Denoising Epigraphical Estampages Using Nested Run Length Count

P. Preethi(✉), K. Praneeth Kumar, M. Sumukha, and H. R. Mamatha

PES University, Bengaluru, India
{preethip,mamathahr}@pes.edu,
praneeth.kumar6699@gmail.com, sumukhamohan6@gmail.com

Abstract. Denoising in epigraphical document analysis helps in building recognition system for fast and automatic processing. However, it is challenging due to the presence of stone texture as a complex background in input samples. In this paper, a nested run length counting with varying block size of 3 * 3, 5 * 5 and 7 * 7 are applied. Computation is carried out on neighboring pixels of the point of interest and discloses whether it is part of the script on inscription or background based on the count value. If it is part of the background, point of interest is set to background value else set to white. The method is tried and tested on 100 samples of epigraphical Estampages collected from archaeological survey of India. A comparative study is derived on the output of the proposed method and on the nonlinear filters such as median and wiener. Human vision perception has evaluated that proposed method is better than median and wiener filters. The quality measures such as Peak signal to noise ratio and Structural similarity indexes are practiced on the sample output for various filters and proposed method.

Keywords: Epigraphical scripts · Run length count (RLC) · Denoising · Peak to signal noise ratio (PSNR) · Structural similarity index (SSIM)

1 Introduction

History and culture are derived from ancient inscriptions and are scribed using regional script. In India inscriptions are available from 3rd to 17th century in the form of stone carvings (immovable), copper plates, coins and pots. The script used on the inscriptions varies from Brahmi, Hoysala & elongated Hoysala, Poorvada Halegannada, Halegannada and Nadugannada [13]. The processing and preservation of inscriptions are in the hands of Archaeology Department. In the process of preservation, Estampages (a rubbed copy of the stone) is formed on a thick sheet of paper applying black ink and same is circulated among Epigraphers for deciphering. Epigraphers are the philosophers who could read and understand the writings on the inscriptions. Finding epigraphers for reading/understanding has become difficult as they are few or extinct. Estampages are stored for future research, which are prone to damage due to environmental conditions. Digitization of Estampage as an image to preserve and automating the reading process is the need of today.

N. Kumar and R. Venkatesha Prasad (Eds.): UBICNET 2019, LNICST 276, pp. 198–206, 2019.
https://doi.org/10.1007/978-3-030-20615-4_15

Digitization deals with storing and processing of scanned Estampages, which are the input image for the automatic reading. The scanned images possess irregular rock structure traces, ink spots and distortion as noise. These noises are random in nature and pose a great challenge for further processing. Elimination of noise and isolating the script from the background texture is the focus of the research. Denoised Estampage images are used for segmentation, feature extraction and recognition.

Preprocessing techniques turns out to be a promising step in optical character recognition as the procedure removes outliers, distortion, binarization, low contrast etc., which supports segmentation, feature extraction and recognition processes. Low level entity noise has to be removed from the image to ensure only object of interest is existent. Denoising images involve standard filtering techniques like frequency domain filters and spatial domain filters which are applied to remove Salt and Pepper, Speckle and Gaussian noise. On the grounds of noise level, different types of filters are applied.

Run length encoding or counting is a compression technique which keeps track of same data value in consecutive elements around the pixel of interest and stores the data value along with the count. Keeping track of data values in the consecutive elements describes the feature of the image. In this paper, instead of tracking horizontal or vertical runs, the neighboring data values are tracked and similar data value count are kept in track. This information designates pixel of reference as actual data or noise. Nested Run length counting with varying size kernel of 3 * 3, 5 * 5 and 7 * 7 is implemented. The detailed explanation to the procedure followed is described later.

The paper describes the challenges in Estampage images, the literature survey in understanding different techniques and estimators used in Sect. 2. Section 3 illustrates procedure followed to implement denoising technique. The last section describes the results and discussion.

2 Literature Survey

In the images of Estampages there is no fixed or defined noise, as the texture and ink spitting, mistakes while writing is also considered as the noise. Some of the experiments were conducted on historical documents with similar kind of noises. Below are the survey detailing techniques and noise estimators used in validating the denoising.

In [1] two stage sequential 10 * 10 sized block filtering is applied. The first stage is to flush out the small noise cluster whereas second step is to remove large noise cluster which are now a part of the background. The output of the denoising technique is used as input for the gradient based segmentation. The authors of [6], experimented by applying the morphological filtering techniques on the segmented input for further processing. The inputs considered for experimentation are degraded historical documents.

Arabic historical documents [10] are subjected to contrast stretching filter and manual line by line pixel tracing technique in removing the noise. In [11], handwritten historical document is considered for the noise removal technique. Application of hybrid Iterative global thresholding to remove noise in the Greek historical documents [8], Application of median filters in [9] for filtering Tamil inscriptional images are discussed. The old palm scripts are also considered to be the historical manuscripts, in preserving and processing them, [2] a set theory based morphological operations are applied and the output is compared against the standard filtering techniques such as wiener and median filters. Copper inscriptions are exposed to nonlinear filtering techniques in [7]. The filters like median, harmonic and contra harmonic filters are applied on the images of copper plates and subjected to estimators such as PSNR and SSIM.

Run length coding is applied on the Chinese tablet to remove noise in the image. The rubbings of tablet are considered as the input, where as a preliminary step the image subjected to smoothening and binarization applying Otsu method. A single character from the tablet is the input and the five basic strokes of the Chinese character are considered for calculating the statistics. It is also assumed that the noise is smaller than the strokes. Using horizontal and vertical scan the noise density is estimated [3]. Horizontal and vertical run length count method is used to remove noise in document analysis. The result of the noise removal method is the input for Optical character recognizer. The author validates the efficiency of the algorithm on handwritten document and printed epigraphical images [4].

Understanding the noise characteristics, estimating the noise level and conclusion plays a key role. In [5], the author describes about the existing noise measures with respect to pixel difference based, Correlation based, edge based, context based, spectral distance based and human vision based measures. Human vision system is considered to be the best metric and is implemented by structural similarity index and proves to be the best out of all in his discussion with some example. In [12], author theoretically understand the similarity and difference between peak signal to noise ratio and structural similarity indexing, and finds that PSNR can be derived from SSIM and vice versa.

Survey reveals variant methods applied on different set of input including handwritten documents, manuscripts, images of inscriptions and epigraphical scripts. Each document has dissimilar noise and processed with linear, nonlinear and block filters, noise being estimated using PSNR, SSIM, MSE measures. Existence of random noise, ink spots, stone texture impression in the images of Estampage pose an open challenge in the field of document analysis.

3 Proposed Method

There are many examples of RLC compression technique applied to reduce image dimension. Compression technique cannot be practiced on epigraphical script as reduction of pixel may remove the details of the image. Use of Run length count on the neighboring pixels can describe information such as noise, edge, text and background. Keeping this information, a novel technique - nested run length count algorithm with varying kernel size is designed and applied.

It is noted from the above Fig. 1 and also from the discussion that, there exist a randomized noise and each input image exist with different challenge.

Run length count algorithm is applied on the horizontal and vertical lines of the image. We have proposed a different version where RLC algorithm will work on block of neighboring pixels of greyscale image. The block size is varied from 3 * 3, 5 * 5 and 7 * 7 to achieve better results. An odd sized kernel is intentionally selected as it

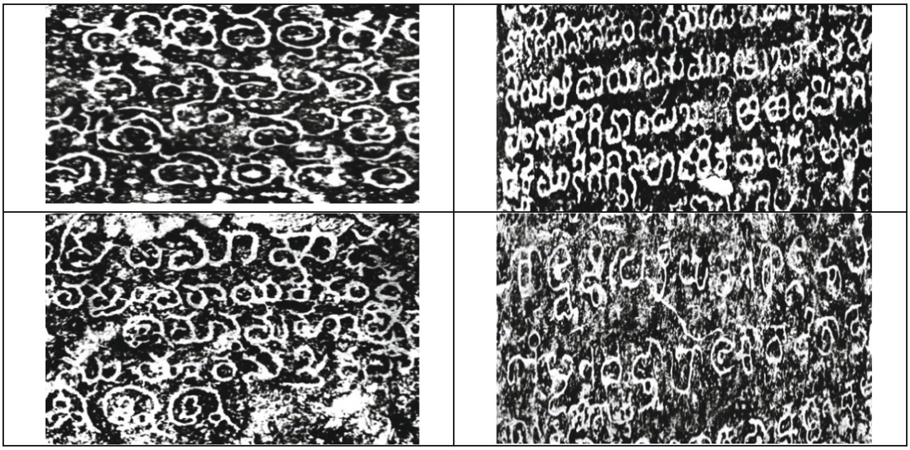
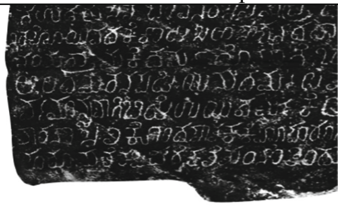
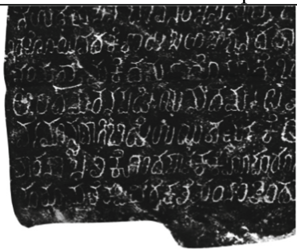
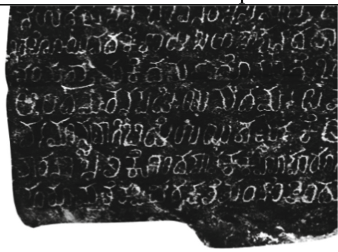
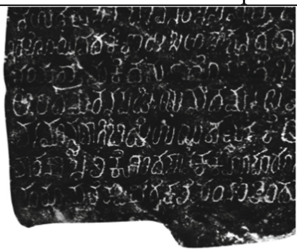


Fig. 1. Depicting the input images majorly having background texture of stone as noise which has considered as preliminary challenge [ASI].

covers all the first, second and third circle neighboring pixels around the pixel of interest. Initially threshold value (by averaging the means) of the image is computed which will be used as a comparative value in deciding the noise level in the neighboring pixels. The marginal pixels of the image are evaluated by considering the out of bound indexing.

Raw input image is first subjected to run length coding algorithm for varying size of the blocks. Each pixel is subjected under 3 * 3 block, covering 9 pixels, where middle pixel pointed by block represents point of input and its remaining 8 pixels are the deciders. Based on computed run length count the value of input pixel is represented as white or black. If white pixel count is maximum than black pixel among 8, the input pixel value is set to white else set to black. The outcome of the 3 * 3 pixel was poor. Hence, changing block size to 5 * 5, 7 * 7 and 9 * 9 was adopted to improvise the Denoise ratio. When 9 * 9 sized kernel was applied on the input, it was removing even the script details. The result of 7 * 7 was comparatively better than 3 * 3, 5 * 5 and 9 * 9. Another problem is, if the number of black pixel is equal to the number of white pixel, then deciding the value of the experimental pixel is tough and random assumption would lead in creating salt and pepper noise. Experimental results are represented in Fig. 2 along with the quality measures.

In proposed nested block model, input pixel value (which is the point of interest) is first compared with 8 neighbors of 3 * 3 block size, if there exists a problem in deciding input pixel value, the same input pixel is subjected for nested experimentation with 5 * 5 block size counting 25 neighboring pixels. The method is continued till block size of 7 * 7 counting 49 surrounding pixels for the conclusion. Experimentation with three nested Run Length Counting has proved with good results and are shown below in Table 1.

Input image	3*3 block sized RLC output
	
5*5 block sized RLC output	7*7 block sized RLC output
	
9*9 block sized RLC output	
	

Size	PSNR	SSIM	MSE
3*3	26.15	0.9591	1.4511
5*5	25.4843	0.9584	1.5013
7*7	24.9958	0.9486	2.0807
9*9	23.14	0.9341	2.0607

Fig. 2. Depicting result of 3 * 3, 5 * 5, 7 * 7, 9 * 9 block sized RLC output along with the PSNR, SSIM and MSE measures for a sample input.

Pseudo code of the Nested RLC is depicted below:

```
I(x, y) is raw input image
Thresholding (I(x, y))
RLC3*3(I(x, y))
If (Count Length with 8 neighbors decides value of I (x, y)) return;
Else RLC5*5(I(x, y))
    If (Count Length with 24 neighbors decides value of I (x, y)) return;
    Else RLC7*7(I(x, y))
        Count Value with 48 neighbors decides the value of I(x, y)
```

The noise estimators are used to validate the work conducted on Estampages and the definition of PSNR, MSE and SSIM are as follows.

Peak signal to noise ratio is the ratio of power of the signal to power of the distorting noise which deteriorates the quality of the image. It is considered as one of the standard noise estimators in both lossless and lossy compression techniques. PSNR value is derived from the mean squared error which is the average squared difference between the estimated values [15].

$$\mathrm{MSE} = \frac{1}{mn}\sum\nolimits_{i=0}^{m}\sum\nolimits_{j=0}^{n}[I(i,j) - K(i,j)]2 \tag{1}$$

Table 1. shows the output of the nested RLC method

INPUT IMAGE	Measures	OUTPUT IMAGE
	PSNR: 21.5825 MSE: 4.9351 SSIM: 0.92223	
	PSNR: 25.2799 MSE: 3.8771 SSIM: 0.9305	
	PSNR: 22.8620 MSE: 3.7935 SSIM: 0.9358	
	PSNR: 21.5585 MSE: 23.3733 SSIM: 0.8742	

For an image I with m * n dimensions, and noise approximation image K of same size as input image. MSE is computed as in Eq. 1, PSNR as in Eq. 2 where MAX is the maximum possible pixel value in the image.

$$PSNR = 20\log_{10}(MAX) - 10\log_{10}(MSE) \tag{2}$$

Image quality assessment is also measured in terms of structural similarity index and it is a perception based method and hence closer to human vision system. It is a measure between the two images x and y of size m * n, and SSIM is computed as in Eq. 3 Using which the quality of the output images are estimated. μ_x and μ_y are the average of pixel in image x and y where variance and covariance are represented as σ_x, σ_y and σ_{xy} [14].

$$SSIM = \frac{\left(2\mu_x\mu_y + C_1\right) + \left(\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{3}$$

Table 1 showcase the input and output images along with computed value of estimators. Input images are also subjected to nonlinear a well-known median and wiener filters and PSNR, SSIM values are recorded. The results are discussed in next section.

4 Results and Discussion

As noise reduction and automation of epigraphical Estampages are new topic of interest, the data samples are collected from the Archaeological survey of India, Mysore. The images considered are of old Kannada Estampages which depicts background stone texture as major noise. These are non-standard random errors and cannot be removed completely by standard filters. The proposed method is compared with median and wiener filters and the results are as shown for 10 images in Table 2.

Table 2. Showing the outcome of the measures applied on the filtered image

Sl. no	Median filters			Wiener filters			Nested RLC filtering		
	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM
1	18.618	26.61	0.942	25.808	29.336	0.952	**4.539101**	**21.58246**	**0.922325**
2	21.890	26.17	0.939	27.351	28.917	0.950	**5.238516**	**21.89441**	**0.9166**
3	21.684	26.92	0.945	27.660	29.045	0.952	**5.508122**	**22.06989**	**0.916554**
4	20.821	26.38	0.943	28.398	28.792	0.952	**5.058379**	**21.89571**	**0.920653**
5	29.876	24.62	0.923	36.8073	27.115	0.939	**9.242608**	**19.64612**	**0.858684**
6	28.376	25.21	0.931	35.0153	27.571	0.943	**8.612306**	**20.18732**	**0.870931**
7	19.535	26.55	0.939	25.8011	29.352	0.949	**4.854342**	**22.02918**	**0.920685**
8	23.139	26.43	0.942	28.538	28.744	0.950	**5.788686**	**21.74733**	**0.910441**
9	15.788	27.29	0.941	21.084	30.444	0.952	**3.793502**	**22.86201**	**0.935805**
10	21.050	27.63	0.950	21.165	30.424	0.958	**5.201315**	**23.1605**	**0.920123**

The experiment is conducted for 100 samples and for all inputs the proposed method has given good results, than median and wiener filters. Human vision could observe better noise removal in proposed method than by standard techniques. The proposed method was also tested on well-known inputs with Gaussian noise and the output is as shown in the Fig. 3. Outcome of the proposed method appears to be distorted.

			
Noisy Leena input	Image after Nested RLC	Noisy Cameraman input	Image after Nested RLC

Fig. 3. The outcome of Nested RLC on Leena and Cameraman images with Noise level of 0.025.

5 Conclusion

A novel method of using Run length counting is used for denoising the epigraphical Estampages to preserve the foundation of history. The proposed method uses a simple 3 * 3, 5 * 5, 7 * 7 neighbors to eliminate the noise based on the run length count of white and black pixels. Application of 9 * 9 sized block had given noisy and inappropriate result when compared with inputs and hence stopped at 7 * 7. The method was tested on 100 samples and could achieve average PSNR as 23.75 and average SSIM of 92.83%. The proposed method is able to remove small noise clusters around the script but the prominent noises are still seen. Our future plan is to extend the preprocessing methods for better filtering of Estampage images.

Acknowledgement. My sincere thanks to The Director and staff members of Archaeological survey of India, Mysore for providing data samples for the research work.

References

1. Baig, A., Al-Ma'adeed, S.A.S., Bouridane, A., Cheriet, M.: Automatic segmentation and reconstruction of historical manuscripts in gradient domain. IET Image Process. **12**, 502–512 (2018)
2. Gangamma, B., Murthy, K.S., Singh, A.V.: Restoration of degraded historical document image. J. Emerg. Trends Comput. Inf. Sci. **3**(5), 792–798 (2012)
3. Zhang, J., et al.: Denoising of Chinese calligraphy tablet images based on run-length statistics and structure characteristic of character strokes. J. Zhejiang Univ. Sci. **7**, 1176–1186 (2006)

4. Karthik, S., Mamatha, H.R., Murthy, K.S.: An approach based on run length count for denoising the kannada characters. Int. J. Comput. Appl. **50**(18), 42–46 (2012)
5. Al-Najjar, Y.A.Y., Soong, D.C.: Comparison of image quality assessment: PSNR, HVS, SSIM, UIQI. Int. J. Sci. Eng. Res. **3**(8), 118–126 (2012)
6. Boudraa, O., Hidouci, W.K., Michelucci, D.: A robust multi stage technique for image binarization of degraded historical documents. In: 5th International Conference on Electrical Engineering. IEEE (2017)
7. Chairy, A., et al.: Image restoration on copper inscription using nonlinear filtering and adaptive threshold. J. Phys: Conf. Ser. **801**, 012043 (2017)
8. Kavallieratou, E., Stamatatos, E.: Improving the quality of degraded document images. In: The Second International Conference on Document Image Analysis for Libraries (DIAL 2006). IEEE (2006)
9. Janani, G., Vishalini, V., Kumar, P.M.: Recognition and analysis of Tamil inscriptions and mapping using image processing techniques. In: International Conference on Science Technology Engineering and Management. IEEE (2016). ISBN 978-1-5090-1706-5
10. Elfattah, M.A., Hassanien, A.E., Mostafa, A., Ali, A.F.: Artificial bee colony optimizer for historical arabic manuscript images binarization. In: 2015 11th International Computer Engineering Conference. IEEE (2015). ISBN 978-1-5090-0275-7
11. Chakraborty, A., Blumenstein, M.: Preserving text content from historical handwritten documents. In: 2016 12th IAPR Workshop on Document Analysis Systems. IEEE (2016). ISBN 978-1-5090-1792-8
12. Horé, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition DBLP (2010)
13. Soloman, R.: Indian Epigraphy, 2nd edn. Oxford Indian Press (1999). ISBN 0-19-509984-2
14. Keiran, G.: Structural similarity index simplified. Occasional Texts in the Pursuit of Clarity and Simplicity in Research. Series 1, no. 1 (2015)
15. http://www.ni.com/white-paper/13306/en/