ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 1; Page No. 59-63

Preservation of Historical Documents by Using Image Thresholding: A Review

Lovendra Solanki[1], Praveen Kr. Sharma[2], Ankit Sharma[3]

[1]Assosiate Professor, B K Birla Institute of Engineering and Technology, Pilani, Rajasthan, India

Dean.ecee@bkbiet.ac.in

[2]Assistant Professor, B K Birla Institute of Engineering and Technology, Pilani, Rajasthan, India

praveen.sharma@bkbiet.ac.in

[3]Student- ME, B K Birla Institute of Engineering and Technology, Pilani, Rajasthan, India

ankit.trex-@live.com

ABSTRACT

History is the thing which enables us to know about the past that has made our present and which will deliver us our future. This history is preserved in the form of documents which can be epics such as Ramayana, pacts preambles etc. We can have the documents but they are not easily available also this information is not safe in the paper form as the paper documentation have many disadvantages such as fragility of paper, fading of ink and different kinds of noises. So digitisation is required to get a permanent copy of the document. The advantage of digitisation is to provide almost permanent and long lasting information which won't fade over time. In this paper we are discussing about the different types of techniques that are used for the digitization of the documents that are given by different scientists and comparison among them. Some of the techniques that are discussed in the paper are Genetic, Niblack, Mean gradient technique etc. that allows us to know more about the quantization and binarization for enhancing the image quality.

Keywords: Binarization, Thresholding Quantization.

INTRODUCTION:

Digitization of document

Historical documents have an age i.e. the paper is very delicate thing to have important documents over it and is susceptible to get wear and tear over a long period of time.

Thus for making a document easily accessible, digitization is the most effective technique for example digitizing image we can visualize document more easily and provide a simple way to produce more and more copies. In this the true color is maintained in the high resolution. Another advantage of digitization is, as the image is directly loaded, the storage space is directly reduced. Thresholding is the technique which is also known as binarization that produce an image in a bi- level format and is stored in GIF file format which is five times less than that of original true color JPEG format.

This binarization technique is effective but there are some problems in the images which lead to the difficulties in proper thresholding of document images. These problems can be termed as noises which can be of several types such as degradation of paper, yellowing the sheet or fading of ink and another problem is both side writing of the paper, in which the ink starts interfering on the both sides of the paper majorly known as "ink bleeding"[1].

By all the difficulties it is very hard to find the best threshold value that could separate the color of the ink from that of the paper or one could say, background.

Quantization of image

When significant set of colors are selected to represent an image the whole process of selection of those set of colors is known as quantization and when this process is attempted on true colors the process becomes color quantization.

We have used quantization because it works upon the logical image and enhances the pixel values or we can say quantizes the pixel values.

When color quantization is done then the resolution of the image is set to a lower level or we can say that it is decreased to a specific quantity, if the technique is not applied with high precision the quality of the image can be drastically reduced.

There are two classes of color quantization: spitting algorithm and clustering algorithm. In spitting algorithm image's color space is divided into disjoint cells until it

*Corresponding author: Lovendra Solanki

reaches the desired number of sets for example popularity algorithm proposed by Heckbert in 1982. In clustering based algorithm the colored space of the image is clustered into k number of clusters. C means algorithm was proposed by parker in 1997 is an example of clustering algorithm.

Another quantization technique which is known as dithering.

This process can also be used in the reduction of color space. Well the thing to remember is Thresholding or binarization is also the technique for color quantization which leads the image to be in two colors displaying just black or white.

Image Fidelity Assessment

There are two basic requirement of image quality: usefulness and naturalness.

Usefulness is defined to be the discreminality of image content and naturalness is the identification of image content. The image fidelity can be defined as Mean Opinion Score (MOS) and Peak Signal 2 Noise Ratio (PSNR) and Mean Square Error. The formula is

$$\mathbf{PSNR = 10 \log_{10} (c^2 / MSE)} \quad (1)$$

Where, c is the maximum colour value. A fidelity index, *Q*, is defined in (Wang & Bovik, 2001) in terms of the linear correlation

Coefficient and the similarities between the mean and variance of two images. This index is defined as:

$$Q = 4u_x u_y \sigma_{xy} / (u_x^2 + u_y^2)(\sigma_x^2 + \sigma_y^2) \quad (2)$$

x= value of original image y= value of tested image where, Q (-1, 1) and it is one when both the images are same.

Implementation Techniques: The techniques that are given by different scientists in the field of image thresholding are as follows:

Algorithms used for Thresholding

1. Genetic Thresholding Algorithm: Main objective of this method is to provide a bi-level image from a historical document. The target image is reduced to palette and remaining colours represent classes for example text, background, and interface. Dark grey level composes text on the other hand light grey level describes background. By the Thresholding the classes are divided into text (darkest) and remaining colour as non-text [8, 9].

2. Niblack's Method : This method is based upon calculating the local mean and local standard deviation i.e.;

$$T(x, y) = m(x, y) + k.\, s(x, y) \quad (3)$$

Where, m(x, y) and s(x, y) shows the average of local mean and standard deviation values of image. The size should be small of local area for detail preservation but it should be large enough for noise suppression [10].

3. Mean Gradient Technique: This is the advancement of the Niblack's method and in this a constant is introduced which is used for grey value. And if constant i.e. (max grey level – min grey level) >= R then,

$$T(x, y) = M(x, y) + K.G(x, y) \quad (4)$$

Else $T(x, y) = .5\, M(x, y)$ (5)

4. Background Subtraction: In this method handwriting is removed from original image separating it from background. This is effective in removing the dark characters that are darker than the background and then background is subtracted from the original document image only leaving handwriting behind [14].

5. Quadratic Integral Method: It is known as two stage thresholding approach. In first step foreground background and fuzzy classes are determined and in second step the threshold value is searched and that threshold value is then used to change the image into bi level grey scale image [14].

6. Yanowitz and Buckstein's Method: They used the grey level values at high gradient regions as known data to interpolate threshold surface of the image. And another important feature is that it uses image segmentation and integration to produce the final binarized image [12].

Results of Genetic Algorithms

The assumption of grey level is a necessity for the quantization to obtain best results. As the number of grey levels increases the Q factor which is compared with original grey levels increases and vice versa.

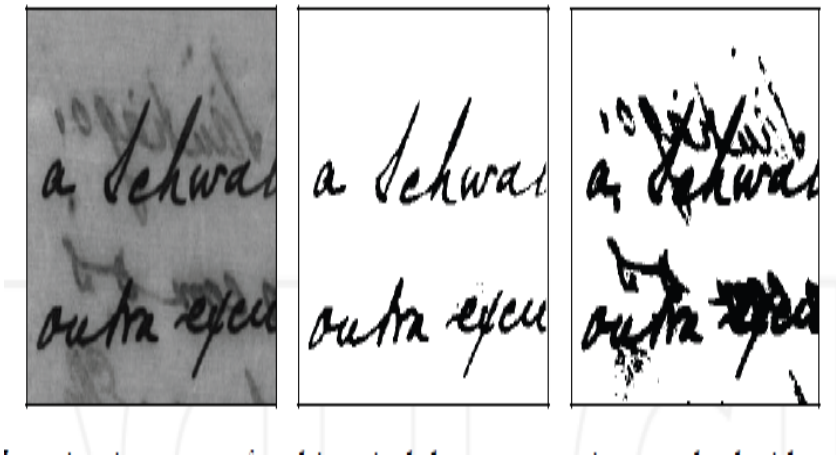

Figure 1: Noises [11]

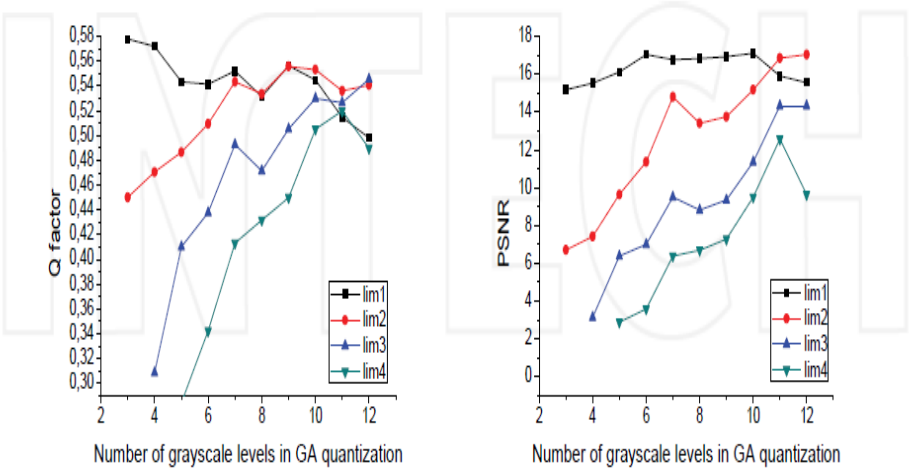

Figure 2: Quantization Results [11]

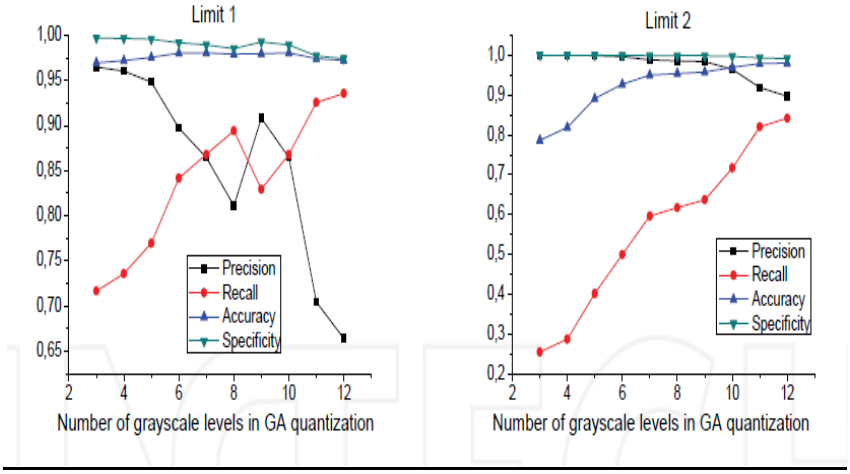

Figure 3: Recall and Precision at different limits [11]

Algorithm	Precision	Recall	Accuracy	Specificity	PSNR	MSE
GA	0.8290	0.8273	0.9851	0.9694	21.5167	0.0306
Bernsen	0.839	0.8135	0.9636	0.9847	21.4086	0.0364
Brink	0.8972	0.7176	0.9378	0.9898	20.5486	0.0622
C-Means	0.9656	0.4752	0.7276	0.9919	15.5124	0.2724
daSilva-Lins-Rocha	0.8882	0.7277	0.9459	0.9908	20.3931	0.0541
Fisher	0.9085	0.7274	0.9329	0.9903	20.5406	0.0671
Huang	0.9174	0.6946	0.9146	0.9904	19.8345	0.0854
Johannsen	0.9398	0.6548	0.9317	0.9942	19.3038	0.0683
Kapur	0.0148	0.4073	0.9021	0.9013	16.8976	0.0979
Kittler	0.8587	0.7982	0.9429	0.9847	21.1809	0.0571
Li-Lee	0.872	0.7313	0.9447	0.9899	20.3123	0.0553
Mean Grey Level	0.9116	0.7518	0.9654	0.9891	21.11	0.0346
Niblack	0.862	0.3911	0.8531	0.985	14.6877	0.1469
Otsu	0.8431	0.7128	0.9422	0.9683	19.1462	0.0578
Percentage of Black	0.9902	0.2156	0.6165	0.9985	10.3754	0.3835
Pun	0.941	0.6513	0.9395	0.9957	19.3791	0.0605
Renyi	0.9894	0.1716	0.3683	0.9011	8.6996	0.6317
Sauvola	0.8925	0.1888	0.6045	0.9755	10.0915	0.3955
Iterative Selection	0.3368	0.3356	0.9418	0.9514	18.9542	0.0582
TwoPeaks	0.9402	0.6288	0.9236	0.9947	18.7636	0.0764
White	0.3855	0.757	0.7348	0.9524	15.6284	0.2652
Wu-Lu	0.7792	0.7711	0.9411	0.9595	19.7098	0.0589
Yager	0.7985	0.2502	0.7475	0.9609	12.0418	0.2525
Yen	0.5276	0.8058	0.9358	0.944	19.0102	0.0642

Figure 4: Results of Different Thresholding Algorithms [11].

Application Areas:

The application area of image thresholding is wide as it can be used in the detection of errors of documents and preserve the information, for the detection of codes on confidential documents plus it provides and aid for the wider scopes in different fields other than information technology such as medical, military services, and most importantly history.

Future scope

With the regular improvement in the technology and newer thresholding algorithms, the image thresholding of documents have become very effective but there are still some areas that can be worked upon such as the time efficiency of thresholding. Second, the letters that are read are unbiased for example there are no capital or small letters so capital letter recognition can be another task to be worked upon, then there should be a dictionary in different office tools about the letters to be recognised. So thresholding is having a great scope for preservation of documents in future.

CONCLUSIONS

The final finding of image thresholding depends upon two major terms.

1. Precision: it is defined as the ratio of correctly detected words to the total detected words.

$$\boldsymbol{precision = \frac{correct\ words}{total\ detected\ wods}}$$

2. Recall: it is defined as the ratio of correctly detected words to the total words.

$$\boldsymbol{recall = \frac{correct\ words}{total\ wods}}$$

Then different results can be obtained by applying these Thresholding methods on different documents such as newspapers, many document images etc.

The Yanowitz and Buckstein's method [12] was taken to be the best binarization method. However, the cost of the thresholding was high as it worked upon the segments of the images. Apart from this the background elimination provides extra or better precision and recall as the handwriting is taken out by subtracting the background. In the end what we can say is the future depends upon the combination of binarization/ thresholding techniques to provide the best results for digitization and prevention of the historical documents.

REFERENCES:

1. Mello, C.A.B.; Oliveira, A.L.I.; Sanchez, A. Historical Document Image Binarization, *Proceedings of the International Conference on Computer Vision Theory and Applications*, pp. 108-113, ISBN 9789898111210, Funchal, January 2008, INSTICC, Portugal.
2. *Proceedings of the International Conference on Image Processing and its Applications*, pp.86-90, ISBN 085296692X, Ireland, July 1997, IEEE Press, New Jersey.
3. Janssen, T.J.W.M. (2001) Understanding image quality, *Proceedings of the International Conference on Image Processing*, pp. 7, ISBN 0780367251, Thessaloniki, Greece, October 2001, IEEE Press, New Jersey.
4. J. M. White and G. D. Rohrer, "Image Thresholding for optical character recognition and other applications requiring character image extraction", *IBM J. Res. Devel*., Vol. 27, No. 4, 1983, pp.400-411.
5. Y. Liu and S. N. Srihari, "Document image binarization based on texture features", *IEEE Transactions on PAMI,*vol. 19 no. 5, May 1997, pp. 540-544.
6. M. Junker, R. Hoch, "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy", Proc. Of 5th ICDAR, India, pp. 713-716, 1999.
7. Parker, J.R. (1997). *Algorithms for Image Processing and Computer Vision*,John Wiley and Sons, ISBN 0471140562, New York.
8. Freisleben, B.; Schrader, A. (1997). Color Quantization with a Hybrid Genetic Algorithm.
9. Mitchell, M. (1998). *An Introduction to Genetic Algorithms*, MIT Press, ISBN 0262631857, Cambridge.
10. W. Niblack, "An introduction to digital image processing", pp. 115-116.
11. McMillan, N.A.; Creelman, C.D. (2005). *Detection Theory*. LEA Publishing, ISBN 0805842306, New Jersey.
12. D.L. Milgram, A. Rosenfeld, T. Willet and G. Tisdale, "Algorithms and hardware technology for image recognition", *Final Report to U.S. Army Night Vision Laboratory*, 1978.
13. Carmelo Basto Filho, Carlos Alexendre Melo, Julio Andrade, Marilia lima, Willington Dos Santos, Adriano Oliveira and Davis Falcao, "Image Thresholding of historical documents based on genetic algorithm", *Department of computing systems, University of Pernambuco, Brazil.*
14. Ghraham Leedham, Chen Yan, Kalyan Takru, Joie Hadi Nata Tan and Li Mian, "Comparison of some Thresholding algorithms for text/background segmentation in difficult document images", school of computer engineering, Nanyang Technical University, Singapore.