Content Extraction from Ancient Document Image

Sweta P. Bharshankar
Department of Computer Engineering
St. Vincent Pallotti College of Engineering & Technology, Nagpur, India.

D. W. Wajgi
Assistant Professor
Department of Computer Engineering
St. Vincent Pallotti College of Engineering & Technology, Nagpur, India

ABSTRACT
Content extraction from badly degraded ancient document is very challenging task due to the different causes of degradation. Ancient documents are of great importance to us and accumulate a significant amount of human heritage over times. These ancient documents are in the degraded form containing vital and valuable information but the contents of the document not recognized easily. There are many causes of degradations such as environmental factors improper handling and the poor quality of the materials. In recent period, the rising interest in the historical document image analysis many researchers are attracting towards extraction of contents from historical document and preserve their contents for future generations. Numerous methods exist but they are not suitable for all types of degraded documents. The proposed method is simple, robust and based on the phase binarisation model. The proposed method is divided into Preprocessing, Post processing and extraction. The Preprocessing helps to separate foreground and background. The post processing enhances the document image and Extraction helps to extract the content from the document image.

Keywords
Degradation, Phase Binarization, local and global thresholding

1. INTRODUCTION
In libraries and archives around the world, historically important documents and manuscripts are being stored. Historical documents are of great importance to us and accumulate a significant amount of human heritage over times which are being suffered high degree of degradation. These documents were written in the period of century and still they survived until today but their quality has been degraded. There are many causes of degradation such as environmental factors (e.g. dust, stain, etc.), improper handling (e.g. scanning and printing multiple times ,torn pages, etc.), and the poor quality of the materials (e.g. ink, paper etc.) used in their creation cause them to suffer a high degree of degradation of various types. Sometimes, the documents have contrast problems such as the foregrounds are usually having damaged ink with different background color, bleed though, faded ink, show through, uneven illumination, deterioration of the cellulose structure, variation in image contrast[1][2]. Due to this the contents of the document are not recognized easily. Now-a-days, these documents are converted into digital form for convenient and easy storage [3] and also to preserve their content for future generation. The huge amount of digital data produced requires automatic processing, enhancement, and recognition. Content extraction of ancient document and processing performs an important role in the analysis for removing the background noise and improving the readability of the document [1].

1. RELATED WORK
In paper [1], a robust phased based binarization model for ancient document images has been proposed. Along with this, post processing method helps in improving binarization method and enhances the ancient document has been proposed. A phase binarization method is mainly based on the three features. The maximum moment of phase congruency covariance, a locally weighted mean phase angle, and a phase preserved denoised image are the phase derived features. The proposed model is divided into three steps: preprocessing, main binarization and post processing. In preprocessing, denoised image is used instead of original image. After that Otsu's method is used to form normalized denoised image by applying linear image transform and Canny operator is applied. In main binarization, these features maps are based on Kovesi's phase congruency model. Just like color, pixels, intensity, phase information is the most important feature of image. The maximum moment of phase congruency covariance is used to estimate the background and foreground of an image, a locally weighted mean phase angle is used to detect the edge accuracy, and a phase preserved denoised image helps to preserve the important phase information in the signal. In Post-processing, Global bleed-through exclusion, Adaptive Gaussian filter and object exclusion maps are used. Global bleed-through exclusion process is applied then Adaptive Gaussian filter is used to enhance the binarized image and object exclusion process is applied which is based on median filter to remove noise. In paper [2], a comprehensive review of the methods for enhancing old document images with damaged background has been given. Three types of enhancement methods have been identified which are (i) image enhancement using binarization/thresholding method, (ii) image enhancement using a hybrid of binarization/thresholding and other methods (iii) image enhancement using non-threshold based methods. Binarization/thresholding method includes entropy_based method and locally adaptive thresholding. Non-threshold based method includes fuzzy logic. A hybrid of binarization/thresholding methods and other method, can produce promising results as compared to the other methods. In paper [3], a hybrid algorithm for thresholding has been proposed. This algorithm consists of both global and local thresholding method. The Global thresholding step has been modified such that the output will not be a binarized image but an intermediate gray level image. It is helpful as most of the background gets eliminated. Local thresholding will be applied on the result given by global thresholding step. This method is simple, robust and effective. The proposed method works better than most of the existing local and global thresholding algorithms and is able to deal with degradations which occur due to strain, ink bleed through, low contrast, water marks, dust, smear and uneven illumination etc. In paper [4], efficient binarization technique to recover the text from the severely degraded document image and also removes the degradation from old document image has been proposed.

The proposed work is based on the fusion of two well-known binarization method Gatos and Niblack method along with dialation and logical AND operation of image processing. As the proposed approach is the fusion of two methos i.e. Gatos and Niblack, Gatos proposed an algorithm which is based on the initial filtering, approximation of foreground and background region and then performs binarization. In initial filtering, Gatos used the Wiener filter which is the very popular filter for document enhancement. Wiener filter is used to reduce noise, smooth the texture but produces blur texture. In approximation of foreground and background region, gray image is filtered by applying Sauvola method and obtained binary image. The reason behind the fusion of Niblack with Gatos is that Niblack method distinguishes the text from background in the area close to the text and labels some part of the background text. The main advantage of this method is that, it completely recovers the text from badly degraded image but produces large noise and can't separate background and foreground. The problem of background and foreground separation is overcome by Gatos method. After the fusion of these two methods, dialation operation and Logical AND are used to find the region close to the text. In order to overcome the flaws of Gatos method, morphological dialation operation applied on final filtered image.In paper [5], a comprehensive survey on different methods such as Otsu's method, local-variance-based method, entropy-based method, Local adaptive method has been given. From these methods Otsu's method is the most successful global thresholding method. Along with these methods multiscale binarization, an automatic threshold approached which is based on fuzzy logic,Balanced histogram thresholding method, a pixel based binarization evaluation methodology and adaptive image contrast method are mentioned.In paper [6], a robust binarization technique by using adaptive image contrast has been proposed. The adaptive image contrast address the issues of inter/intra variation between foreground and background document image. The adaptive contrast is a combination of local image contrast and local image gradient. The local image contrast and the local image gradient are very useful feature for segmenting the text from the document. The local image gradient has been widely used for edge detection and also to detect the text strokes of the document image having uniform background. Both features are very effective and have been used in many document binarization methods. In the proposed technique, the adaptive contrast map is constructed first for an input degraded document image. Then this constructed map is binarized and combines with the canny edge map to identify the text stroke edge pixels. The document text is further segmented by local threshold which is estimated on the basis of detected text stroke edge pixels. Some post processing is applied to improve the quality of document.In paper [7], various text extraction techniques such as Adaptive Local Connectivity Map (ALCM), Expectation Maximization (EM), Maximum Likelihood (ML), Spiral Run Length Smearing Algorithm (SRLSA) etc. have been discussed. The performance comparison of these methods for document text extraction on the basis of accuracy, precision rate, recall rate, processing time has been done. It is observed that in case of accuracy is best for hybrid (connected component (CC) and texture Analysis) approach and edge based text extraction techniques. Precision and recall rate is best in case of EM algorithm and ML segmentation method. Processing time is best in case of digital filter using Haar wavelet. For handwritten text document images, accuracy is best for ALCM and histogram projection based method.In paper [8], a modified iterative global thresholding approach to separate the clusters of foreground and background has been proposed. After image equalization the relative closeness towards background intensity is computed in each iteration. Camera captured images of ancient printed documents, stone carvings and palm leaves are evaluated in this paper. In each iteration ,the average intensity of the document image is adopted as midpoint between the clusters. In the next step the remaining pixels are equalized so as to compare the histogram. The number of iterations depends on the sensitivity of successive thresholds. This algorithm is found to be effective on historical document images as well as camera captured stone carvings.In paper [9], a method which deals with recognition and matching of text in ancient and cartographic document has been proposed. Based on statistical and global features, this proposed method finds similar text. First a phase of normalization is done followed by a phase of words spotting. For extracting the textual contents of image documents, author worked on describing the word or character rather than worked on whole page of document because document image have different properties and didn't contain the same noise. For font recognition, the optical font recognition (OFR) is used. For text matching local descriptors are used such as projection profile, Euclidian Distance Map (EDM), XOR operator. Firstly EDM applied to XOR operator which obtained vector measure to calculate the error between the requested images and the images of basis. Secondly projection profile of image of characters calculated.In paper [10], survey on text line segmentation of historical documents has been discussed. Projection-profiles are commonly used for printed document segmentation. This technique can also be adapted to handwritten documents with little overlap. For printed and binarized documents, smearing methods such as the Run-Length Smoothing Algorithm can be applied. Consecutive black pixels along the horizontal direction are smeared: i.e. the white space between them is filled with black pixels if their distance is within a predefined threshold. The bounding boxes of the connected components in the smeared image enclose text lines. The Hough transform is a very popular technique for finding straight lines in images. This method is developed on a hypothesis-validation scheme.

2. THE PROPOSED APPROACH

In this paper, the proposed approach may enhance the old document image and extract the content in order to read that document easily. The proposed method works on the English ancient degraded document. It is basically divided into preprocessing, post processing and extraction. The flowchart of the proposed system is as follows:

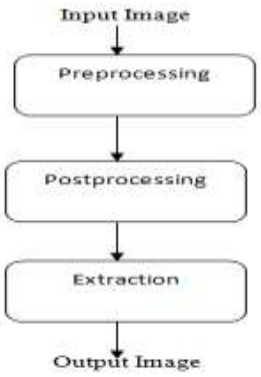

Fig.1 Overview of Proposed Approach

2.1 Preprocessing

In preprocessing, original image is converted into gray image. The denoising process consists of determining a noise threshold at each scale and shrinking the magnitudes of the filter response vector appropriately. After removing the background noise, Canny edge detector will be used to detect the edges. Canny edge detector gives better result as compared to other edge detector. In Preprocessing, the foreground will be differentiated from background and performs binarization.

2.2 Post processing

Using the result of preprocessing as input, the post processing is performed. In Post processing, Gaussian filter and median filters are used for filtering the noise. After filtering the noise the threshold value is maintain according to the image and this is save for the extraction process.

2.3 Extraction

For extracting the character Optical Character recognition (OCR) method is used. OCR is the mechanical or electronic conversion of image of type document into machine encoded text whether from scanned document or a photo of document. It is a common method digitizing the printed texts so that it can be electronically edited, searched and stored.

3. IMPLEMENTATION

The basic implementation is based on the Phased binarization model and extraction method. The Phased binarization helps to enhance the image and recovers the most of the characters in the image and the extraction method, extract the character which are not recover during the binarization. This method works on English document so it is easy to obtain characters.

Step 1: Load original Image

The original image is in degraded form having some background noise and also yellowish black in color.

Fig2. Original image

Step 2: Preprocess the original Image

Preprocessing is to differentiate the background and foreground. The preprocessed image is as follows:

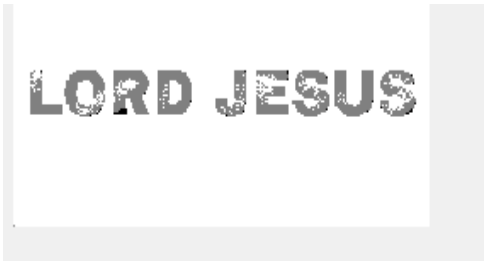

Fig3. Preprocessed Image

Step 3: Post processes

In Post processing, two filters namely Gaussian filter and median filter is used in order to remove the noise and to enhance the image. The Post processed image is as follows:

Fig4. Post processed Image.

Step 4: Extraction

In the extraction method, extract the characters which are not recovered during the phased binarization and these characters matches with the database

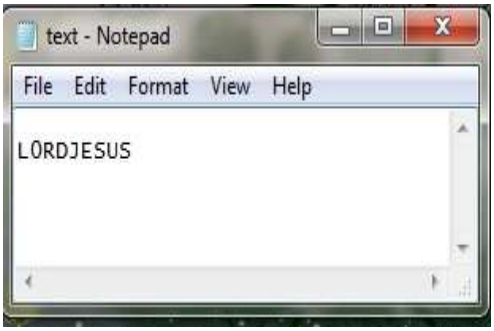

Fig5. Output Image

4. CONCLUSION

Many techniques have been proposed to deal with different types of problems of degradations. In this paper, several methods have been studied and compared, where the main idea is to enhance the image and extract the contents of the ancient document image. The proposed method works on English document. With the proposed approach, some enhancement can be done as compared to the existing methods.The proposed approach is divided into preprocessing, post processing and extraction. In preprocessing, the foreground and background of imageis differentiated and binarized.For post-processing, median filter is used to remove noise and Gaussian filter is used to further separate foreground from background and also to improve the final binary output. OCR method is used for extraction. This proposed method is simple, robust and also it works for different types of degraded documents.In future this project workcan be extended for different language documents and alsoensure the stable behavior of document image.

5. REFERENCES

[1] Hossein Ziaei Nafchi, Reza Farrahi Moghaddam Member, IEEE and Mohamed Cheriet, Senior Member, IEEE," Phase-basedbinarization of ancient document images: Model and applications" 10.1109/TIP.2014.2322451, IEEE Transactions on ImageProcessing.

[2] Sitti Rachmawati Yahya, S. N. H. Sheikh Abdullah, K. Omar, M. S. Zakaria, and C. -Y. Liong," Review on

Image EnhancementMethods of Old Manuscript with Damaged Background", International Journal on Electrical Engineering and Informatics - Volume 2Number 1, 2010.

[3] Prashali Chaudhary, B.S. Saini," An Effective And Robust Technique For The Binarization Of Degraded Document Images", International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308Jun-2014.

[4] Brij Mohan Singh, Mridula,"Efficient binarization technique for severely degraded document Images'' @ CSI Publications 2014.

[5] N. Chaki, "A Comprehensive Survey on Image Binarization Techniques", Exploring Image Binarization Techniques, Studies in Computational Intelligence 560, DOI: 10.1007/978-81-322-1907-1_2, © Springer India 2014.

[6] Bolan Su, Shijian Lu, and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE Transactions On Image Processing, Vol. 22, No. 4, April 2013.

[7] Deepika Ghai, Neelu Jain "Text Extraction from Document Images- A Review", International Journal of Computer Applications (0975 – 8887) Volume 84 – No 3, December 2013.

[8] N.Venkata Rao,A.V.Srinivasa Rao, S. Balaji and L. Pratap Reddy,"Cleaning of Ancient Document Images Using Modified IterativeGlobal Threshold", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2, November 2011.

[9] N. Zaghden , B.Khelifi , A.M Alimi , R.Mullot ," Text Recognition In Both Ancient And Cartographic Documents", Digital Heritage – Proceedings of the 14th International Conference on Virtual Systems and Multimedia VSMM 2008.

[10] Laurence Likforman-Sulem, Abderrazak Zahour, Bruno Taconet,"Text Line Segmentation of Historical Documents: a Survey", International Journal on Document Analysis and Recognition, Springer, 2006.