Performance Evaluation of Novel Historical Documents Restoration Algorithm

Er **Neha Kundal** [1],Er **Anantdeep** [2]

[1] Student, [2] Assistant professor, Department of computer engineering,
Punjabi university Patiala, India

Abstract— Historical documents contain important contemporary information about a person, place, events of that era. A beautiful work on medicine, religion and science written by the scholars, called Vedas, is preserved in India. The histories of civilizations are stored in libraries and museums. Around the world, there is a treasure of excellent literature, which cannot be accessed by most of the people in the world because of time and travel cost. By restoring these historical documents digitally and putting them into a digital library, powerful opportunities for improving knowledge and providing historical background can be available.

Image restoration is the process which is used to restore a degraded document back to the original image. Image Restoration involves denoising using filter techniques like Gaussian Filter, Fuzzy Filter etc.

In this Paper Images have been restored using OTSU Thresholding, SAUVOLA Thresholding and Hybrid Method.

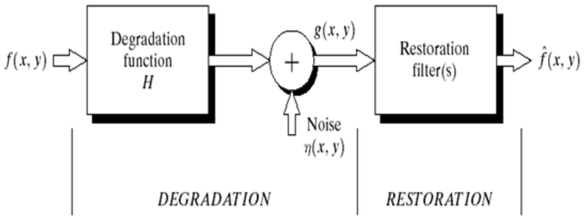
I. Introduction

A.Image Restoration

Image restoration is the process of taking corrupted/noisy image and producing clean original image. corruption may come in many forms such as camera misfocus, motion blur, noise. In this paper we will introduce and implement several of methods used in the image processing to restore images.

Image degradation occur when image undergoes loss of stored information either due to digitization or conversion decreasing visual quality.

Degradation model: In degradation model, the image is blurred using degradation function and additive noise. The following Figure 1 represents the structure of degradation model .



Degradation Model

The degraded image can be described by the following equation:

$$g(x, y) = f(x, y) * h(x, y) + \eta(x, y)$$

In equation (1), g is the degraded image, h is the degradation function, f is an original image and n is the additive noise.

B. Historical Documents

Historical documents are that documents which contains information about person, place.

A collection of rare books and manuscripts, including early copies of works by Aristotle, Dante, Euclid, Homer, and Virgil, are available in the Vatican Library in Rome. And a beautiful work on medicine, religion and science written by the scholars, called Vedas, is preserved in India. The histories of civilizations are stored in libraries and museums. Around the world, there is a treasure of excellent literature, which cannot be accessed by most of the people in the world because of time and travel cost. By improving access to scientific, educational and historical documents and information, digital libraries can create powerful opportunities for revamping education, improving knowledge and providing historical background.

In the past few years the technology advances in two areas, computers and communication networks have helped in creating the Internet today. This advancement has provided the opportunity to make this literature available to the people from all around the world. By developing image databases of these manuscripts and documents, and making them available on the Internet it will be easier for the people to access them. Another significant advantage of image databases is we can save the documents from further degradation by preserving them and at the same time make them acc perpetuity in the digital library.So far we are able to access images of historical artwork such as old paintings, sculptures, prints etc. through the digital libraries.

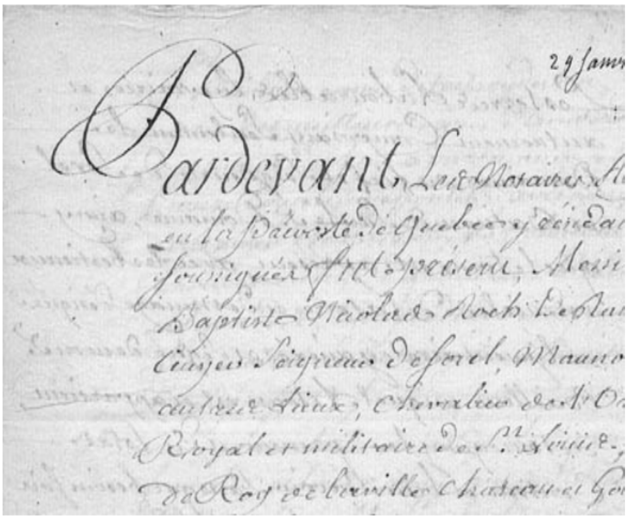


Figure 1. 1: Sample of an old degraded document

II. LITERATURE SURVEY

The need for efficient image restoration methods have grown with the massive production of digital images of all kinds, often taken in poor conditions. Even though good cameras are available, images may not be in a condition to directly use for the analysis. From the literature survey, it is observed that, there are several techniques like median filter, Gaussian filter for noise suppression are available. However, the techniques effectively suppress the noise but fail to preserve many useful details.

Md. Iqbal Quraishi (2013) said that the old degraded historical documents carry various important information regarding our culture, economics etc. Proper restoration of these documents is very necessary. They 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. Both the approaches are then tested visually and quantitively to show the effectiveness of the approach.

K. Shirai (2013) 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. This allows us to apply conventional smoothing methods for continuous intensity, i.e., gray scale, images to the discrete intensity, i.e., binary, image implicitly. For instance, by using anisotropic diffusion together with geodesic dilation, anisotropic dilation along the stroke direction is obtained and brings better results.

Akihito Kitadai (2012) said that shape features of character patterns on the documents are unstable or missing because most of the documents have been stained and degraded deeply. Digital archives of the documents with accurate character pattern retrieval methods are helpful for archaeologists and historians. They proposed a similarity evaluation method for character patterns with missing shape parts. It collaboratively works with non-linear normalization for such patterns, and modifies the templates for each trial of the retrieval efficiently. In the experiences using 4,911 Kanji (Chinese origin) character patterns from the Japanese historical documents called mokkans, the method shows improvements of the retrieval accuracy. They also presented a simple implementation of gradient feature extraction to compare the chaincode feature with the gradient feature in the retrieval.

B.Gangamma (2012) presented the combination of special domain method along with set theory operations are used to enhance the historical image. The proposed method eliminate noise, background and enhance the contrast of image. The result of proposed method is compared with mean and Gaussian filter. The restored image will have clear uniform background and foreground with enhanced character appearance.

B.Gatos (2006) presents a new adaptive approach for the binarization and enhacement of degraded documents. The proposed method does not require any parameter by user and deal with degradation which occur due to shadows, non uniform illumination, low contrast,smear and strain.we follow several steps: a pre-preprocessing procedure using low paas weiner thresholding by combining the calculated background surface with original image while incorporating image up sampling and finally post proceesing step in order to improve the quality of text regions and preserve stroke connectivity.

III. PROPOSED WORK

Basically our purpose is to extract text from the degraded documents by separating background from the foreground by using method called binarization. In this paper the digital image of the document shall be converted into the binary form i.e. 0 & 1. The resultant binary image so obtained shall be having black text and a white background. This text will be more legible and require lesser storage space. First of all a threshold level has to be identified for this binarization.

This will involve the following steps:

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.)sauvola thresholding 7.) fuzzy filters and 8.) comparison with parameters.

The various types of historical images will be coming into the account under this research. The proposed research is based upon tackling the different levels and types of the noise in the historical noises. To achieve such objectives, the fuzzy based filters are the best options. The proposed model will be using the fuzzy filters for the purpose of the denoising will be designed to remove the various levels of noise from the historical images when they will be available with the various classes of images. The hybrid filter will be using the combination of three filters such as, Histrogram adaptive fuzzy filter (HAF), weighted fuzzy mean filter and minimum maximum fuzzy filter. The performance parameters will be PSNR, NAE (Normalized absolute error),MSE(Mean square error) and elapsed time.

Historical documents are generally in degraded form because of ink bleed, watermarks, mutilation, stains, smudge and cracks etc. So it is difficult to read these documents since these historical documents are of great importance, we have to restore these documents. We may restore it manually or digitally. In base paper Particle Swarm Optimisation technique for restoration of these documents is used. In this paper, The geodesic morphological model will be compared with the fuzzy noise filter and shape restoration model using colour and texture for the purpose of historical document restoration will be compared with each other on the basis of various performance parameters.

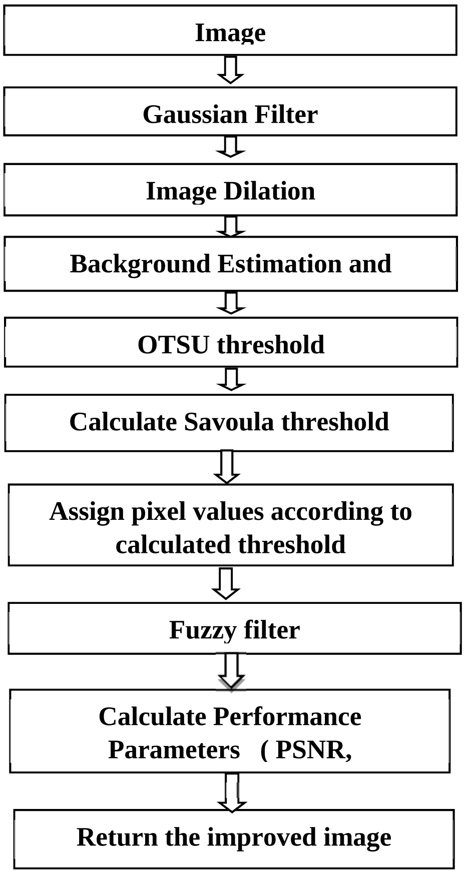


Image
Gaussian Filter
Image Dilation
Background Estimation and
OTSU threshold
Calculate Savoula threshold
Assign pixel values according to calculated threshold
Fuzzy filter
Calculate Performance Parameters (PSNR,
Return the improved image

IV.PERFORMANCE PARAMETERS & EXPERIMENTAL RESULTS

1) Peak Signal to Noise Ratio (PSNR)

Peak signal-to-noise ratio, often abbreviated PSNR, is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR value is drawn by calculating the formula on the extracted image and original secret image. The PSNR values testify the quality of the images produced before and after the algorithm processing. Higher PSNR shows the better quality of the results. The PSNR can be calculated using the below formula:

$$\text{PSNR} = 10.\log_{10}\left(\frac{Rn * Cn}{MSE}\right) \qquad (1)$$

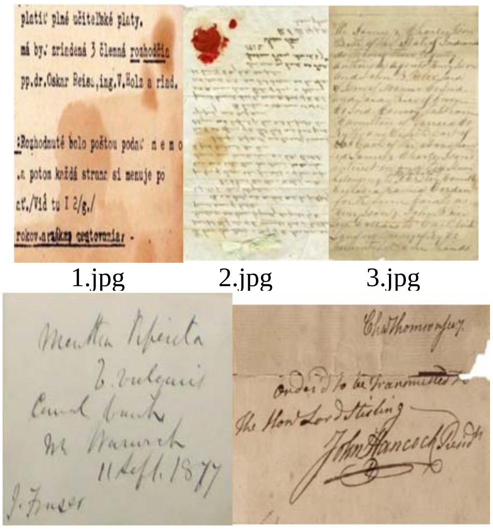
Where, **Rn** is the number of rows of the image, and **Cn** is the number of columns of the image.

2) Normalized Absolute Error (NAE):-

Normalized absolute error should be the minimum in order to minimize the difference between original and filtered image.

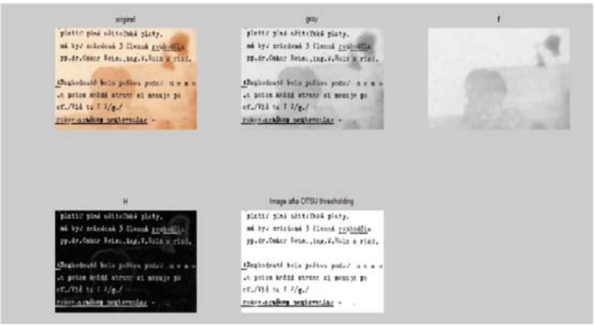
3) **Mean Square Error(MSE) :-**The mean square error measures the average of the squares of the errors i.e. the difference between the actual and the estimated signals. The MSE is the second moment (about the origin) of the error, and thus incorporates the variance of the estimator and its bias. For an unbiased estimator, the MSE could be the variance of the estimator. Just like the variance, MSE has the same units of measurement as the square of the quantity being estimated.

INPUT IMAGES:-

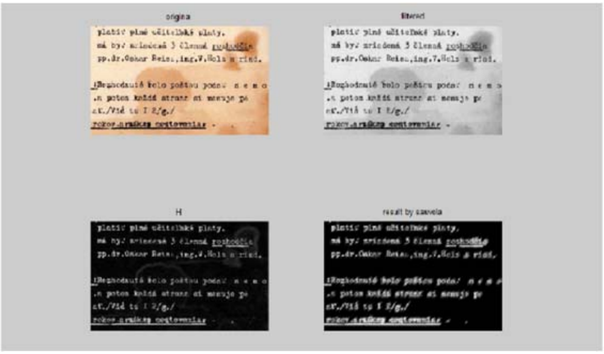


jpg image

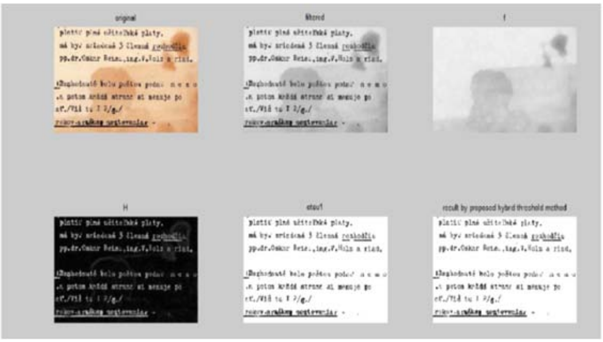
OUTPUT IMAGE



Result by otsu method



Result by sauvola method



Result by proposed method

V. Comparison Of Otsu, Savoula And Hybrid Algorithm

Table 1: The peak signal to noise ratio based performance evaluation

PSNR		
OTSU	**SAUVOLA**	**PROPOSED MODEL**
49.7548	81.75958	89.19921811
55.92627	90.81418	87.51800475
59.55622	96.72239	86.8946317
50.39101	83.01345	91.01173765
50.51256	83.72943	88.31812463

Table 2: The mean square error based performance evaluation

MSE		
OTSU	**SAUVOLA**	**PROPOSED MODEL**
2.435699	1.311102	0.430207391
2.904881	1.979748	0.111119223
2.40872	1.594896	0.602843089
2.221747	1.117418	0.318778302
2.507858	1.085516	0.801014623

Table 3: The normalized absolute error based performance evaluation

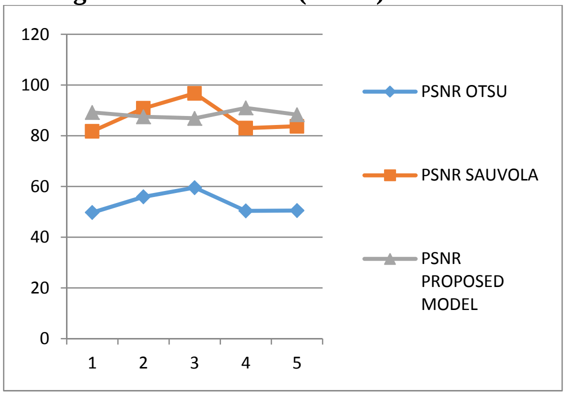
NAE		
OTSU	**SAUVOLA**	**PROPOSED MODEL**
0.772336	0.992144	0.076620358
0.815184	0.998724	0.561130027
0.741935	0.994406	0.737788252
0.288784	0.990208	0.703324127
0.575833	0.999011	0.737517765

Table 4: The elapsed time based performance evaluation

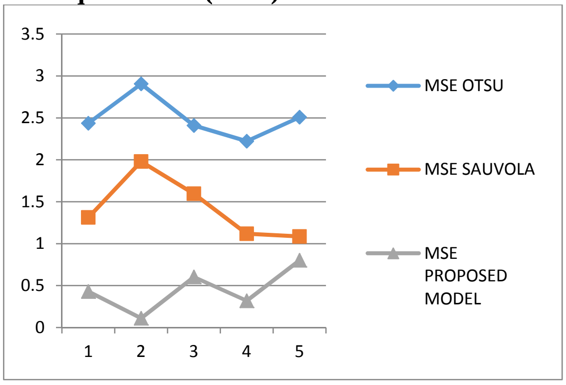
ELAPSED TIME		
OTSU	**SAUVOLA**	**PROPOSED MODEL**
0.04677	1.52589682	0.91317702
0.056657	7.294770534	4.631505461
0.119584	16.04184357	10.71769306
0.01321	1.672019492	1.035159029
0.014323	1.874616242	1.128058227

VI. Graphical Representation

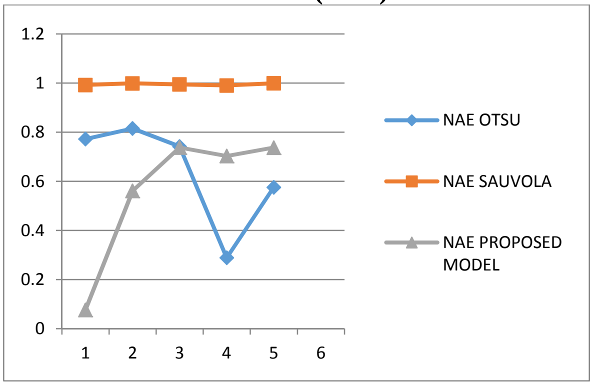
1)Peak Signal To Noise Ratio(PSNR)



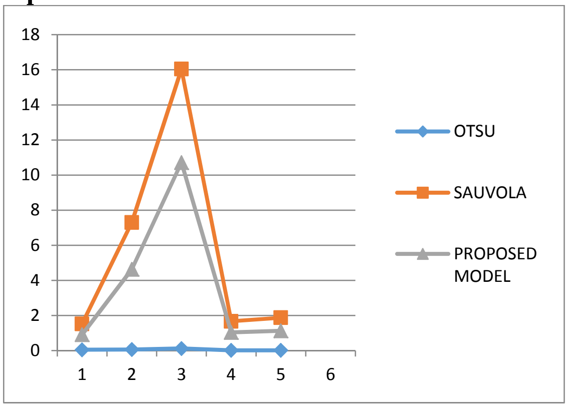
2)Mean square error(MSE)



3)Normalized Absolute Error(NAE)



4)Elapsed Time



VII. CONCLUSION & FUTURE SCOPE

The proposed algorithm has been written in the MATLAB simulator using a combination of morpohological operations, de-noising filters, image de-blurring techniques and shape reconstruction, etc. The proposed algorithm has been aimed to solve the problem of digital historical document restoration and to produce better results from the existing algorithms. The proposed algorithm has proved to be efficient when tested with a set of historical images. The historical image restoration algorithm has been tested with the images of various sizes, and the results have been collected and publish in this document along with the comparison table with the existing techniques table on the basis of normalized absolute error (NAE) and peak signal to noise ratio (PSNR),mean square error(MSE) and Elapsed time which is used to compute the running time. The proposed algorithm has performed well in the terms the parameters. The PSNR value has been improved and enhanced more than the existing two techniques. In the proposed algorithms, the Hybrid method perform better than the exixting method that is otsu method and sauvola method. However, our proposed algorithm has been improved on both of the fronts.The proposed algorithm has produced the stronger and better results against both of existing algorithms in the terms of PSNR. The PSNR value is used to measure the quality of the image after image processing images. The proposed technique has been proved to be better on both fronts when compared to the existing,otsu method and sauvola method.

VIII.FUTURE WORK

The proposed algorithm is capable of producing the quality results in terms of document clarity. The proposed algorithm has a critical set back of elapsed time. In the future, the proposed algorithm can be enhanced to produce the results quicker than the existing work. In the future, the authors/researchers can improve or develop a new model to enhance the performance better than the proposed model

REFERENCES

[1] Md. Iqbal Quraishi,Mallika De, Krishna Gopal Dhal, SahebMondal, 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] 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.

[4] B Gangamma and Srikanta Murthy K, "Enhancement of Degraded Historical Kannada Documents", International Journal of Computer Applications, Vol. 29 No.11, 2011.

[5] B. Gangamma, Srikanta Murthy K, and ArunVikas Singh," Restoration of Degraded Historical Document Image", Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 5, 2012.

[6] 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.

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

[8] Chen Yan & Graham Leedham," The Multistage Approach to Information Extraction in Degraded Document Images", IEEE, 2004.

[9] Kusum Grewal and RenuMalhan, " DESIGN OF MORPHOLOGICAL APPROACH TO DETECT AND ELIMINATE INK BLEED IN DOCUMENT IMAGES", IJREAS, Volume 2, Issue 2, 2012.

[10] Krisda Khankasikam, " Restoration of Degraded Historical Document Image: An Adaptive Multilayer-Information Binarization Technique", JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, 2013.