly.

## Chapter 2

# Theory of the adaptive method to read degraded document

### 2.1 Introduction

The steps involved in carrying out the adaptive method for improving quality of the images of the degraded documents.[5]It briefly consists of a pre-processing procedure using a low-pass Wiener filter, an approximate estimation of foreground regions, a background surface estimation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing that improves the quality of text regions and preserves connectivity among stroke.

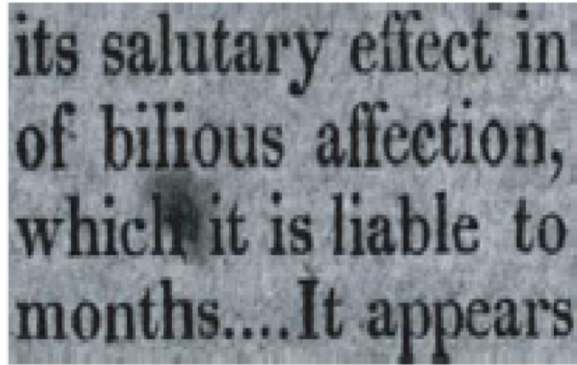


Figure 2.1: Image of a degraded document

## 2.2 Pre-Processing

For degraded document images, it is necessary to improve the visibility of the text on the image by carrying out steps that help in smoothening the background, improving the contrast between text and background and help in removing the noisy areas from the document image. To help an image go through such a step, it is necessary that the grayscale image  $I_o$  of the degraded document should be converted into an improved grayscale image  $I$ . This is done by:

$$I(x, y) = \mu + (\sigma^2 - \nu^2)(I_o(x, y) - \mu)/\sigma^2$$

where  $\mu$  is local mean,  $\sigma^2$  is the variance of pixels in  $3 \times 3$  neighborhood and  $\nu^2$  is the square of mean of pixels in  $3 \times 3$  neighborhood.

Carrying out this step is similar to applying a  $3 \times 3$  low pass Weiner filter to the image.

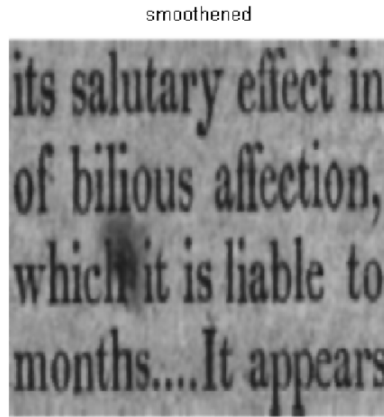


Figure 2.2: Result of applying the Wiener Filter to the image

## 2.3 Approximate Estimation of Foreground Regions

During this step, an approximate guess of text regions is obtained. This generates an initial segmentation of foreground and background areas which give a bigger set of all correct foreground points. This image is then improved at a later step. This particular method for binarization (Sauvola's approach) assumes the value for  $k = 0.2$ . [7] At this step, image  $I(x, y)$  is processed in order to get the binary image  $S(x, y)$ , where darker regions correspond to foreground regions.



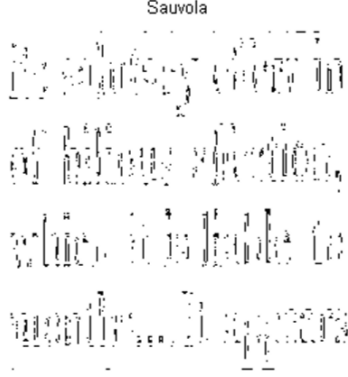


Figure 2.3: Result of applying Sauvola thresholding to the smoothed image

## 2.4 Background Surface Estimation

In this step, a closely rough background surface  $B(x, y)$  of the image  $I(x, y)$  is calculated. Background surface estimation is monitored by evaluating the  $S(x, y)$  image. For pixels that lie in the 0 regions of the  $S(x, y)$  image, the respective pixel value of  $B(x, y)$  equals to  $I(x, y)$ . For rest of the unattended pixels, the pixel point values of  $B(x, y)$  is calculated by neighboring pixel interpolation, as shown in the following equation://

$$B(x, y) = I(x, y) \quad \text{if } S(x, y) = 0$$

$$B(x, y) = \frac{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} I(ix, iy)(1 - S(ix, iy))}{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (1 - S(ix, iy))} \quad \text{if } S(x, y) = 1$$

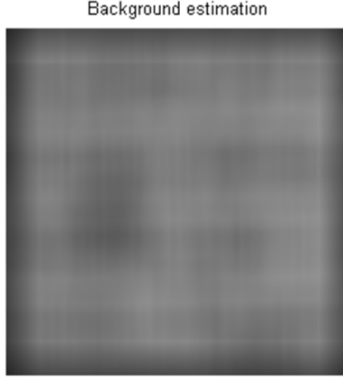


Figure 2.4: Estimated background surface of the input image

## 2.5 Final Thresholding

In this step, final thresholding is carried out by collaborating the estimated background surface  $B(x, y)$  with  $I(x, y)$ , pre-processed image. In this, we employ a formula where text are assumed to be present at those positions where the distance between the pre-processed image  $I(x, y)$  and calculated background  $B(x, y)$  is greater than a selected threshold  $d$ . This threshold is location dependent and changes as the background gray-scale level changes. For this reason, we propose a threshold  $d$  that has lower values for text containing regions. The final binary image  $T(x, y)$  is given by:

$$T(x, y) = 1 \quad \text{if } B(x, y) - I(x, y) > d(B(x, y))$$

$$T(x, y) = 0 \quad \text{otherwise}$$

In order to calculate the threshold for each portion of the image. We first need to find the average distance  $\delta$  between the foreground and the background using the following formula:

$$\delta = \frac{\sum_x \sum_y (B(x, y) - I(x, y))}{\sum_x \sum_y S(x, y)}$$

Then the threshold is decided using the equation here:

$$d(B(x, y)) = q\delta\left(\frac{1 - p_2}{\exp\left(\frac{1-4B(x,y)}{b(1-p_1)} + \frac{2(1+p_1)}{(1-p_1)}\right)} + \delta\right)$$

where previous researches have found that  $p_1 = 0.5$ ,  $p_2 = 0.8$  and  $q = 0.6$  are suitable for a degraded document processing.

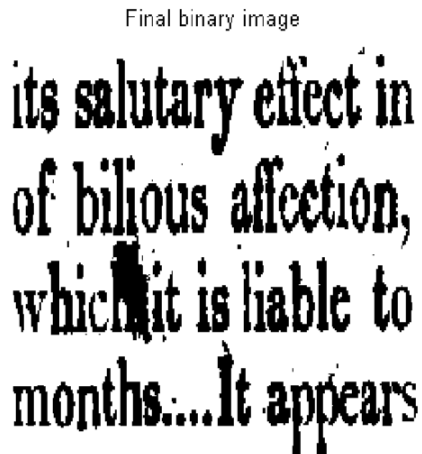


Figure 2.5: Final thresholding image

## 2.6 Post-Processing

[5] In the final step, we proceed to post-processing of the resulting binary image in order to eliminate noise, improve the quality of text regions and preserve stroke connectivity by isolated pixel removal and filling of possible breaks, gaps or holes. Below follows a detailed step-by-step description of the post-processing algorithm that consists of a successive application of shrink and swell filtering

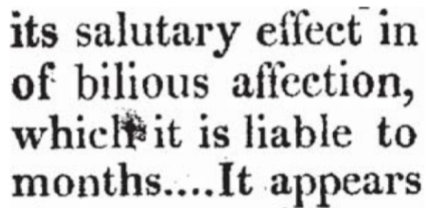
### 2.6.1 Shrink Filter

A shrink filter is used to remove noise from the background. The entire binary image is scanned and each foreground pixel is examined. If  $P_b$  is the number of background

pixels in a sliding  $n \times n$  window, which has the foreground pixel as the central pixel, then this pixel is changed to background if  $P_b$  is greater than  $k$  where  $k$  can be defined experimentally.

### 2.6.2 Swell Filter

A swell filter is implemented in order to fill visible breaks and gaps in the foreground. There are some text where the connectivity among the strokes have underwent discontinuity. In order to preserve this, the swell filter is implemented. The entire binary image is scanned and each background pixel is examined. If  $P_f$  is the number of foreground pixels in a sliding  $n \times n$  window, which has the background pixel  $(x, y)$  as the central pixel, then this background pixel is converted into foreground pixel if most of the surrounding pixels belong to foreground layer.



**its salutary effect in  
of bilious affection,  
which it is liable to  
months...It appears**

Figure 2.6: Post-processed image

## 2.7 Proposed Algorithm

---

**Algorithm 3** :Adaptive Method

---

- 1: Get the gray scale image of the degraded document image
  - 2: Pre-process the image using Wiener filter
  - 3: Apply Sauvola threshold to get rough estimation of foreground details
  - 4: Apply neighborhood interpolation to estimate the background
  - 5: Carry out final thresholding to get an improved image
  - 6: Post process the image by applying swell and shrink filter to remove noise and fill broken connectivity among strokes
-

# Chapter 3

## Experimental Evaluation

### 3.1 Implementation

The various methods to process the input image have been carried out using Matlab. This step involves taking the input image , processing it until an image readable by the OCR is obtained.The OCR system is utilised from the Wolfram Mathematica to obtain the string present in the image of the document provided.The Matlab Part of the experiment involves usage of Otsu ,Niblack and Adaptive method to generate the images which are then fed into the OCR for string recognition.

### 3.2 Experimental Setup

The image is fed into the different methods to get the output better image. The images obtained are better as compared to the degraded document image. This experiment involves comparing how good each of the image obtained is at getting detected by the OCR.

### 3.3 Results

The comparison results between the different methods have been briefly shown here as:

Method	Image	String Detected
Original:	its salutory effect in of bilious affection, which it is liable to months....It appears	-No text detected-
Otsu:	its salutory effect in of bilious affection, which it is liable to months....It appears	its salutory sffsct in sl* bi 'sus silsction. whim: tis liable tu months.-. .It _appeals
Sauvola:	its salutory effect in of bilious affection, which it is liable to months....It appears	"its salutory cllgzf in of bil'ous allcslion. whim: it is liable to months.-. .It -appears"
Niblack:	its salutory effect in of bilious affection, which it is liable to months....It appears	its salutory el'l'ect` in of bi 'ous alfcction. Whiz _t is liable to months.-. .It -appears
Adaptive:	its salutory effect in of bilious affection, which it is liable to months....It appears	its salutory effect in of bilious affection, whicliifit is liable to months .... It appears

Figure 3.1: Post-processed image

# Chapter 4

## Conclusion

The adaptive method for improving the images of degraded document provides better results as compared to the other binarization techniques. This method aims at improving the textual image by removing noise and making the unwanted gaps within the text to be filled and hence provide a complete readable image which is detected easily by the OCR. The binarization techniques only aim at converting the gray-scale, or sepia- kind of image into black and white which also converts the damaged portion of the image into black color and hence do not provide much facility to be detected easily by the character recognizer. Use of this adaptive method can always be helpful in extracting textual information from the degraded documents although the time complexity involved in this method is certainly higher than those of binarization techniques. But the kind of result that is obtained from the adaptive method is worth the time complexity involved.



# Bibliography

- [1] A. Rosenfeld, A.C. Kak, Digital Picture Processing, second ed., Academic Press, New York, 1982.
- [2] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Systems Man Cybernet. 9 (1) (1979) 6266.
- [3] J. Kittler, J. Illingworth, On threshold selection using clustering criteria, IEEE Trans. Systems Man Cybernet. 15 (1985) 652655.
- [4] P.K. Sahoo, S. Soltani, A.K.C. Wong, A survey of thresholding techniques, Comput. Vision, Graphics Image Processing 41 (2) (1988) 233260.
- [5] J. Yang, Y. Chen, W. Hsu, Adaptive thresholding algorithm and its hardware implementation, Pattern Recognition Lett. 15 (2) (1994) 141150.
- [6] W. Niblack, An Introduction to Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1986 pp. 115116.
- [7] J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition 33 (2000) 225236.