Development of an Advanced Technique for Historical Document Preservation

Latika Kumawat[1], Navneet Agrawal[2]

[1]M.Tech Scholar, Department of ECE, CTAE College, Udaipur, Rajasthan, INDIA
[2]Assistant Professor, Department of ECE, CTAE College, Udaipur, Rajasthan, INDIA

Abstract—*In this paper, technique used for historical document preservation is explored. In this paper a noise estimation technique is applied to know noise standard deviation. We first estimate or detect level of noise present in noisy images by selecting weak textured patches in image on the basis of gradient matrix and its statistical properties, then eliminate that noise through non local means(NLM) denoising technique that will use estimated noise level as filtering parameter for eliminating noise from the image. This technique is based on weighted average of the similar pixels in historical image. Non local means techniques removes noise from images without taking care of noise level ,it is mandatory to take care of noise level for best preserving Historical document images.*

Keywords— Filter, Historical document, Noise, Noise removal method, Noise level estimation.

I. INTRODUCTION

Historical documents play a very important role as it expresses the difference between past and present as well as future too.So it is important to retrieve the information in order to preserve the Historical document images.There are many factors which affect the accuracy of text documents. When historical documents are digitized by Scanning of the documents, they generally contain noise due to use of scanner , printer and age of the document. Historical documents are different from the other documents in terms of foreground and background,so they are to be treated in different ways.[4]
The noise estimation and the noise removal technique [1]proved itself much better as a important part in enhancing or to make historical documents readable.This paper has been organized in number of sections. Section 2 involves basic steps used. In section 3 noise level estimation is explained. Section 4 having the details of noise removal technique .Section 5 deals with the output of the used techniques. and section 6 comprises of conclusion.

II. STEPS INVOLVED

Steps involved in the preservation of Historical document preservation are as follows in fig 1:

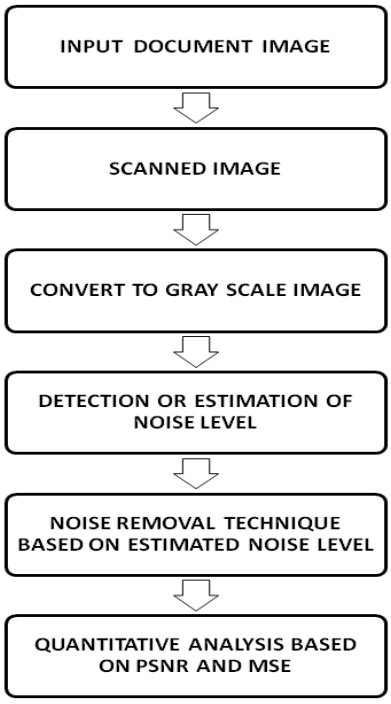


Fig 1: Block diagram of the steps involved in preservation

It is very complicated to get useful information from noisy images. That is why it is necessary to denoise images to get useful information from them. But performance of denoising techniques or degree of removal of noise from noisy images usually depends on the level of noise present in images. So it is an important task to know about the true level of noise present in images and then depending on that standard deviation of noise, noise can be removed from images by using specific image denoising techniques.

III. NOISE LEVEL ESTIMATION

In order to remove noise we must know the level of the noise present in the scanned historical document. Estimation or detection of the noise level for image is very important parameter to improve the effectiveness of the denoising.
A patch based noise level estimation algorithm[6] is proposed in the current work, with patches generated

from noisy image .We can estimate the image patches, if the image consist only weak textured patches or smooth patches in the image.the challenge in front of the patch based noise estimation is to select weak textured patches. To select weak textured patches from noisy image based on gradient of patches and their statistics. We estimate noise level from selected weak textured patches using PCA(principal component analysis).Where we can write the image model as :

$$z_i = y_i + n_i \quad (1)$$

z_i = Original image patch, y_i = Observed vectorized patch corrupted by i mean Gaussian noise vector i.e n_i The goal of the noise level estimation is to determine unknown standard deviation σ_n of the noisy image. Consider variance of data projected on to a definite axis, the minimum variance direction computable by using PCA.

The variance of data projected onto the minimum variance direction is equals the minimum eigenvalue of covariance matrix.Then the equation derived is

$$\lambda_{min}(\mathbf{\Sigma y}) = \lambda_{min}(\mathbf{\Sigma z}) + \sigma_n^2 \quad (2)$$

Where $\mathbf{\Sigma y}$ = covariance matrix of noisy patch $y_{i.,}$
Σ_z =covariance matrix of noise free patch z_i
λmin($\mathbf{\Sigma}$) =minimum eigen value of matrix $\mathbf{\Sigma}$.
The noise level can easily be estimated if we can decompose minimum eigen value of covariance matrix of noisy patches as eq.(2). As decomposition is ill posed problem because minimum eigen value of covariance matrix of noise free patches is unknown.we can easily estimate the level of noise if we can select weak textured patches from noisy image. So we can calculate the noise level[6] by

$$\hat{\sigma}_n^2 = \lambda_{min}(\Sigma y') \quad (3)$$

$\mathbf{\Sigma}_{y'}$= selected weak textured patches and we can estimate noise level easily we are known of weak textured patches.

3.1 Selection of weak textured patches

For selection of weak textured patches firstly, Image pattern can be measured efficiently by gradient covariance matrix,the gradient covariance matrix ,i.e .C for image patch y is defined as:

$$C_y = G_y^T G_y \quad (4)$$

$$G_y = [D_h y \;\; D_v y] \quad (5)$$

Where D_h = horizontal derivative operator, D_v = vertical derivative operator, both represent a matrix

More information about the image patch can be given through eigen vectors and the eigen values of gradient covariance matrix. Like maximum eigen value of gradient covariance matrix shows the potency of dominant direction of that patch. Larger maximum eigen value shows the richer texture. In this reading as the quantitative measure for texture strength of the image patches we have used this maximum eigen value of gradient covariance matrix. As eigen values of covariance matrix is very responsive to noise , so that we should select the weak textured or smooth patches from that noisy image.

Consider a flat patch having N pixels where Gaussian noise with standard deviation *σn is* added,flat patch.then as the gradient of the perfectly flat patch is zero,so we can compute the probable gradient covariance matrix of the noisy flat patch :

$$E(C_y) = E(C_n) = E\left(\begin{bmatrix} n^T D_h^T D_h n & n^T D_h^T D_v n \\ n^T D_v^T D_h n & n^T D_v^T D_v n \end{bmatrix}\right) = \begin{bmatrix} E(n^T D_h^T D_h n) & 0 \\ 0 & E(n^T D_v^T D_v n) \end{bmatrix} \quad (6)$$

Two diagonal component have same arithmetical properties. For that reason ,we specially examine upper-left component.Let $\xi(\boldsymbol{n}) = \boldsymbol{n}^T \boldsymbol{D}_h^T \boldsymbol{D}_h \boldsymbol{n}$ we estimated distribution of $\xi(\boldsymbol{n})$ by gamma distribution to make simpler the problem. The moment generating function of variable $\xi(\boldsymbol{n})$ can be derive as:

$$M_\xi(t) = E(e^{t\xi(n)} \prod_{i=1}^{N} \frac{1}{(1-2\sigma^2 t\lambda_i)^{1/2}} \quad (7)$$

Where as λ_i is the i-th eigenvalue of matrix $\boldsymbol{D}_h^T \boldsymbol{D}_h$.The moment generating function of Gamma distribution by the shape parameter α and scale parameter β can be written as:

$$M_g(t) = (\frac{1}{1-\beta t})^\alpha = \prod_{i=1}^{N} \frac{1}{(1-\beta t)^{\alpha/N}} \quad (8)$$

Comparing equations (7) and (8) we estimate the MGF of variable $\xi(\boldsymbol{n})$ through that gamma distribution parameters:

$$\alpha = \frac{N}{2} \;\; , \beta = \frac{2}{N}\sigma_n^2 tr(\boldsymbol{D}_h^T \boldsymbol{D}_h) \quad (9)$$

Here tr($\boldsymbol{D}_h^T \boldsymbol{D}_h$) is the trace of matrix $\boldsymbol{D}_h^T \boldsymbol{D}_h$. Trace is a function in matlab software which represents the sum of the diagonal elements of matrix. For selecting weak textured patches, we name the null hypothesis as"the given patch is flat patch with the noise".Null hypothesis is acknowledged if the maximum eigen value of gradient covariance matrix is lower than some threshold. The threshold τ is given with the significance level δ and the noise level σ_n as:

$$\tau = \sigma_n^2 F^{-1}(\delta, \frac{N}{2}, \frac{2}{N} tr(\boldsymbol{D}_h^T \boldsymbol{D}_h) \quad (10)$$

Where F^{-1} (δ, α, β) is inverse gamma cumulative distribution function with shape parameter α and scale parameter β.In the weak textured patch selection algorithm, we choose the patches of the maximum eigen value of the gradient covariance matrix is lower than the

threshold value given in equation (10).The considerable level δ should be given manually for ex.0.99.

Table 1:Estimated noise level

Images	Noise level
Image_1	20.03
Image_2	20.43
Image_3	27.79
Image_4	29.74
Image_5	29.92
Image_6	32.75
Image_7	34.48

IV. NOISE REMOVAL TECHNIQUE

Scanning itself introduces noise in the document and to make it free from the noise, noise removal methods implemented.After estimating noise level we eliminate noise from images using non local means algorithm .Thisalgorithm use noise level calculated through Noise level estimation techniques as a filtering parameter to remove noise from images. It works on the self similarity assumption that is Adjacent pixels tend to have the similar neighborhoods, but non-adjacent pixels can also have similar neighborhoods in the noised image. The self-similarity assumption is exploited to denoise an image. This method is based on weighted average of the similar pixels in Historical image.[4] In proposed algorithm two windows are placed on image: one is the similarity window and the other is search window. In this first of all using similarity window, the similarity between the central pixel and all its neighbors in the area of search window, is calculated. Then calculate the average weight of like or similar pixels. Then weight of central pixel is replaced with that average weight. This process is repeat for every pixel of image. Proposed algorithm described as:

$$NL[u](x) = \frac{1}{c(x)} \int_{\Omega} e^{-\frac{G_a * |u(x+.) - u(x-.)|^2(0)}{\Delta^2}} u(y)dy \quad (11)$$

Where $x \epsilon \Omega$ and

$$c(x) = \int_{\Omega} e^{-\frac{G_a * |u(x+.) - u(x-.)|^2\ (0)}{\Delta^2}} dz \quad (12)$$

c(x) is a normalizing constant ,$G_a$is a Gaussian kernal and Δ work as filtering parameter.

Δ=degree of filter =estimated noise level (σ)

Δ i.e. used as a filtering parameter in noise removal technique is the output of the estimated noise level technique.computed as a weighted average of all the pixels in the image,

$$NL[v](i) = \sum_{j \in i} w(i,j)v(j) \quad (13)$$

Where weights depend on similarity between the pixels i and j

$$0 \leq w(i,j) \leq 1 \quad and \sum_{j} w(i,j) = 1$$

This similarity is measured using Euclidean distance and weight can be defined as:

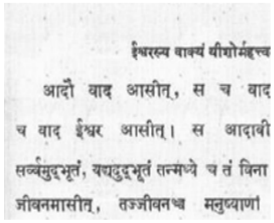
$$w(i,j) = \frac{1}{z(i)} e^{-\frac{\|v(N_i) - v(N_j)\|^2_{2,a}}{\Delta^2}} \quad (14)$$

the parameter Δ acts as a degree of filtering which is equal to the standard deviation of noise which is calculate by Noise estimation technique[6]. Consequently use estimated noise level and apply NL means algorithm [1] to remove noise from image while preserving contents of Historical document image.

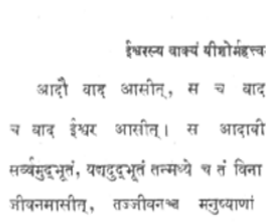
V. RESULTS

The performance of the proposed algorithms was evaluated in terms of the visual quality, the peak-signal-to-noise-ratio (PSNR) ,mean signal error (MSE).

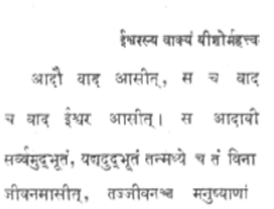
ORIGINAL IMAGE



EXISTING APPROACH



PROPOSED APPROACH



MSE: (Mean square error) Here it is just used to calculate the difference betweena original image with the restored image.Mean square error is given by:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{i=1}^{N} [g(i,j) - f(i,j)]^2 \quad (15)$$

Where M and N are the total number of pixels in the horizontal and the vertical dimensions of image, g denotes the Noise image and f denotes filtered image. The lowest mean square error represents the best quality image.

PSNR:(Peak signal to noise ratio) is measured in decibel (dB) and for Gray scale image it is defined as:

$$PSNR = 10log_{10}\left(\frac{255^2}{MSE}\right) \quad (17)$$

For the image quality measures, if the value of the PSNR is very high then is best quality image. PSNR calculation

of two images, one original and an altered image, describes how far two images are equal.

Table 2: Performance Metrices Based On PSNR And MSE

Images	Existing approach		Proposed approach	
	PSNR1	MSE1	PSNR2	MSE2
Image_1	37.22	12.31	37.51	11.53
Image_2	37.07	12.74	37.66	11.13
Image_3	36.83	13.57	37.16	12.50
Image_4	36.13	15.84	36.37	14.98
Image_5	34.80	21.51	34.99	20.56
Image_6	33.90	26.47	34.15	24.99
Image_7	33.42	29.54	33.63	28.14

The above shown results are all are the outputs of the noise removal technique used in order to get enhance and clear image of Historical document images. Above shown results shows that proposed technique gives better results than the existing technique [4].

Fig 2: a.) original noisy text image b.) non local means image c.) non local means image based on estimated noise level

Through the images we can clearly see that proposed approach removes noise better from the scanned noisy document image .

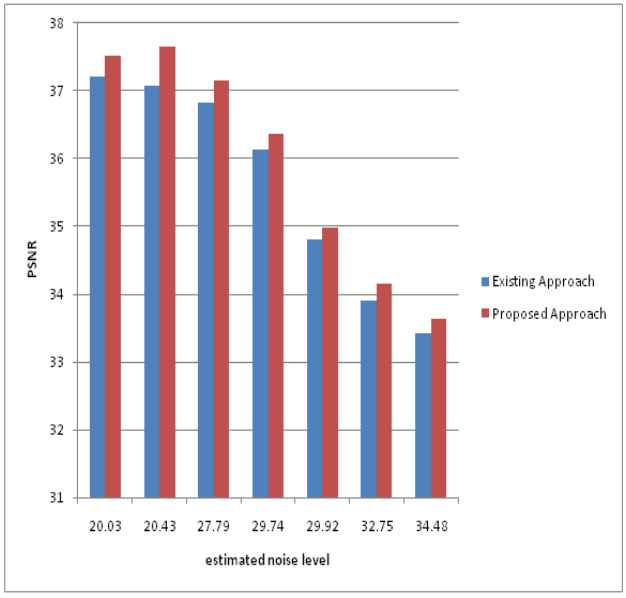


Fig 3: Graphical representation of PSNR

This graph shows that PSNR of proposed approach is better than the PSNR of existing approach.

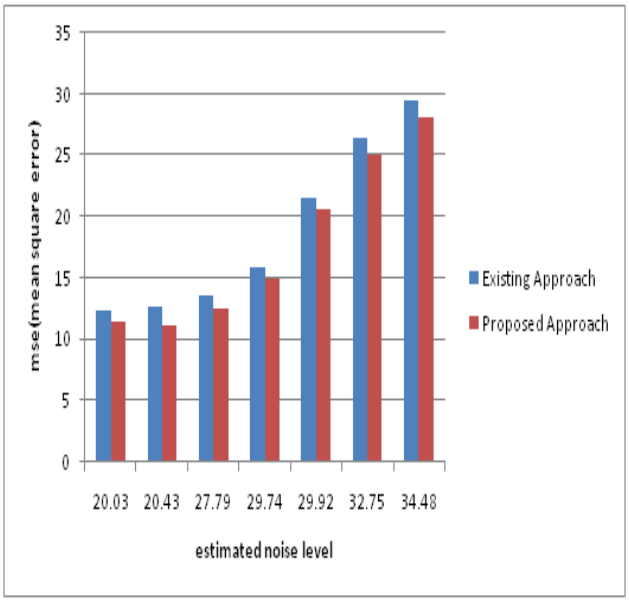


Fig 4: Graphical representation of MSE

MSE of proposed approach is lower than the existing approach. And the images having low MSE values are better images or enhanced inages.

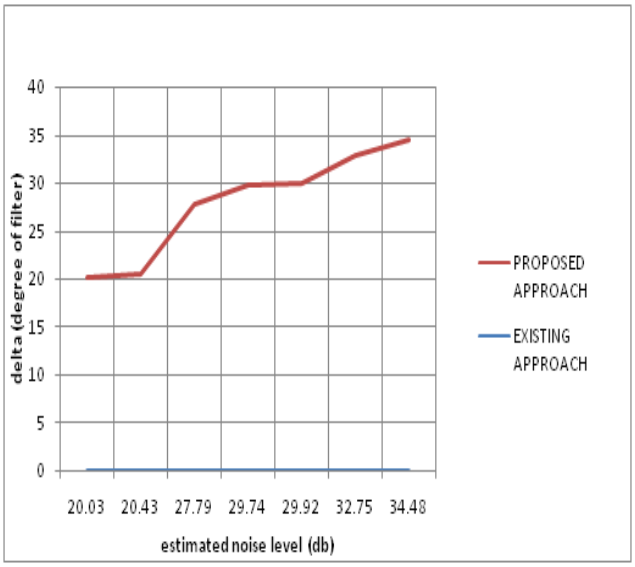


Fig 5: Graph of existing approach filtering parameter and proposed approach filtering parameter

This graph shows existing approach filtering parameter i.e delta as 0.02 for all the images but in proposed approach filtering parameter i.e delta changes in every image because delta depends on estimated noise level.

VI. CONCLUSION

This paper presents a system that enhances the quality and readability of Historical document through noise detection and noise removal methods.Noise level estimation technique estimates noise present in image and

Non Local Means Agorithm [1]removes noise from image using that estimated noise level. Non Local Means algorithm provides outstanding results in denoising images and also preserves edges, lines and fine details present in image But this techniques is very time consuming. Noise removal techniques have been successfully tested on document images.

Experiments shows best result for Non- local means filter based on estimated noise level rather than the existing approach which is non local means filter .

REFERENCES

[1] A. Buades, B. Coll, J.M. Morel, "A non-local algorithm for image denoising," : IEEE CVPR,2005,pp. 60–65.

[2] D.N. Satange, S. Bobde and S. Chikate, "Historical Document preservation using image processing technique" International Journal of Computer Science and Mobile Computing, 2013, pp .247-255

[3] L. Likforman, J. Darbon and E .Smith" Enhancement of Historical Printed Document Images by Combining Total Variation Regularization and Non-Local Means Filtering." Boise State University ScholarWorks ,2011,pp.1-41.

[4] K. Balakrishnan, K. Sunil, A.V Sreedhanya, and K.P Soman." Effect Of Pre-Processing On Historical Sanskrit Text Documents" International Journal of Engineering Research and Applications (IJERA) ,2012 ,pp.1529-1534.

[5] R.C. Gonzalez, R. E. Woods, and S. L. Eddins,.Digital
image processing using Matlab.2nd Edition.

[6] X. Liu, M. Tanaka and M .Okutomi," Noise Level Estimation Using Weak Textured Patches of A single noisy image." *:IEEE* ,2012,pp .665-668.

[7] P.Dubey, S.S. Sharma," Noise estimation and removal through principal component analysis in a segmented singly image,": international journal of electronics,Communication & instrumentation engineering Research and development (ijecierd),2013,pp.51-56.

[8] S. Kaur, R. Singh, "Noise estimation and removal from gray-scale image using non- local means algorithm.": International journal of advanced research in Computer science and software engineering, 2015,pp.1461-1467.

[9] B.R. Prasad, S. Choudhary, "Survey paper on different approaches for noise Level estimation and denoising of an image": international journal of science and research (ijsr), 2014,pp.618-622.

[10] S. Pyatykh, J. Hesser, and L. Zheng, "Image noise level estimation by principal component analysis.": IEEE Trans. Image Process., 2013,vol. 22, no. 2, pp. 687–99.

[11] D. Raghuvanshi, D. Jain , P. Jain, 2013. " Performance Analysis of Non Local Means Algorithm for Denoising of Digital Images ": International Journal of Advanced Research in Computer Science and Software Engineering,2013,pp.94-100.