RESEARCH ARTICLE OPEN ACCESS

Content Restoration of Degraded Termite Bitten Document Images

Vishnupriya Raj*, C Arunkumar**
*(Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Coimbatore-641112)
** (Department of Information Technology,Amrita Vishwa Vidyapeetham, Coimbatore-641112)

ABSTRACT
Document images are often obtained by digitizing the paper documents like books or manuscripts. Due to degradation of paper quality, aging spreading of ink etc their appearance may get poor. Document may undergo termite bite due to aging and will get severely degraded .This work tries to find some solutions to increase the recognition rate of degraded characters after applying a best preprocessing technique for removing the noise due to termite bite. Main degradation for a degraded document are due to non-rectilinear camera positioning, blur, bad illumination, non-uniform backgrounds, non-flat paper surface, spreading of ink, document aging, extraneous marks, broken characters. There are different systems which are able to deal with degradations which occur due to these type of degradations. Restoration is highly useful in a variety of fields such as document recognition, historic document analysis etc. In this paper ,we propose a method to remove the noise due to termite bite . The work has the ability to deal with larger patch sizes and allows to deal with severe degradations.

Keywords- Restoration, Preprocessing, Thresholding, Niblack, Gabor Filter.

I. INTRODUCTION

Restoration plays a very important role in enhancing the degraded noisy images. Numerous algorithms have been designed to enhance the degraded image. The main reason for OCR failure is that the preliminary step of character segmentation is made difficult by nonuniform reduction of image contrast, nonuniform spacing of characters, presence of broken and touching characters, and presence of strong background noise.In this work special emphasis is given to ancient printed documents, where the degradation is mainly caused by the termite bite and the patches occurred due to the termite bite. Binarization and noise removal can be helpful in recognizing the text of a highly degraded document.Many algorithms are tested for binarization and noise removal.The best preprocessing method has to find which will work even in variable background intensity and the noise occured due to the termite bite. These requirements lead us to evaluate the best preprocessing method for a highly degraded termite bitten document.This paper is organized as follows. Section II presents the general architecture for the best preprocessing method for termite bitten document image. Section III presents the related works in this field. Section IV describes the algorithms used, and the corresponding results. Section V discusses about the future enhancements and section VI makes the conclusion.

II. GENERAL ARCHITECTURE

Different preprocessing methods have tested.The steps for the best preprocessing method for noise removal of termite bitten document image.

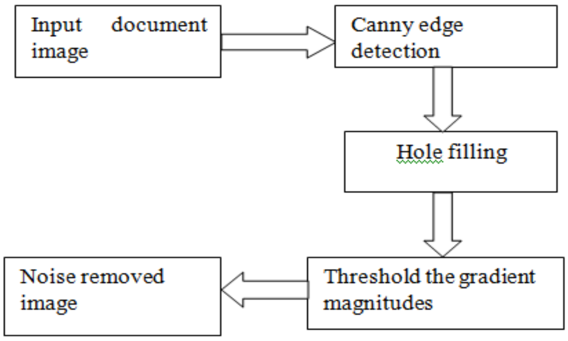


Fig 1: Basic architecture for best preprocessing method

III. RELATED WORKS

The works on the preprocessing of degraded document images can be dated back to the early works by

Paper [1] propose an algorithm which is based on the Markov Random Field for the binarization of document images degraded by uneven light distribution.For more robust binarization, they propose a soft decision algorithm based on the MAP-MRF framework which is widely used in signal processing area. The MAP-MRF problems can be solved by the combinatorial optimization if the corresponding energy is de- fined. In this paper, they formulate the energy using a new likelihood model and a generalized Potts prior model, and then solve the problem by the graph cut algorithm. In the graph cut, a binary graph G is defined as a set of nodes. They presented the results of experiments on two different types of images. The proposed algorithm is a soft decision method that do not need parameter control.As this method is based on optimization this will be time consuming.

Y. Chen in[2] a decompose algorithm, which recursively decomposes a document image into subregions until appropriate weighted values can be used to select an appropriate single-stage thresholding algorithm. The initial step of the algorithm tests whether the image has a bimodal histogram. After that decompose image into four equal size local regions.Then extract feature vectors from each local region and classify it.Smooth the edges of each region and finally threshold method is applied to each region. If the image is not bimodal it is recursively decomposed using quad-tree decomposition into smaller regions and a local threshold method with appropriate weighted values is applied to each different class region. This continues until the whole image has been decomposed and a threshold technique assigned to each region. It uses local feature vectors to analyse and find the best approach to threshold a local area.This method is suitable for historical handwritten documents.

In [3] starting with a single page of the document, wavelet-based decomposition and filtering is used to reduce the noise,Here the preprocessing of the image should make use of both filtering techniques for noise removal and deconvolution techniques for blur removal. The basic idea behind the wavelet transform is the multiple decomposition of a signal using banks of highpass and low-pass filters connected in cascade and spaced by decimation stages. For thresholding the high-frequency terms, [3] adopted a soft thresholding approach to discard terms that are lower than the noise power . This method is useful for preserving the boundaries of the characters that are usually associated with frequency terms higher than the noise power. The preprocessing explained in this paper has been tested on pages of the Opera Omnia by Girolamo Cardano.

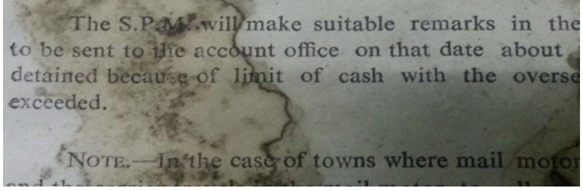
In [4] they propose a hybrid binarization approach for improving the quality of old documents using a combination of global and local thresholding. First,the entire document image is applied with global thresholding methods. If,background noise are detected again, the same technique is re-applied to each area separately. Image analysis systems use binarization as a standard procedure to convert a grey-scale image to binary form. An ideal binarization algorithm would be able to perfectly discriminate foreground from background, thus, removing any kind of noise that obstructs the legibility of the document image. In order to evaluate the proposed approach, they collected an amount of historical document images. In more detail, they formed a collection of 183 document images taken from Korgialenios Library of Cephalonia (KLC) and the Daratos Private Archive (DPA) of old documents. The parameters of the proposed algorithm, a medium size window (n=50) provides the most appropriate solution for detecting the areas with remaining noise.

Paper [5] describes a system which is able to separate the two regions of the document. De-noising step for the purpose of de-noising and a rough estimation of foreground region and background region. Binarization step is applied by computing an approximate background surface of an original document image.By combining the calculated background surface with the preprocessed original image ,thresholding is calculated by using a threshold parameter for predefined local window of specific size. For local adaptive thresholding using luminance value,the luminance value uses a global single-stage thresholding technique that finds the optimum threshold value for document images, using luminance value. After applying the thresholding technique, a manual thresholding approach is done to ensure good output quality. The main complexity here is in predicting the stroke direction and the disconnected edges.

IV. ALGORITHMS

Many algorithms are tested for binarization and noise removal. It is essential to find the best preprocessing method which will correctly gives all the information present, even in variable background intensity and the noise occured due to the termite bite. These requirements lead us to evaluate the best preprocessing method for a highly degraded termite bitten document.

Binarization as a standard procedure to convert a grey-scale image to binary form. An ideal binarization algorithm would be able to perfectly discriminate foreground from background, thus, removing any kind of noise that obstructs the legibility of the document image.Here for the termite bitten document we have applied different binarization techniques such as Savoula and Niblack .But it gave very poor output.



with reference to the entries in the Diary made b
offices from the time the overseer left his headquarter
cribed limit of cash to be conveyed by the overseer is

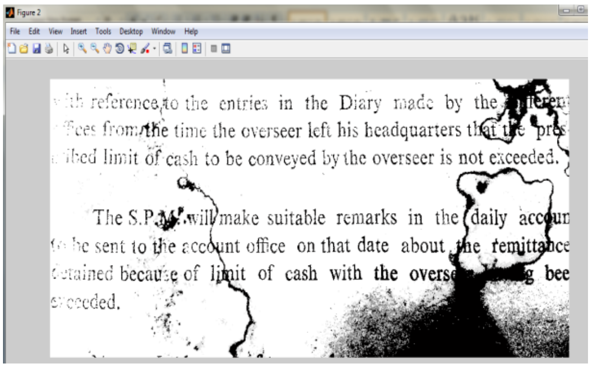
The S.P.M. will make suitable remarks in the
to be sent to the account office on that date about
detained because of limit of cash with the overse
exceeded.

NOTE.—In the case of towns where mail motor

Fig 2:Input image

A. GLOBAL THRESHOLDING METHOD

The pixels of the image are classified into text or background according to a global threshold. Usually, global thresholding methods are very simple and fast. In case if the background noise is unevenly distributed in the entire image ,they cannot be easily adapted.



with reference to the entries in the Diary made by the differen
offices from the time the overseer left his headquarters that the pres
cribed limit of cash to be conveyed by the overseer is not exceeded.
The S.P.M. will make suitable remarks in the daily accoun
to be sent to the account office on that date about the remittance
detained because of limit of cash with the overse bee
exceeded.

Fig 3:Result of Otsu's method

B. LOCAL THRESHOLDING METHODS

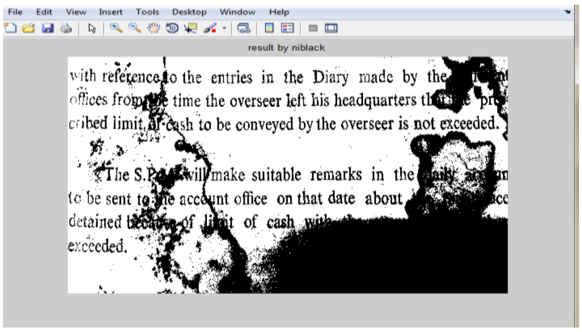
The pixels of the image are classified into text or background according to a local threshold determined by their neighboring pixels. Such methods are more adaptive and can deal with different kinds of noise existing in one image. On the other hand, they are significantly more time-consuming and computationally expensive.

Niblack

Local threshold for each pixel is calculated using Niblack method This method depends on the local mean value and the local standard deviation in the neighborhood of the pixel.

$$T(x,y) = m(x,y) + k * s(x,y)$$

where m(x, y) is the average of a local area and s(x, y) and standard deviation values, and k is used to adjust how much the total print object boundary is taken as a part of the given object.



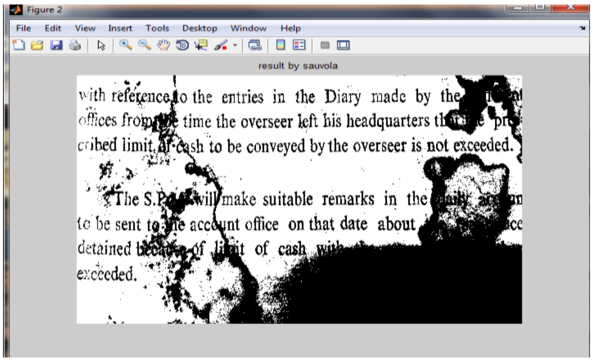
with reference to the entries in the Diary made by the
offices from the time the overseer left his headquarters t pr
cribed limit of cash to be conveyed by the overseer is not exceeded.
The S.P. will make suitable remarks in the
to be sent to the account office on that date about
detained because of limit of cash with
exceeded.

Fig 4:Result of Niblack method Savoula

A modification of Niblack algorithm which takes a grayscale image as input.
Threshold value is calculated using:

$$T = m(1 - k(1 - \sigma/R))$$

m is the mean
k is a constant between 0 and 1
R is the range of gray levels
R=128 and k=0.5



with reference to the entries in the Diary made by the
offices from the time the overseer left his headquarters t pr
cribed limit of cash to be conveyed by the overseer is not exceeded.
The S.P. will make suitable remarks in the
to be sent to the account office on that date about
detained because of limit of cash with
exceeded.

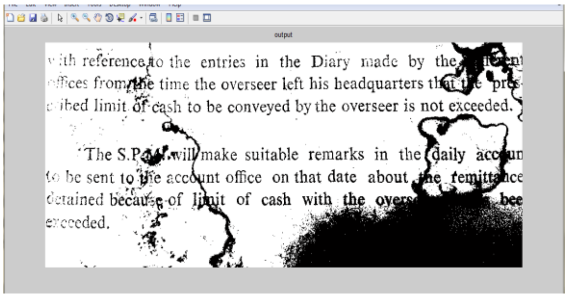
Fig 5:Result of Savoula method

C. PREPROCESSING PROCEDURE USING GAUSSIAN FILTER

Gaussian filtering is a linear convolution algorithm . It employs a convolution kernel that is Gaussian function is defined as follows:

$$g(x,y) = {}^{1}\!/_{\sigma\sqrt{2\Pi}}\, e^{-x^2+y^2/2\sigma^2}$$

where σ is the standard deviation of the Gaussian function. The Gaussian filtering method allows user to make fine changes to the amount of spatial averaging that occurs in the document image. Gaussian filtering with σ value gives about the degree of noise reduction.



with reference to the entries in the Diary made by the different
offices from the time the overseer left his headquarters that the pres
cribed limit of cash to be conveyed by the overseer is not exceeded.
The S.P.M. will make suitable remarks in the daily accoun
to be sent to the account office on that date about the remittance
detained because of limit of cash with the overs bee
exceeded.

Fig 6:Result after applying Gaussian filter

D. PREPROCESSING USING GABOR FILTER

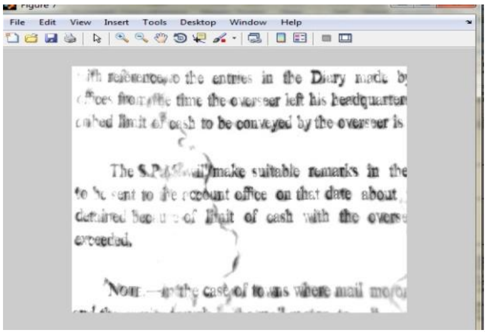
Gabor Transformation includes :

Orientation φ Frequency f Sigma (standard deviation of gaussian distribution)

The value of sigma involves a tradeoff Larger values which is more robust to noise but more likely to create spurious rings Smaller values which is produce less spurious rings but less effective in removing noise

1. Take the input image.
2. Apply the gabor filter.
3. Enhance the contrast of the image by transforming values of intensity image.

4. Enhance contrast of small regions of the document image.
5. Subtraction is done between input image and the contrasted image.
6. Find compliment of the result.

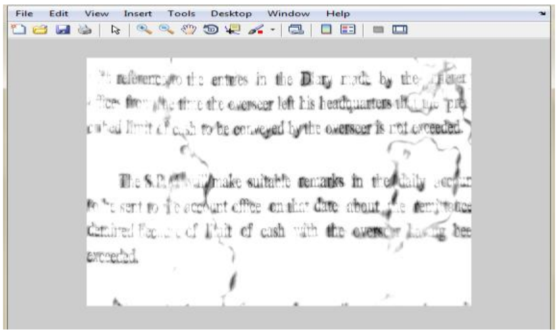


ith reference to the entries in the Diary made by
offices from the time the overseer left his headquarter
cribed limit of cash to be conveyed by the overseer is

The S.P.O. will make suitable remarks in the
to be sent to the account office on that date about
detained because of limit of cash with the overse
exceeded.

NOTE.—In the case of towns where mail mo

Fig 7:Result after applying Gabor filter $_{\theta}$=160^{0}



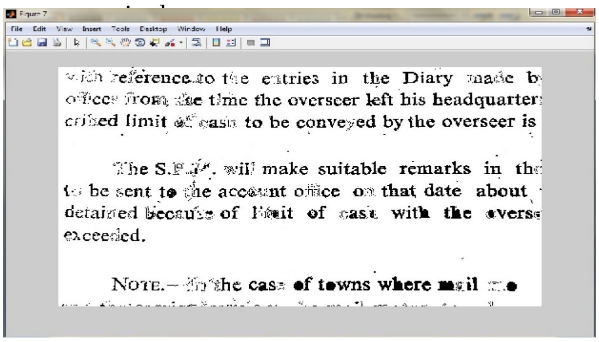
reference to the entries in the Diary made by the overseer
offices from the time the overseer left his headquarters till the pre-
cribed limit of cash to be conveyed by the overseer is not exceeded.

The S.P.O. will make suitable remarks in the daily accoun
to be sent to the account office on that date about the remittance
detained because of limit of cash with the overseer having bee
exceeded.

Fig 7:Result after applying Gabor filter $_{\theta}$=120^{0}

E.PROPOSED SYSTEM

1. Apply canny edge detection to detect the edges.
2. Apply hole filling after the edge detection.
3. Estimation of grey level gradient at a pixel is done.
4. Inorder to made the background uniform and to remove the termite bite stain
5. To identify the meaningful edges from the data, a simple approach is to threshold the gradient



with reference to the entries in the Diary made by
offices from the time the overseer left his headquarter
cribed limit of cash to be conveyed by the overseer is

The S.P.O. will make suitable remarks in the
to be sent to the account office on that date about
detained because of limit of cash with the overse
exceeded.

NOTE.—In the case of towns where mail mo

Fig 8:Result of proposed method

V. FUTURE WORKS

This work can be extended further to line and character process, texture extraction using template matching. It can be noted that there is missing of letters present in the result . So, for future work, the missed letters have to be restored and more enhanced output has to be developed.

VI. CONCLUSION

In the given degraded document image dataset several preprocessing algorithms have been applied and their results have been subjectively compared.Global thresholding method is not suitable for this type of document as it has varying background.Local thresholding method is suitable for varying background,but local adaptive thresholding approaches such as Niblack and Savoula donot gave good result for the dataset .Different filtering methods such as gabor filter ,wiener filter ,guassian filter are applied in the dataset.But it does not give good result.In the proposed method after canny edge detection and hole filling,estimation of grey level gradient at a pixel is done.To identify the meaningful edges from the data ,a simple approach is to threshold the gradient magnitudes.

REFERENCES

[1] Jung Gap Kuk, Nam Ik Cho and Kyoung Mu Lee **“MAP-MRF Approach for Binarization of Degraded Document Image”***Seoul National University European Journal of Scientific Research Vol.28, pp.14-32,2009*

[2] Y. Chen and G. Leedham”**Decompose Algorithm for Thresholding Degraded Historical Document Images***” IEE Proc.-Vis. Image Signal Process., Vol. 152, No. 6, December 2005*

[3] Anna Tonazzini, Stefano Vezzosi, Luigi Bedini**“Analysis and recognition of highly degraded printed characters***”Institution of Science and Technology, Digital Object Identifier (DOI)Pisa, Italy c_ Springer-Verlag 2003.*

[4] Ergina Kavallierato, Efstathios Stamatatos **“Improving the Quality of Degraded Document Images’’** *University of the Aegean, Greece Second International Conference on Document Image Analysis for Libraries 2006*

[5] Yahia S. Halabi, Zaid SA, Faris Hamdan”**Modeling Adaptive Degraded Document Image Binarization and Optical Character System”** *Seoul National University European Journal of Scientific Research Vol.28, pp.14-32, 2009*

[6] Mahendar.R, Navaneetha Krishnan.S, Roshini Ravinayagam, Prakash.P

"Degraded Documents Recovering by Using Adaptive Binarizations and Convex Hull Concept" *International Journal of Computer Science and Mobile Computing Vol. 3, Issue. 2, February 2014, pg.187 – 196*

[7] Y. Huang, M. S. Brown, and D. Xu. **"A Framework for Reducing Ink-Bleed in Old Documents"***In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.*

[8] Tonazzini.A, Bedini.L and Salerno. E, **"Independent component analysis for document restoration"***International Journal on Document Analysis and Recognition 7(1), 17–27 (2004)*

[9] B. Su, S. Lu, and C.L. Tan, **"Combination of Document Image Binarization Techniques,"** *Proc. Int'l Conf. Document Analysis and Recognition, pp. 22-26, 2011.*

[10] J.G. Kuk, N.I. Cho, and K.M. Lee**, "MAP-MRF Approach for Binarization of Degraded Document Image,"** *Proc. Int'l Conf. Image Processing, pp. 2612- 2615, 2008.*

[11] R. Cao, C.L. Tan, Q. Wang, and P. Shen, **"Segmentation and Analysis of Double-Sided Handwritten Archival Documents,"** *Proc. Fourth IAPR Int'l Workshop Document Analysis Systems, pp. 147-158, 2000.*

[12] Q. Chen, Q.-s. Sun, P.A. Heng, and D.-s. Xia, **"A Double-Threshold Image Binarization Method Based on Edge Detector,"** *Pattern Recognition, vol. 41, no. 4, pp. 1254-1267, 2008.*

[13] N.R. Howe, **"Document Binarization with Automatic Parameter Tuning,"** *Int'l J. Documant Analysis and Recognition, doi: 10.1007/s10032-012 0192-x, 2012.*

[14] B. Gatos, K. Ntirogiannis, and I. Pratikakis, **"ICDAR 2011 Document Image Binarization Contest (DIBCO 2011*),"*** *Proc. 11th Int'l Conf. Document Analysis and Recognition, 2011.*