IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2, November 2011
ISSN (Online): 1694-0814
www.IJCSI.org 

# Cleaning of Ancient Document Images Using Modified Iterative Global Threshold

**N.Venkata Rao[1] A.V.Srinivasa Rao[2] S. Balaji[3] and L. Pratap Reddy[4]**

**[1] Department of Electronics and Computers Engineering**
**Research Scholar, KL University, Vaddeswaram – 522502, Andhra Pradesh, India**

**[2] Department of Electronics and Communication Engineering,**
**Research Scholar, JNTUH, Hyderabad - 500085, Andhra Pradesh, India**

**[3] Department of Electronics and Computers Engineering**
**KL University, Vaddeswaram - 522502, Andhra Pradesh, India**

**[4] Department of Electronics and Communication Engineering**
**JNTUH College of Engineering, Hyderabad – 500085, Andhra Pradesh, India**

## Abstract

Ancient document Image processing is an important area attracting many researchers in the recent period. Binarization is the first step while cleaning the document for further processing. Based on the degradation of the original document, either global or local thresholding methods are preferred. Thresholding phenomenon is a simple and practical approach to identify the cluster of pixels that are most likely associated with background information, while separating the object information. In this paper we propose a modified iterative global thresholding approach to separate the clusters of fore ground and background. The relative closeness towards background intensity is computed in each iteration after image equalization. Camera captured images of ancient printed documents, stone carvings and palm leaves are evaluated in the present paper.

***Keywords:*** *Binarization, Noisy Documents, Threshold, Modified IGT, Image Document Analysis.*

## 1. Introduction

The rising interest in historical document image analysis created many challenges for researchers. Degraded conditions of historical documents (e.g., bleed-through, ink stains, torn pages, etc) motivated researches towards binarization and enhancement algorithms suitable for these challenges. Binarization is often the first stage in all systems of image processing and analysis. Cultural heritage document collection are mostly digitized images. These precious documents are available on the Internet for manual annotations to make their content accessible. Documents may be damaged by light, particularly ultraviolet light which is present in daylight. Having original documents at home or visiting a local archive or history center gives us an opportunity to handle old material with historical evidence, a price to pay. Frequent handling of these original documents results in steady physical wear and tear with eventual loss of the document. In addition, the documents are vulnerable to damage caused by fluctuating environments and light. In this context distinction of ancient document as well as processing plays an important role in the image analysis for removing the background noise and improving the readability of the document.

There are two simple and practical Binarization techniques adopting, global and or local threshold. The Global threshold defines a global value for all the pixel intensities of the image in order to separate them as text object or background [1]. This method fails to remove non uniformly distributed noise in the image. On the other hand, local threshold provides an adaptive solution for the images with different background intensities, so

that the threshold varies according to the properties of the local region [6]. There are many general purpose Binarization methods capable of dealing with any document image with complex background. These methods fall under local or Adaptive thresholding. Bernson's proposed[2] local threshold using the neighbors, Niblack evaluated[3] the threshold at each pixel using local mean and Standard deviation, Sauvola applied two algorithms[4] in order to calculate a different threshold for each pixel. João Marcelo Monte da Silva et., all proposed [7] a fast entropy-based segmentation method for generating high-quality binarized images of documents with back-to-front interference. Xiao et., all proposed [8] an entropic thresholding algorithm based on the gray-level spatial correlation(GLSC) histogram. They revised and extended Kapur et., all algorithm. Syed Saqib Bukhari et., all proposed [9] an adaptation of local binarization method such that two different set of free parameters values are used for foreground and background regions respectively. They present the use of ridges detection for rough estimation of foreground regions in a document image. This information is then used to calculate appropriate threshold using different set of free parameter values for the foreground and background regions respectively. Chien-Hsing Chou a et., all proposed [11] a method that divides an image into several regions and decides how to binarize each region. The decision rules are derived from a learning process that takes training images as input. Rachid Hedjam .A et., all proposed [12] an adaptive method based on the maximum-likelihood(ML) classification and uses apriori information and spatial relationship on the image domain to recover weak text and strokes along with main data. Mehmet Sezgin et., all proposed[5] a extensive evaluation on existing local and global thresholding methods

A general technique for cleaning the degraded documents is proposed in this paper by using modified iterative global threshold algorithm. A simple approach in the separation of object information from fore ground is to compute a global threshold of intensity value with which two clusters can be separated. We adopted an Iterative approach[10] which can handle various degraded conditions. In each iteration the intermediate tones are shifted towards background there by providing efficient distinction between foreground and background. It is better suited for the documents having non-uniform distribution of noise.

The paper is organized as 4 sections. In the first section briefly discussed the introduction, literature survey and problem definition. In the second section discussed the algorithm for cleaning the noisy documents. In the third section discussed the experimental results and performance evaluation. Section four describes the conclusions and future scope of work.

## 2.Methodology

Presently we described a technique(Fig-1) for binarization of ancient degraded documents comes under the clustering of pixels. In this class of technique the grey level data under goes a clustering analysis, with the number of clusters being set always to two. These two clusters corresponds to the two peaks of histogram. In this technique we are trying to find mid point of the pixels. The flowchart of the proposed model evolved currently with details of binarization technique and the benchmark for its efficiency are illustrated in Fig-1

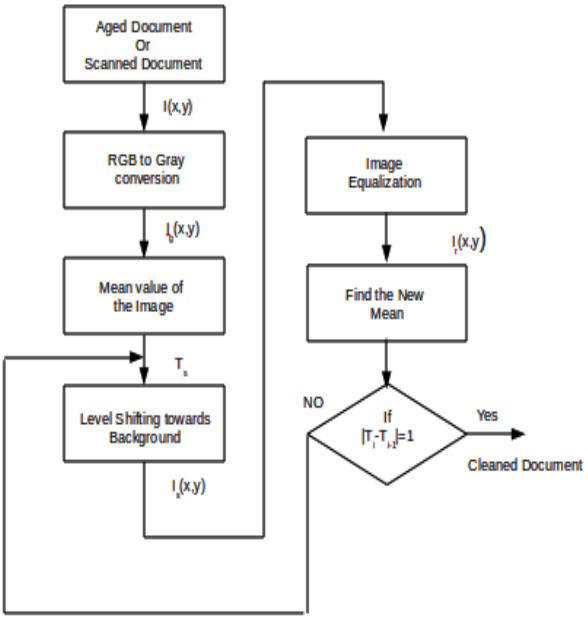



Fig.1 flowchart of the proposed model

The conventional approach in a Global thresholding technique[13] is aimed at finding a unique threshold to eliminate all pixels representing image background , while preserving others in the form of image foreground. Many real world images posses complex backgrounds or weak image foregrounds (some foreground pixels have gray values very close to some background pixels). In such cases it is difficult to find a single threshold that can completely separate the object information from the background. The same holds good in local thresholding also, where the threshold values are determined locally, e.g. pixel by pixel, or region by region. In the proposed algorithm after each thresholding operation, image equalization is carried out while evaluating the relative importance of a respective pixel intensity toward background. The new threshold is

computed to perform background elimination. This process will be repeated up to sensitive threshold.

### 2.1.Modified IGT Algorithm

The series of sequential steps are necessary for the Modified IGT algorithm consists of various steps suitable for noisy documents are viz., Extraction of degraded(noisy) document; Conversion of noisy document into Gray-scale image; Average intensity of background+object; Intensity shifting of pixels in the image towards background; Equalize the image based on the influence of foreground object intensity on background information; Determine the average intensity of the resultant image; Evaluate the threshold between the iterative average intensities.

A gray-scale image of the noisy document is generally represented by I(x,y)

$$I(x,y)= S,\ S \epsilon s\ [0,1] \quad (1)$$

Where x and y are the horizontal and vertical coordinates of the image I(x,y), and S can take any value between 0 and 1 where S=1 stands for white and S=0 stands for black. In the proposed algorithm contain intermediate tones are shifted to background. In general the fact is any document image includes few pixels of useful information (foreground) compared to the size of the image (foreground+ background). Rarely the amount of object information exceeds 10% of the total pixels in the document. Taking this advantage, it was assumed that the average value of the pixels will be determined mainly by the background even if the document is quite clear. There are two parts in the proposed Modified IGT. In the first part the level shifting of the pixels of an image is evaluated, while the second part of the algorithm, determines the relative importance of pixels with respect to object information. After each iteration some amount of pixels will be moved from fuzzy region to background. The iteration process will continue as long as the following criterion is satisfied is expressed by the Eqs-2

$$\left|T_i - T_{i-1}\right| = t \quad (2)$$

where

$T_i$ is the threshold used in $i_{th}$ iteration and

$T_{i-1}$ is the threshold before the $i_{th}$ iteration

t – is the sensitivity parameter of threshold

The threshold $T_i$ is the average intensity of background+object for an MxN document image, expressed in Eqs-3

$$T_i = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} I_i(x,y)}{M * N} \quad (3)$$

Where I(x,y) is the gray-scale image, 'M' is the number of rows and 'N' is the number of columns of an image.

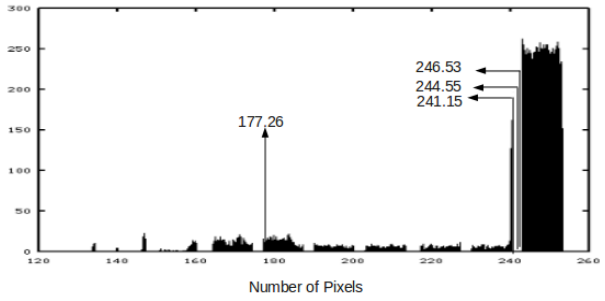
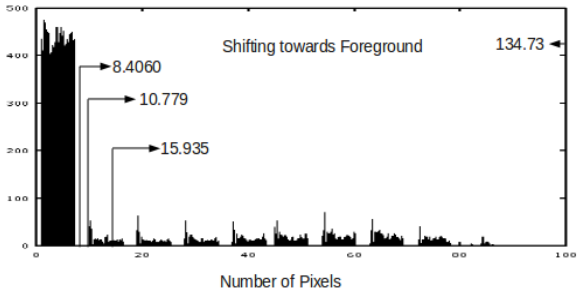



Fig.2. Mean/Threshold shift in each iteration towards Background(white)

Fig.3. Threshold shift in each iteration towards Foreground(Black)




Depending on the type of object representation in gray-level, the following assumptions are made '0' is black as text, with background as the highest luminance of document that is 255 or conversely the foreground as white and the background as black. The initial threshold is computed as first separation point of background and foreground layers. In each iteration the threshold value of the image will try to catch the background cluster. when the background is white the threshold value is shifted towards background and the converse operation is presented in Fig-2 and 3 respectively.

During the process of cleaning a clusters of pixels which are in the fuzzy region (these belongs to either foreground or background) are shifted towards background cluster when the image background is white is expressed by the Eqs-4, on the other hand the image

background is black the fuzzy pixels are shifted towards foreground as expressed in Eqs-5. Where 'L' is a variable to represent the maximum or minimum luminance value. After all iterations due to the level shifting of the pixels approximately 85% of pixels belongs to fuzzy region will be shifted to background or foreground.

$$I_s(x,y) = [L - T_i] + I_i(x,y)$$

For white Background (4)

(or)

$$I_s(x,y) = I_i(x,y) - T_i$$

For black Background (5)

After the level shifting of pixels, the leftover pixels in the fuzzy region will undergo equalization process to determine the influence of foreground object intensity on background information and is expressed in the Eqs-6 for white background images Eqs-7 is for black background images. Where "k" is the sensitivity parameter. The amount of intermediate tones shifted to background or foreground cluster depends on the value of 'k'.

$$I_r(x,y) = I_s(x,y) - k\left[\frac{L - I_s(x,y)}{L - E_i}\right]$$

For white Background (6)

(or)

$$I_r(x,y) = I_s(x,y) + k\left[\frac{L - I_s(x,y)}{L - E_i}\right]$$

For black Background (7)

Where $I_r(x,y)$ is the resultant image after image equalization $I_s(x,y)$ is given by the Eqs-5&6 and $E_i$ is the minimum (for black background images) or maximum (for white background images) pixel intensity value in the image $I_s(x,y)$ during i-th iteration just before image equalization.

By substituting the Eqs-4 into Eqs-6 and Eqs-5 into Eqs-7 the resultant equations are expressed as Eqs-8 and Eqs-9

$$I_r(x,y) = L - T_i + I_i(x,y) - k\left[\frac{T_i - I_i(x,y)}{L - E_i}\right]$$

For white Background (8)

(or)

$$I_r(x,y) = I_i(x,y) - T_i + k\left[\frac{T_i - I_i(x,y)}{L - E_i}\right]$$

For black Background (9)

The PSNR ratio is often used as quality measurement between the original (gray-scale) and cleaned image after binarization. In Eqs-10 $I_1(x,y)$ and $I_2(x,y)$ are original(gray scale) and cleaned image, M and N are dimensions of the images respectively and R is the highest variation in the input image, is equal to 255. Now the PSNR is expressed by the Eqs-12

$$MSE = \frac{\sum_{M,N}[I_1(x,y) - I_2(x,y)]^2}{M * N} \quad (10)$$

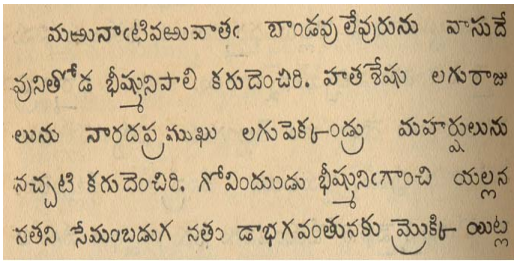
$$PSNR = 10\log 10\left[\frac{R^2}{MSE}\right] \quad (11)$$

$$PSNR = 10\log 10\left[\frac{255 * 255}{MSE}\right] \quad (12)$$

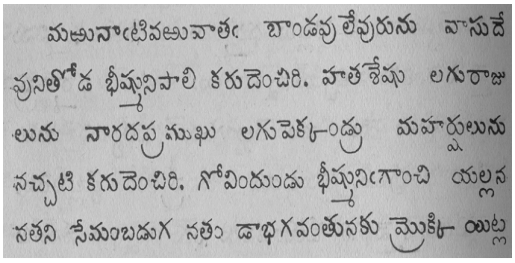
## 3. Results and Discussions

The Modified IGT algorithm is evaluated on a set of 60 document images. They are collected from the Net (Telugu old book named "Thiagarajaswami Krithis" is published in 1933, at Kesari Printing Press, Chennapuri) and the scanned copies of old story books(Telugu old book named "vydula kathalu" is published in 1942 at Madras printing press) which are of 50 to 60 years old. A typical noisy document is presented in Fig.4(a) having non-uniformly distributed noise. The background of the noisy image is golden brown in color. It is converted into gray-scale(black and white) image before applying defined IGT algorithm, is illustrated in Fig.4(b). After applying the Modified IGT algorithm on the gray-scale image continuously until the condition satisfied is given by the Eqs-2. In each iteration it removes some part of noise from the document. After completion of 4 iterations, the intensity values in the middle of the histogram are stretched back to background, so the resultant IGT images are presented in Fig.4(c,d,e,f). If the process continuous further ie.,|Ti-Ti-1|>1 the left over noise is removed but the loss of information is more means the clarity of the image is

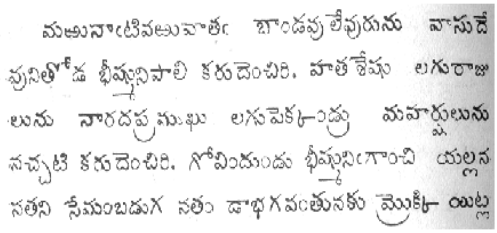
gradually decreased. Now compare the result of our method with the Otsu's method and Niblack method which are illustrated in Fig.4(g) and (h). In both these methods still some noise is presented in the document without missing the clarity of the image.



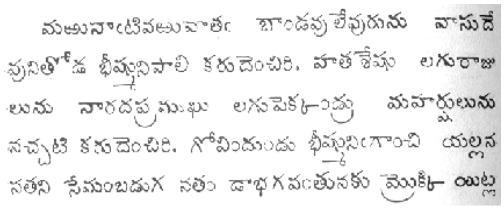
**(a)**



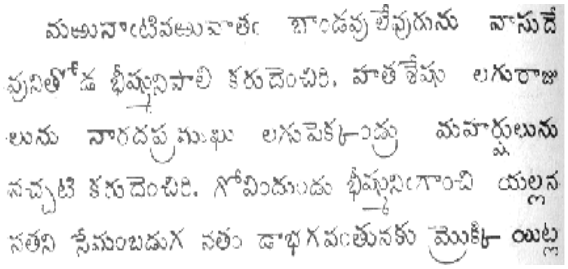
**(b)**



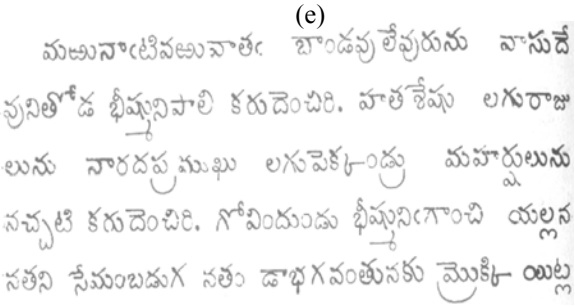
(c)



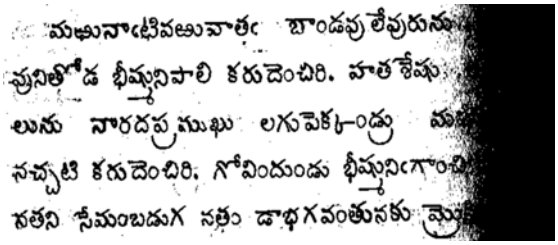
(d)



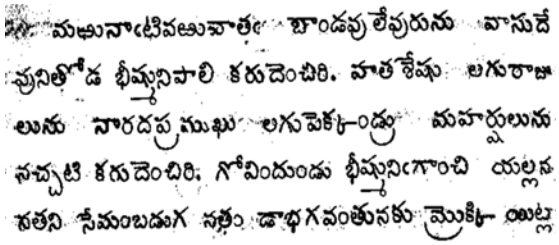
(e)



(f)



(g)

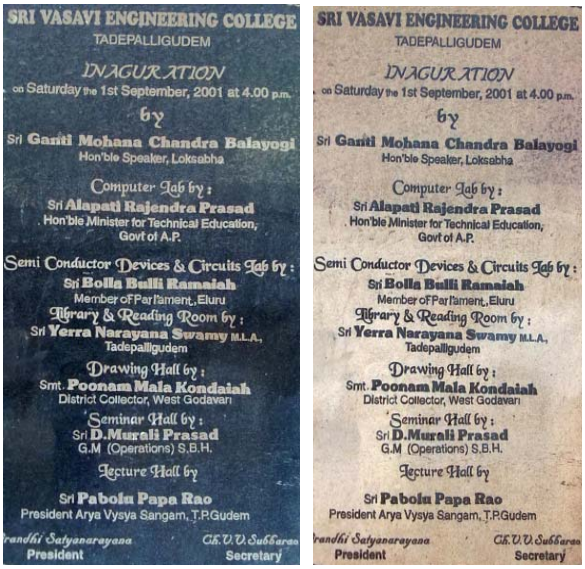


(h)

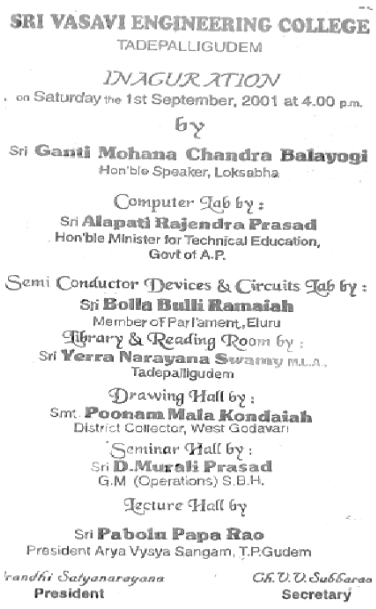
Fig-4(a) Original Image (b) Gray scale Image (c,d,e,f) Resultant images of Modified IGT(after 4 iterations) (g) Resultant image of Otsu's method (h) Resultant image of Niblack method

Modified IGT algorithm is applied on the another class of camera captured stone carved images, which are 10 years old. A typical stone(noisy) image is presented in Fig-5(a). The inverted version of the image

is presented in Fig-5(b). In the original image the background is highly dark which is dominated by the black pixels. So that the image is inverted and then applied the modified IGT for cleaning the background noise. In this image the noise is non-uniformly distributed over the image. After applying the algorithm on the image the resultant image is presented in Fig-5(c).



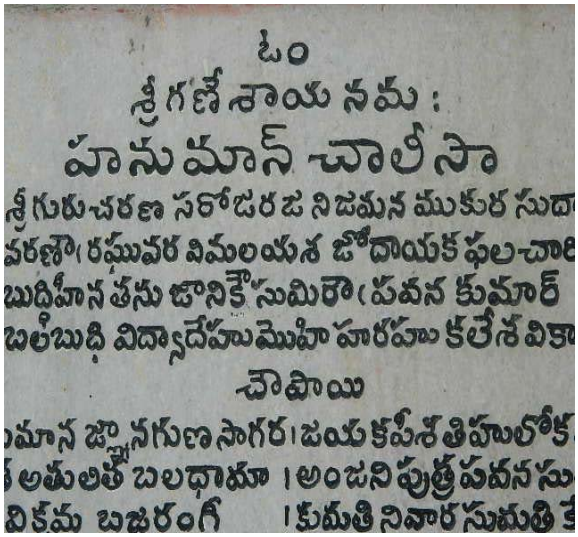

**(a)** **(b)**



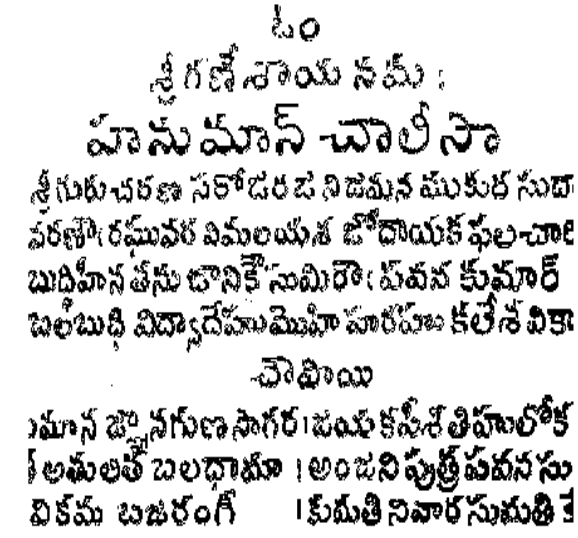

(c)

Fig-5 (a) Stone carved Image (b) Inverted image of (a), (c) Resultant image of (b)

Now the Modified IGT is applied on the a class of images with white background. Let us consider an image with white background is presented in Fig-6(a). Resultant image after applying Modified IGT is presented in Fig-6(b)



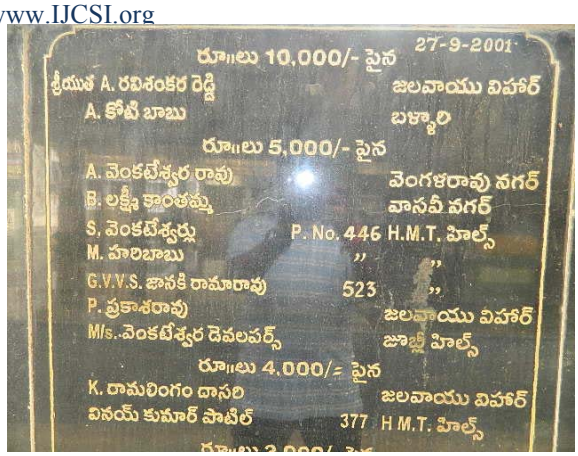

(a)



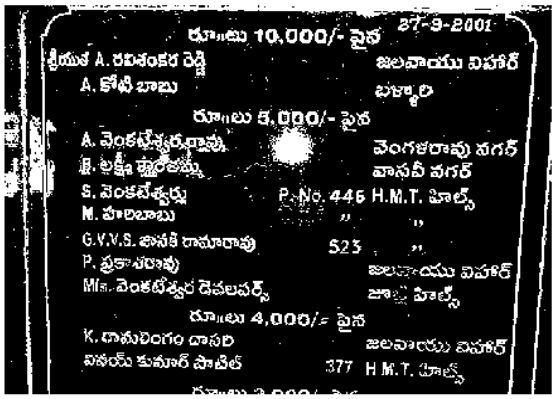

(b)

Fig-6 (a) Image with white background (b) Resultant image of (a)

Modified IGT is applied on another class of images with black background. Let us consider an image with black background is presented in Fig-7(a). After applying the defined algorithm the resultant image is presented in Fig-7(b).



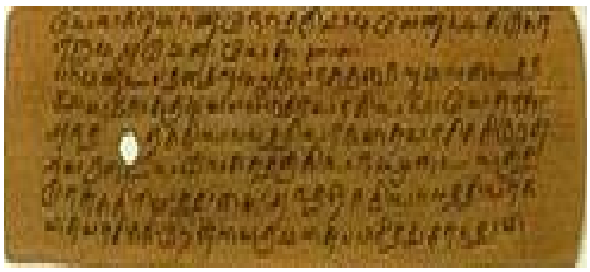

(a)



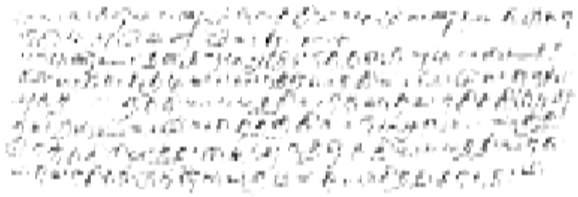

(b)

Fig-7 (a) Image with black background (b) Resultant image of (a)

The Modified IGT is applied on a set of palm leaf manuscripts. They are collected from the NET, which are 300 years old. A typical palm leaf manuscript is presented in Fig-8(a). Resultant image after applying the algorithm is presented in Fig-(b).



(a)



(b)

Fig-8 (a) Palm leaf manuscript

(b) Resultant image of (a)

## 3.1.Performance Evaluation

The performance of the algorithm is evaluated on 60 text samples. Many qualitative measures are available in the literature for measuring the quality of the image. Signal to noise ratio is one measure, used to estimate the quality of the image. There is an improvement in the S/N ratio of an image in the each iteration. This improvement is due to the fact that the tones are at intermediate stage are slowly moved towards background of the image. The S/N ratio of 10 samples in each iteration are presented in Fig.9. After analysis the S/N ratio of first 10 samples are in the range of 10db-18db in the 1st iteration, it goes on increasing up to in the range of 39db- 45 db in the 4th iteration. For higher order iterations, the S/N ratio improved gradually but the quality of image in terms of illumination gradually decreases. From the Fig.9 bottom curve represents the S/N ratio of samples undergone to the binarization of adapted algorithm. Rest of the lines represent the S/N ratio of the samples undergone to binarization of the algorithm in sucessive iterations. There is a gradual change, observed in the S/N ratio of samples.

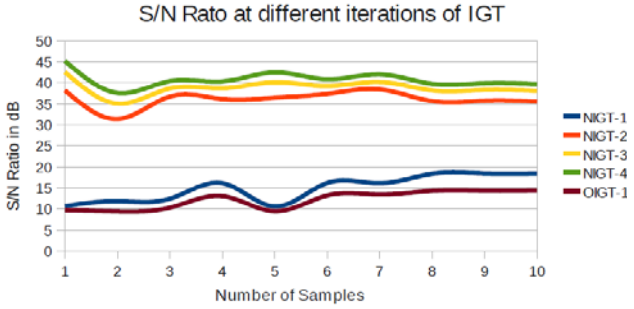



Fig.9. S/N Ratio at different iterations of IGT Algorithm in comparison with the 1st iteration of old IGT

The performance of Modified IGT algorithm

primely depends on the mean value of the image. So the effect of noise corresponding to the mean value in each iteration is illustrated in Fig.10. In the first iteration the mean value of a given 10 samples vary form sample to sample based on the presence of noise intensity and the size of the text document. In the higher order iterations, mean value is almost constant and it tries to reach the background intensity, so that $2^{nd}$ ,$3^{rd}$ and $4^{th}$ iterations are close to each other(Fig.10).

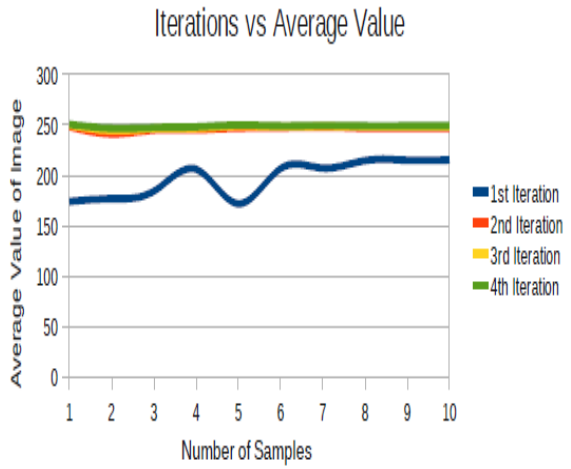



Fig.10. Number if iterations vs Average value

The amount of pixels in fuzzy region shifted towards background cluster in each iteration is presented in the Fig-11. Let us consider a sample of size(262x682), the percentage of pixels from fuzzy region to background in each iteration is presented in the table-1

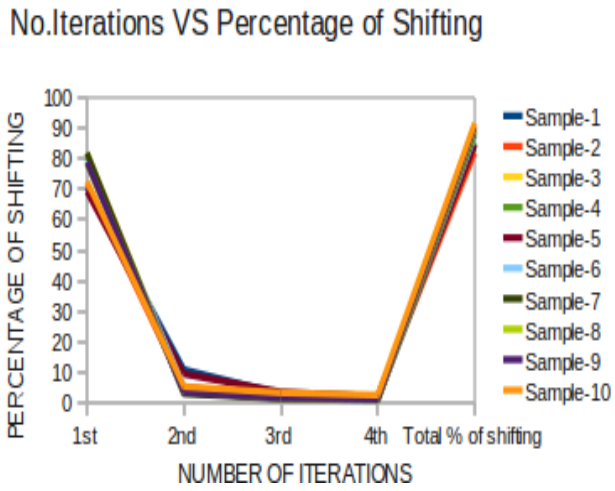



Fig.11. Percentage of pixels shifted from fuzzy region to background

Table.1 .Percentage of pixels move towards background in each iteration

| Sl. No | Iterations | Percentage of pixels shift towards background |
|---|---|---|
| 1 | $1^{st}$ | 72.59 |
| 2 | $2^{nd}$ | 5.42 |
| 3 | $3^{rd}$ | 3.38 |
| 4 | $4^{th}$ | 2.61 |
| | | Total = 83.99 |

The proposed algorithm is found to be ineffective on palm-leaf manuscripts when compared with printed and stone carvings.

## 4.Conclusions

Modified Ieterative Global thresholding is proposed in the present work. The document image under test is attempted to binarize with the help of clustering apparoach while estimating most likely background information using iterative algorithm. In each iteration the average intensity of the document image is adopted as midpoint between the clusters. In the next step the remaining pixels are equalised so as to compand the histogram. The number of iterations depends on the sesitivity of succesive thresholds. This algorithm is found to be effective on historical document images as well as camera captured stone carvings. However it is observed that further improvement is necessary on palm leave manuscripts.

## Acknowledgements

The author would like to thanks Government of India for the research grant. This work is funded by University Grant Commission, New Delhi. Under the Major Research Project UGC-MRP on Lr. No 37-1/2009(AP)

## References

[1] N.Otsu, “A threshold selection method from a gray level histograms”, IEEE Trans.Systems, Man, Cybernet., 9(1),1979, pp.62- 66
[2] J.Berson, “ Dynamic thresholding of gray-level images,” 8 th Int. Conf. on pattern recognition, 1986, pp.1251-1255
[3] W. Niblack “ An Introduction to Digital Image Processing”, Prentice Hall, 1986, pp. 115-116
[4] J.Sauvola, M.Pietikainen, “ Adaptive Document Image Binarization ,” Pattern Recognition, 33, 2000, pp.225-236
[5] Mehmet Sezgin, Bulent Sankur, “ Survey over image thresholding techniques and quantitative performance evaluation,” 146 / *Journal of electronic Imaging/* January

2004/ vol 13(1)
[6] B.Gatos, I. Pratikakis, and S.J.Perantoni, " Adaptive degraded document image binarization," Pattern recognition, vol.39, pp.317- 327, 2006
[7] João Marcelo Monte da Silva, Rafael Dueire Lins, Fernando Mário Junqueira Martins, Rosita Wachenchauzer, "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference," Journal of Universal Computer Science, vol. 14, no. 2 (2008), 299-313
[8] Xiao. Y, Cao. Z.G, and Zhang, "Entropic thresholding based on gray-level spatial correlation histogram," Proc. 19th Int. Conf. On Pattern Recognition (ICPR 2008), Tampa, FL, USA, 8-11 December 2008, pp. 1-4
[9] Syed Saqib Bukhari, Faisal Shafait, Thomas M. Breuel, " Adaptive Binarization of Unconstrained Hand-Held Camera-Captured Document Images," Journal of Universal Computer Science, vol. 15, no. 18 (2009), 3343-3363
[10] A.V.S.Rao, Tinnati Sreenivasu, N.V.Rao, T.S.K.Prabhu, A.S.C.S.Sastry, L.P.Reddy,"Binarization of Dcuments with complex Backgrounds," Procedings of International Conference on Machine Vision, 978-0-7695-3944-7/10, 2010.
[11] Chien-Hsing Chou. a, Wen-Hsiung Lin .b, Fu Chang .b.Ã, " A binarization method with learning-built rules for document images produced by cameras," Pattern Recognition 43 (2010) 1518–1530
[12] Rachid Hedjam Ã, Reza Farrahi Moghaddam, Mohamed Cheriet, " A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," Pattern Recognition(Elsevier), 2011.
[13] Maythapolnun Athimethphat , "A Review on Global Binarization Algorithms for Degraded Document Images ," AU J.T. 14(3): 188-195 (Jan. 2011)

Nekkanti Venkata Rao obtained his B.Edegree in Electronics and Communication Engineering from Bangolore University, Bangolore, India and M.Tech in Instrumentation and Control Systems from JNT University, Kakinada, India. He is working as Professor in ECE Department in Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India. He has 26 years of teaching experience in various Engineering colleges. He has 11 publications in various International Conferences and journals. His area of interests includes Image Processing, Pattern Recognition. He is active member in professional bodies like ISTE, IETE

Adabala Venkata Srinivasa Rao obtained his B.Tech degree in Electronics and Communication Engineering from JNT University, Kakinada, India, AMIE Electrical from Institute of Engineers (India), Kolkotta,India. and M.Tech in Instrumentation and Control Systems from JNT University, Kakinada, India. He is currently working as an Associate Professor in Kakinada Institute of Engineering and Technology, Korangi, Kakinada, AndhraPradesh, India. He has 7 years teaching and 4 years industrial experience. He has 13 publications in various National, International Conferences and journals. His area of interests include Pattern Recognition , Image Processing and VLSI. He is an active member in professional bodies like AMIE,IACSIT

S.Balaji received his B.E in Electronics and Communication Engineering from Andhra University, India in 1990, M.Tech. degree Osmania University, Hyderabad, India in 1995 and Ph.D from G.S. University, USA in 2003. Presently he is working as Professor and Head of the Department of Electronics and Computers Engineering at KL University, Vaddeswaram, Guntur Dist, India. He has 3 years of Industrial experience and 17 years of teaching experience. He published more than 26 publications in various National, International conferences and Journals.His area of interests include Image Processing and Pattern Recognition.

L.Pratap Reddy received his B.E degree in Electronics and Communication Engineering from Andhra University, India in 1985,M.Tech. degree in Electronic Instrumentation from Regional Engineering College,Warangal, India in 1988 and Ph.D. Degree from Jawaharlal Nehru Technological University, Hyderabad, India in 2001. From 1988 to 1990 he was lecturer in Department of Electronics and Communication Engineering at Bangalore Institute of Technology, Bangalore, India, from 1991 to 2005 he was faculty member at JNTU College of Engineering, Kakinada, India. Since 2006 he is with Department of Electronics and Communication Engineering and presently he is Professor and Head of the department at JNTUH College of Engineering, Hyderabad, India. His current activity in research and development includes, apart from telecommunication engineering subjects, Image Processing, Pattern Recognition and Linguistic processing of Telugu language. He published more than 120 research publications in various National, Inter National conferences and Journals. He is active member in professional bodies like ISTE, IE, IETE, and CSI