A NOVEL IMAGE RESTORATION ALGORITHM FOR DIGITIZED DEGRADED HISTORICAL DOCUMENTS

Rupinder Kaur, Research Scholar, Guru Kashi University, Talwandi Sabo (Bathinda).

Jaspreet Kaur, Assistant Professor, Guru Kashi University, Talwandi Sabo (Bathinda).

Abstract— **the historical documents are the important traces of the history. Historical documents carry the information about the tradition, religion, science, literature, war strategies of the ancient times, which are very useful to understand the life and culture of the ancient civilizations. These historical documents are usually recovered from the archaeological findings. Historical documents are usually found degraded due to the straining because of many factors. The historical documents are found in hundreds of thousands in numbers, so take a lot of time for manual restoration of these documents. In this paper, a new algorithm is proposed, which can be used to restore the historical documents using the combination of digital image processing based image restoration techniques.**

Keywords- **Archaeological findings, Historical documents, Image restoration techniques, Show-through, straining.**

1. INTRODUCTION

Old historical documents are quite important because they contain information regarding our culture, economics etc. In libraries and museums, the histories of civilizations are stored. These historical documents cannot be accessed by the most of the people in the world because of time and travel cost. As the historical documents contain very important information, so they need to make accessible by the people around the world anywhere. For this purpose, digital libraries should be created to improve the access to scientific, educational and historical documents and information. Digital libraries can provide powerful opportunities for promoting education, improving knowledge and providing historical background.

In past few years the use of internet technology has been increased, this advancement provided the opportunity to make this literature available to the people from all around the world. To make accessible historical documents on the internet, image databases of these manuscripts and documents should be created. By creating these image databases, we can save the documents from further degradation and at the same time make them accessible in the digital library. It would be very difficult to create a large volume of text data because it would require a lot of labor and time if typed in manually. Therefore, we need to find another technique to store the historical documents in digital library. One approach is to convert the document into a digital image and save the image in the digital library that can save a lot of time and work.

Noises and other low-resolution components appear on the historical documents after digitization of these documents. These components affect the overall visual appearance of the documents. Historical documents suffer from several degradations, which have been introduced along time and can be of very different nature. A few examples of such degradation types are: bleed-through effect, ink fading, deterioration of paper material and cellulose structure. However, one major problem of developing a digital library of historical documents is the *show-through*. Most of the documents are written on both sides. When ink impression from one side appears on the other side, is known as the *show-through* problem which makes documents difficult to read. We need to restore these documents to make them easily readable. By removing the *show-through*, the compression time for the images can be reduced and thus leading to faster download over networks. Figure 1.1 shows the old degraded historical document with high *show through.*

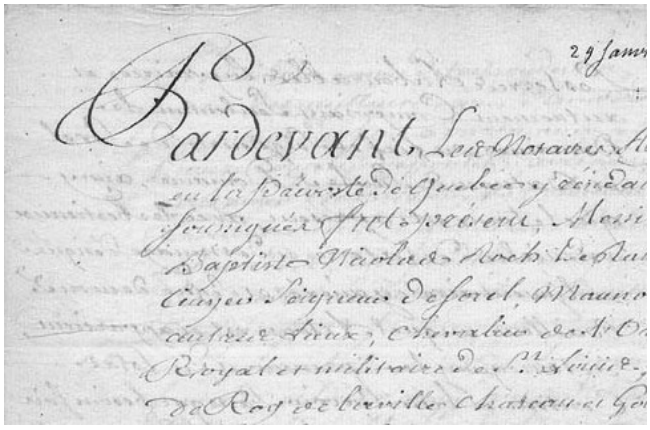

Fig 1.1: Sample of an old degraded document with high *show-through*

By removing the *show-through* from this image, a clear background is obtained. By applying a contrast enhancement technique the blur edges of the text, sketch or any other image can be improved. Some techniques have been developed to remove the noise from historical documents. But these techniques work with only one side of the document at a time and only for a pale *show-through.*

The objective of this thesis to develop a technique which works with both sides of the document at the same time and which will give best results than existing techniques to remove the *show-through* from historical documents. For this purpose, a digital scanner is used to get the image, but if the document is in too poor condition then we can use a digital camera to get the image. After getting the both sides of the image restoration technique is applied on both sides with minor modification. The image of the front of

document is called as recto and image of back is called verso. Then results are tested to find the best results for the restoration of image.

2. LITERATURE REVEIW

Donohue presented a soft thresholding method for denoising in 1-D signal using wavelets pyramidal filtering in 1995. Chang, Yu, and Vetterli introduced an adaptive wavelet thresholding for image denoising and compression. Shijian Lu and associates developed a technique, which estimates the document background surface using an iterative polynomial smoothing procedure. Using L1-norm image gradient, the text stroke edge is detected from the compensated document image. Finally, the document text is segmented by a local threshold that is estimated based on the detected text stroke edges. Napa Sae-Bae and Somkait Udomhunsakul presented adaptive BSVD method to denoise the image. It is found that these techniques alone cannot improve the visibility of the degraded images. Yahia S. and associates presented an enhanced system for degraded old document. The developed system was able to deal with degradations, which occur due to shadows, non-uniform illumination, low contrast, and noise. Laurence Likforman-Sulem developed a novel method for document enhancement, which combines two recent powerful noise-reduction steps. The first step was based on the Total Variation framework. Second step was based on Non-local Means. Non Local Mean filter computational complexity depends on the size of the patch and window. K. Shirai et al. presented a method, which performs anisotropic morphological dilation via implicit smoothing for the purpose of restoring the degraded character shapes of binarized images. Exploiting the idea of geodesic morphology that the binary image and its distance-transformed image are interconvertible, they applied a smoothing method not to the binary image but to the distance-transformed image, and then reconvert it by binarization. Md. Iqbal Quraishi and Mallika De proposed a novel approach to enhance ancient historical documents. To enhance these digital format documents a two way approach is considered. At first Particle Swarm Optimization (PSO) and bilateral filter is applied. At second level Non-Linear Enhancement with bilateral filter is applied. Reza Farrahi Moghaddam et al. combined two complementary approaches. First, multi-level classifiers, which take advantage of the stroke width a priori information, allow to locate candidate character pixels. Second, a level set active contour scheme is used to identify the boundary of a character. Then they have been tested these approaches on a set of ancient degraded Hebraic character images.

3. PROBLEM FORMULATION

Historical documents are often degraded by ink bleed through, stains, smudge, smear, cracks, watermarks, dust marks etc. These various types of degradations make it difficult to read the text. As historical documents contain very useful information, this data has to be preserved. The degraded historical documents are available in higher amounts. Archeologist need to restore the historical documents manually or digitally. It is very time consuming task to rewrite the larger number of documents. In the base paper, authors have used Particle Swarm Optimization (PSO) algorithm for image quality enhancements. We are proposing a combination of novel image enhancement techniques to improve the quality of historical documents.

4. PROPOSED SYSTEM

The aim is to segment the text from the degraded document images. That means the foreground and background should be separated. For this purpose we have chosen a method known as binarization. It means we will be converting the digital image into a binary form i.e. '0' and '1'. The resultant image will be having the background as white and text as black. This method is helpful because now the text is more legible and also it requires very less memory for storage. The basic and most important part for this procedure is to determine a threshold for binarization. Various steps followed will be: 1) normal image to grayscale conversion, 2) A Gaussian filter to remove the noise, 3) Image dilation to estimate the background, 4) Estimated background subtraction from grayscale image, 5) Global thresholding (OTSU) for modification, 6) Savoula threshold for local area or window, 6) Calculate parameters (PSNR, MSE, RMSE, and NAE), 7) Comparison with existing technique.

5. OBJECTIVES

- To remove the show-through from digitized historical documents, so that documents should be easily readable
- To remove the stains, smudges, cracks, watermarks, dust marks from the digitized historical documents
- To produce the more accurate results for digitized historical documents restoration than existing algorithms
- To improve the quality of wide variety of digitized historical documents

6. RESEARCH METHODOLOGY

Image acquisition will load the image in our program on which the enhancement has to be applied. Image acquisition will convert the image from its existing digital format in the form of a matrix. Second step will contain the feature extraction, which will use image binarization, boundary tracing, background removal and/or region growing in the perfect arrangement of these techniques. Binarization is an ordinary image processing technique to convert the image matrix values in to 1 and 0, which are called binary values; hence this technique is called binarization. Binarization will convert the image into black and white image according to the threshold value. Boundary tracing will select the objects (text in this case). Objects inside the boundaries will be extracted and then the feature enhancement will be performed on the extracted text. The boundary removal will be applied on the image before or after the feature extraction, wherever it will produce the best results. The performance parameters considered will be Elapsed Time, Peak Signal to Noise Ratio (PSNR), Normalized Absolute Error (NAE) and Mean Square Error (MSE) and Root Mean Square Error (RMSE).

We will start our research project by conducting a detailed literature review on the historical document and other similar document digital restoration to know the problem in detail. Then, a detailed algorithm designed would be generated for the restoration of historical documents. The simulation would be implemented using

MATLAB. The obtained results would be examined and compared with the existing security mechanism to address the similar issues.

7. CONCLUSION

In this research project, a new algorithm is being developed for the restoration of the digitized historical documents. The new algorithm will be faster and accurate than the existing algorithm. Also, it will work with a wider range of the historical documents. The new algorithm will be developed using MATLAB simulator. Performance of the new algorithm will analyzed based on elapsed time and accuracy of the alphabet/object shape restoration. In addition, the new system will be analyzed with Type 1 and Type 2 errors to collect the statistical data to judge the performance of the new algorithm.

REFERENCES

[1] Md. Iqbal Quraishi, Mallika De, Krishna Gopal Dhal, Saheb Mondal, Goutam Das "A novel hybrid approach to restore historical degraded documents", ISSP, vol. 1,pp. 185-189, IEEE 2013.

[2] K. Shirani Y. Endo, A. Kitadai, S. Inoue, N. Kurushima, "Character Shape Restoration of Binarized Historical Documents by Smoothing via Geodesic Morphology", ICDAR, vol. 12, pp. 1285-1289, IEEE 2013.

[3] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," ACM TOG (Proc. SIGGRAPH Asia), vol. 31, no. 6, pp. 139:1–139:10, 2012.

[4] A. Kitadai, M. Nakagawa, H. Baba, and A. Watanabe, "Similarity evaluation and shape feature extraction for character pattern retrieval to support reading historical documents," in Proc. IAPR Intl. WS. DAS, pp. 359–363, 2012.

[5] R. F. Moghaddam, D. R.-H´enault, and M. Cheriet, "Restoration and segmentation of highly degraded characters using a shape-independent level set approach and multi-level classifiers," in Proc. IAPR ICDAR, pp. 828–832, 2009.

[6] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "Ocr binarization and image pre-processing for searching historical documents," Elsevier Trans. Pattern Recogn., vol. 40, no. 2, pp. 389–397, 2007.

[7] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," Elsevier Trans. Pattern Recogn. vol. 39, no. 3, pp. 317–327, 2006.

[8] D. Tschumperl´e and R. Deriche, "Vector-valued image regularization with PDE's: A common framework for different applications," IEEE Trans. PAMI, vol. 27, no. 4, pp. 506–517, 2005.

[9] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in Proc. IEEE ICCV, pp. 839–846, 1998.

[10] Extrapolation, interpolation, and smoothing stationary time series. New York: Wiley, 1949.

