Content Retrieval from Historical Manuscript Images: A Review

Kitty Gupta
Student (ECE Department)
BGIET Sangrur, Punjab

Rishav Dewan
Assistant Professor (ECE Department)
BGIET Sangrur, Punjab

ABSTRACT
Ancient documents play an important role in history. Various information regarding the literature, tradition and culture is kept in these documents. These heaps of documents are degraded because of some climatic circumstances, low quality and inappropriate holding. This paper reviews on the techniques used to retrieve the necessary content from these ancient documents. The techniques include preprocessing, image binarization, thresholding methods and post processing methods. Further, during scanning the document it can get corrupted with some unwanted lines or signals termed as noise that should be eliminated.

General Terms
Document Image Processing, binarization, PSNR

Keywords
Degraded document image; preprocessing; thresholding; post processing.

1. INTRODUCTION
There are many libraries and places in which degraded Historical documents are preserved. These manuscripts are degraded because of various fault conditions in the environment or low quality of paper used. Another problem with these documents is that because of past decay of time the ink of front page get disfigure with last page. Such types of problems must be corrected by different techniques. Image binarization is that technique by which the text may be retrieved from the document. Binarization breaks the document image in two parts: image background and foreground text. Document text edges are digitized by using image binarization. Thresholding methods are used for binarization of image. Further, two types of thresholding methods: local and global thresholding. Separation of foreground and background of degraded document image is done by global thresholding method. To get information about the pixels and local area of the document image, local thresholding is used. This was also seen that global thresholding proves best with comparison of local thresholding. Another thresholding method is Otsu's method named after Nobuyuki Otsu. This is used to detect the text edges. To find text edges constructing contrast is very important and after this the edges of text can be easily identified. Clear bimodal patterns are not obtained in the degraded document. To find text stroke edges in the image, correct contrast construction is very important. To detect the text edges provide uniform background to the degrade document image. After this, text edges can be easily identified and detected by separating text and background from the image. Edge detection methods are used for edge detecting.

Now, by comparing the intensity of document image contents can be easily retrieved from the image. For comparison, assign two values '0'for background and '1' for edges. Detected edges have clear bimodal pattern which are obtained by binarization. Bimodal patterns results in the text edge sharpening. In this way useful contents can be retrieved from the document images. This paper focuses on review of such types of methods or algorithms by which quality of degraded documents can be enhanced.

(a)

(b)

Figure 1: Degraded Historical Documents

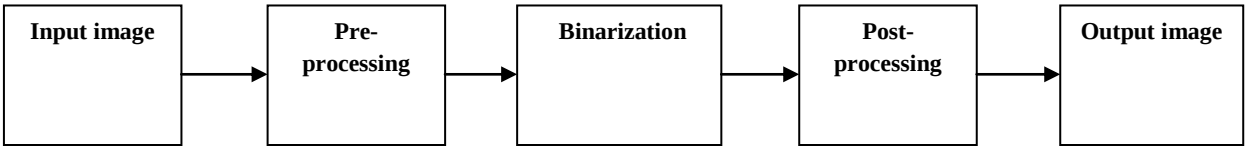

Figure 2: Block diagram for enhancing degrade documents

2. METHODOLOGY

Denoising and Enhancing of Degraded Historical documents are very important task. In the existing technology some steps are concluded. These include (a) Pre-processing (b) Binarization and (c) Post-processing. Each method is explained as below:

2.1 Pre-processing

Pre-processing is also known as pixel level processing. This processing includes conversion of coloured image into Grey scale image. Unwanted noise and lines are reduced by using noise removal filters. The noise present in the document images are margin noise, Gaussian noise and impulse noise. For reducing these noises the filters include Gaussian filter, median filter, and bilateral filter and guided image filter. Bilateral filter reduce noise without preserving edges and this limitation may be overcome by using guided image filter.

2.2 Binarization

Binarization is a processing of converting grey image into binary form. Foreground and background parts of the image are separated from each other. It is done by thresholding methods. Global and local thresholding comes in this process. Another thresholding method is Otsu method that gives the weighted sum of variance of two pixels.

2.3 Post-processing

To enhance the performance of binarization post-processing technique is applied. During scanning, text can be deviated from the base line. This kind of problem is corrected in post processing step. Text extraction and edge sharpening is also included in this.

3. LITERATURE SURVEY

In 2010, Shazia Akram et al. [11] give an overview on various techniques to enhance the document images. The techniques include Preprocessing, feature extraction and classification. Preprocessing is that state that enhances the quality of image. It is also known as pixel level processing. It includes image acquisition, noise removal and image de-skewing methods. Useful information or data may be extracted by using feature extraction method. Classification includes distribution of documents into various categories that improves indexing efficiency of document storage places**.** This paper concludes various techniques for document image processing and new methods must be developed to enhance the document images.

C. Patvardhan et al. (2012) [3] proposed the method of discrete Curvelet Transform for denoising the document images. Document images corrupted by Gaussian and impulse noises are denoised by curvelet transform. These noised gets added during scanning and transmission. Hard thresholding and global Otsu methods were also used to smooth the boundaries. In wavelet scheme Haar wavelet was used as it does not blur the image. Wavelet scheme, wavelet based scheme with edge preservation and curvelet transform was compared with (a) F-Measure, (b) Negative Rate Metric, (c) Normalized correlation, (d) Peak signal to noise ratio (PSNR). It was concluded that Curvelet transform performs better results as it preserve edge features of the noisy image and reduce Gaussian and impulse noise in the image.

B. Gangamma et al. (2012) [1] combines two methods of image processing, filtering and mathematical morphology. Bilateral filter reduce the unwanted noise in the images but unable to smooth the edges of image. Mathematical morphology helps in the extraction of edges, shapes and cracks in the texts. Further, to binarize the image global thresholding was used. The results of proposed method are comparing with Gaussian and Average filter. It was concluded that the proposed method proves better to degrade the document image.

Hossein Ziaei Nafchi et al. (2013) [5] proposed a post processing method which is based on phase- preserved denoised image and phase congruency extracted features from the document image. Non-orthogonal log-Gabor filters were used to get the information of phase and amplitude value at each point of the image. Maximum moment of phase congruency covariance (MMPCC) and locally weighted mean phase angle (LWMPA) were used detect the edge and structure of foreground text respectively. Then Otsu's method was applied to get the binarized image and median filter was used to remove the noise from the image. The results were evaluated in F-Measure, recall and Distance reciprocal distortion (DRD). These methods were tested on DIBCO and H-DIBCO datasets and the proposed methods showed better results on DIBCO datasets in terms of F-Measure, recall and DRD.

Md Iqbal Quriashi et al. (2013) [7] compare two approaches to degrade historical document images. In first case, Particle Swarm Optimization (PSO) with bilateral filter was applied. In second case, Non-linear filter and bilateral filter was applied. Then both the techniques were compared in terms of PSNR and NAE. The results conclude in favor of Particle Swarm Optimization.

Then, in 2014 Jagroop Kaur et al. [8] proposed a new filter known as guided image filter as it is an edge preserving filter. The proposed method in this paper was worked in various steps. In first step, guided image filter was applied to smooth the degrade image. Secondly, adaptive image contrast enhancement was applied for the grouping of contrast and gradient of local image. Then, final binarization was done with thresholding methods. Proposed method was compared with some old methods in terms of F-Measure, Specificity, Geometric Accuracy and PSNR. The proposed method show better results. Another advantage of using guided image filter is that it reduces noise at higher extent from the degraded document.

S. Tamilselvan et al. (2014) [12] proposed a binarization technique for retrieving contents from the degraded document images. This binarization was performed in various steps. Correct contrast of the image was constructed. Then, by using Otsu's & Canny edge detection method edges of the image

was detected. After edge detection, necessary text was extracted from the image. At last post processing method was applied to sharpen the text edges. Clear bimodal pattern of the text was extracted without blurring the image. After experimental results threshold value of output image was calculated and value ranges from 0.3-0.9. It was also concluded that contrast construction is more valuable step among other steps in the proposed method.

Haneen Khader et al. (2014) [4] describes a novel Annotation tool for handwritten historical images. This was done on English and Arabic texts for text segmentation. K-means and Otsu's thresholding methods were used in image binarization which comes under pre-processing step. Suitable binarization method was selected depending on the quality of image. On the output of thresholding, segmentation was applied to detect the lines and texts from the image. Finally, Annotation tool was applied which finds whether the text is English and Arabic. A rectangular box was appeared on each word. Last step of proposed method is saving of Annotation by creating an xml file. The aim of this tool is to eliminate segmentation errors.

Hossein Ziaei Nafchi et al. (2014) [4] introduced a phase based binarization method which worked in three steps (a) pre-processing (b) binarization (c) post processing. Denoising of image is considered in preprocessing step. Median filter was used to remove unwanted noise and lines and Gaussian filter to separate foreground from background. Main binarization is based on MMPCC and LWMPA. In post processing, to enhance the binarization Gaussian filter was applied. Further, to get ground truth image PhaseGT was used to simplify and speed up the ground truth generation process. These methods had been analyzed on dataset of DIBCO, H-DIBCO, PHIBD and BICKLEY DIARY.

4. CONCLUSION

This review paper analyzes the various algorithms and techniques for enhancing the degraded historical documents or manuscripts. The documents get degraded due to various environmental conditions. Every technique has own advantages and disadvantages. Usually for improving the quality of image binarization performs better results. It is done by various thresholding methods. Filters are used to remove the noise from degraded image and edge preserving filter show better results. Every algorithm is compare with parameters as PSNR, F-Measure and NC. In future new algorithms would be developed for improving other historical images like text on stone monuments. In future, further techniques for binarization may be used. As thresholding is a universal problem so another denoising filters may be used to improve the degraded image.

Table 1: Comparison of various methods

S.no	Reference	Method Used	Conclusion
1.	Shazia Akram et al. [11]	Preprocessing, feature extraction and classification	Gives an overview on these techniques used to degrade the digital document images.
2.	C. Patvardhan et al. [3]	Discrete Curvelet transform wavelet scheme and wavelet scheme with edge preservation. Hard thresholding and Otsu's method in wavelet scheme	After comparing three methods, curvelet transforms gives better results as it reduce noise and preserve edge features.
3.	B. Gangamma et al. [1]	Filtering and Mathematical Morphology. Bilateral filter and Binarization is done by global thresholding.	One disadvantage of bilateral filter is that it not edge preserving filter. Another filter can be used to preserve edges.
4.	Hossein Ziaei Nafchi et al. [5]	Post processing method, MMPCC, LWMPA, Otsu's method and median filter.	On DIBCO dataset proposed method shoe better results
5.	Md Iqbal Quriashi et al. [7]	PSO, bilateral filter and non-linear filters	PSO performs better.
6.	Jagroop Kaur et al. [8]	Guided image filter, Binarization	Guided image filter is edge preserving filter and will used to detect brain tumor.
7.	S. Tamilselvan et al. [12]	Contrast construction, Otsu's method and post processing.	Contrast construction gives better results.
8.	Haneen Khader et al. [4]	Annotation tool, image segmentation and Otsu's method	Annotation tool find whether the text is English or Arabic.

5. REFERENCES

[1] B. Gangamma, Srikanta Murthy K, Arun Vikas Singh, "Restoration of Degraded Historical Document Image", Journal of Emerging Trends in Computing and Information Science, Vol. 3, Issue 5, May 2012.

[2] Bolan Su, Shijian Lu, Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE Transactions on Image Processing, Vol. 22, Issue 4, April 2013.

[3] C. Patvardhan, A.K. Verma, C.V. Lakshmi, "Denoising of Document images using Discrete Curvelet Transform for OCR Applications", International Journal of Computer Applications(0975-8887), Vol. 55, Issue 55, October 2012.

[4] Haneen Khadar, Abeer Al-Marridi, "An Interactive Annotation tool for Indexing Historical Manuscripts", IEEE transactions on image processing, 2014.

[5] Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, "Application of Phase –Based Features and Denoising in post processing and binarization of Historical Document Images", IEEE 12th International Conference on Document Analysis and Recognition", pp. 220-224, 2013.

[6] Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, "Phase Based Binarization of Ancient Document Images: Model and Applications", IEEE Transactions on Image Processing, Vol. 23, Issue 7, July 2014.

[7] Iqbal Quraishi, Mallika De, "A Novel Hybrid Approach to Restore Historical Degraded Documents", IEEE International Conference on Intelligent Systems and Signal Processing (ISSP), 2013.

[8] Jagroop Kaur, Rajiv Mahajan, "Improved Degraded Document Image Binarization using Guided image filter", International journal of Advance research in Computer Science and Software Engineering, Vol. 4, Issue 9, September 2014.

[9] Konstantinos Ntriogiannis, Basilis Gatos, Ioannis Pratikakis, "Performance Evolution Methodology for Historical Image Binarization", IEEE Transactions on Image Processing, Vol. 22, Issue 2, February 2013.

[10] Manish Yadav, Swati Yadav, Dilip Sharma. "Image Denoising Using Orthonormal Wavelet Transform with Stein Unbiased Risk Estimator", IEEE student's Conference on Electrical, Electronics and Computer Science 2014.

[11] .Shazia Akram, Mehraj-Ud-Din Dar, Aasia Quyoum, "Document Image Processing- A Review", International Journal of Computer Applications (0975-8887), Vol. 10, Issue 5, November 2010.

[12] S. Tamilselvan, S.G. Sowmya, "Content Retrieval from Degrade Document Images using Binarization Technique", IEEE International Conference on Computation of Power, Energy, Information and Communication(ICCPEIC), pp.422-426, 2014.