Historical Document Image Enhancement by Using Digital Image Processing

Roopa S Rachoti, Prof K. D. Hanabaratti

Abstract— **Historical document contain useful information. It gives chance to know about history of individual, family and conventions with old authentic confirmation. Many manuscripts are written on leaves which may contain ancient culture. They have very limited life span. The causes for degradation of document are environment condition and human negligence. Protection of these old documents is required for our future reference. Image processing technology can help to enhance these documents. It allows extracting text information from document. In proposed system Retinex enhancement technique is used. To get better improvement further Gaussian filter is applied. To do extraction of text from document background variational edge thresholding is used. As it is sensitive to edge most of border areas are preserved. The final image is binary image. It is analyzed that the proposed technique produces better results by removing background noise and improves the quality of document.**

Index Terms— **Image processing, Multiscale Retinex, Variational Image thresholding, noisy documents.**

I. INTRODUCTION

Worthy source of learning is obtained by historical records. They are utilized as reference for revelation of the cultural anthropology.

They are identified with workmanship, science, antiquated design, mathematics, Ayurveda and many more. They frequently experience the ill effects of debasements issue as a result of mischievous age, shadow, and stains on paper, low enlightenments, low quality paper, ecological awful condition and foundation with old authentic confirmation. Regular usage of these original documents may results in physical depreciation with loss of information. In addition, the document harm is caused by physically explosion to environment and light. Protection of these documents is necessary for our future reference, for current society, for study of archaeology and for country development. By digitalizing these physically stored documents, they can be made open around the world .In spite of the fact that method of improvement looks basic yet enhancing these poor adapted documents is test for advancement without bounds.

The main objective of the research is enhancement of degraded documents without affecting the other clearer contain. ‘Digital image processing technique’ is field of computer science which improves quality of the documents. Digital documentation is solution for document degradation by which worldwide and permanent access is possible.

Thresholding is prevalent method used as ideal worth choice. It aims to separate foreground object or text based on the threshold value. It transforms gray scale image into binary image. Digital image is formed from number of pixels and values foreground and background consist of different count. Gray scale images have values within range 0 to 255. Each pixel point in image is supplanted by the black color pixel, if value of image intensity is lower than consistence limit esteem i.e. some constant threshold value T and pixel supplanted by a white color pixel if the image intensity is higher than that fixed constant threshold value. In thresholding 1 refers to foreground and 0 refers background [1].

Proposed method uses Multi scale Retinex [14] as initial preprocessing to upgrade the corrupted documents and variational thresholding [22][23] to separate foreground from document background. Finally it converts given damaged test color image to gray scale picture or image and then from gray scale picture or image to binary image which is having value 0 and 1.

1.1 MULTISCALE RETINEX TECHNIQUE

In the past year 1986 Mr. Edwin land proposed Retinex philosophy as look up to for hominal color constancy[14]. Retinex belongs to section "surround and center" field where each produced value is obtained by interchangeable input value. This hypothesis is called as "retinex" [15] framed from "retina" and "cortex” of human. Retinex is an image improvement algorithm used as preprocessing pictures that is used to recover the lightness lucidity, brightness and hue between colors of an image primarily at the hand of "dynamic range compression".

In Retinex explanation, to ideal human visual strategy in the field of computer vision course of action, the image is disintegrated into the image illumination I(x,y) and the reflectance R(x,y) segments. Mathematically, the image or conception in retinex theory is presented by as yield of its reflectance with its illumination as in equation 1.

$$L(x,y)=I(x,y)*R(x,y). \qquad (1)$$

Reflectance of R(x,y) boot be obtained by computing the approximation between the image intensity and its illumination component as in equation 2.

$$E\ (x,y) = L\ (x,y)/\ I(x,y) \qquad (2)$$

The algorithm furthermore simultaneously provides color constant output and for hereafter it removes the any effects caused by illumination.

1.2 Variational Thresholding:

In locally adaptive thresholding technique, variational thresholding is one among that. It is through “minimax”

Manuscript received June 30, 2016

Roopa S Rachoti, Computer science, VTU/GIT College,Belagavi, India
Prof.Kavita .D. Hanabaratti, Computer Science,VTU/GIT College, Belagavi , India

optimization of which consists of aggregation a “non- linear” collection terms known as regularization and edge confidential data fidelity. All the earlier energy rational based adaptive image thresholding algorithms consider manually terrain weighting parameters to gain a offset between the terms regularization and data fidelity [22]. In regard to this, proposed minimax factor to automatically face this weighting parameter rate, as cleanly as the threshold surface.

II. RELATED WORK

Iterative based thresholding method is the topic described by **N Venkat Rao , A V srinivasa Rao; S Balaji and L pratap Reddy[2]** where frontal area cluster is divided from background document image is done. Clustering study is done for given set of damaged images. Based on clustering technique approaches are made to binarize image. Binarization of image is performed where pixels divided into two groups of clusters as frontal information foreground second one as background. Various different types of degraded documents were studied and they conducted experiment on documents like simple papers, high quality camera snapped images, different stone carvings images and old palm leaf manuscripts and performance is improved. The results are compared by knowing performance and hence it is evaluated based on mathematical signal to noise ration as key judgment factor.

M preetham Krishna, A sriram, N B Puha[4] have explained clustering based Binarization. The main objective is correct data extraction from foreground of old palm leaf image was the only intention. To do experiments ground truth image of ten old palm leaf images is prepared. Every pixel in a test image is marked and read as to be 1 or 0 by human observation. They have used latest k- mean clustering algorithm as base. After making comparative studies with many other thresholding methods this method views to be more correct and applicable among all other.

In this **paper J. Ramya and B. parvathavarthini[25]** have described the new efficient binarization method and it is based on the “Divide and conquer”. It recognizes threshold point calculation on the bases of gray level pixel intensities. It makes use of low pass Weiner filter method as a first preprocessing step to enhance the image from deblurring the image to look bright. It makes uses of Median Filter method as a last post processing step to reduce any other possible remaining noise in the image. For the better results uniform and equally distributed image is required so it has become its one of the weakness it the given document is not even illumination. And it does not clear all noise.

B.Gatos, I. Pratikakis and S.J. Perantonis[5] explained adaptive approach for document improvement like any other efficient Binarization technique in this process also they have used low pass wiener filter and denoising of the poor old documents done. The filter improvement to document is based on local nearby statistics of image. In the second step, it conducts premier rough esteem of foreground area regions of image. In third step, it measures the background base of the given image. In the fourth step, last binarization is done by integrating information from the computed background base and the original image or picture. The new data and text is protected using excess thresholding. After that post processing is done in which any stroke joining connectivity is kept, noises are tried to remove completely and entire image quality of diminished image is improved.

Cleaning and enhancing historical documents is proposed by **Ergina Kavallieratoul and Hera Antonopoulou**[19]. In this method it is considered that the actual foreground image pixel is not more than 1 0 % of its overall image. There is no any necessary to do further more processing. And hence making improvement to document is very simple and easy method. In the beginning where the initial stage computation of average pixel value of image is done and then further subtraction of these average pixel value is generated in the next step. Then further histogram equalization is done in final step where binary image comes as an output image. This method in very low in cost because of simple calculations and it has effective robustness behavior.

Zia-ur Rahrnan, Daniel J. Jobson,Glenn A. Woodell[14] implemented Multi-Scale Retinex is instinct image improvement algorithm which has been discussed by them. "Dynamic range compression" and also "color constancy” are two techniques which are used at the very same time for image enhancement. It is inspired by hominal visual framework and compared with computer vision course of action. And duplication of this concept to computer vision is tried to obtain. Some calculations are made on illumination and light where reflectance is the output. This concept gives maximum results and became one of the theories in which maximum noise is removed.

New variational thresholding technique proposed by **Nilanjan Ray and Baidya Nath Saha[22].** Computation of weights on the data fidelity is done automatically and energy functional is regularized. Automatic tuning of weighing parameters is objective of this technique. This process is unique and good among all processes and hence texture effect as a result of this process is better than any other technique.

Text extraction from historical documents is obtained by **Toufik Sari, Abderrahmane Kefali [13] and Halima Bahil**. Hybrid thresholding method by integrating many other Binarization techniques is discussed. They proposed a hybrid thresholding method for to do any enhancement to degraded document. One of the two theories is used as foundation otsu's algorithm is used for all the global thresholding of image. Then threshold values are marked as fixed value. The pixel count which are below some first threshold value they are considered as foreground image pixel and preserved. Then the pixel counts which are bigger than the second threshold value are received as background pixels and removed. Then local calculation based on neighborhood information of the remaining pixels is done.

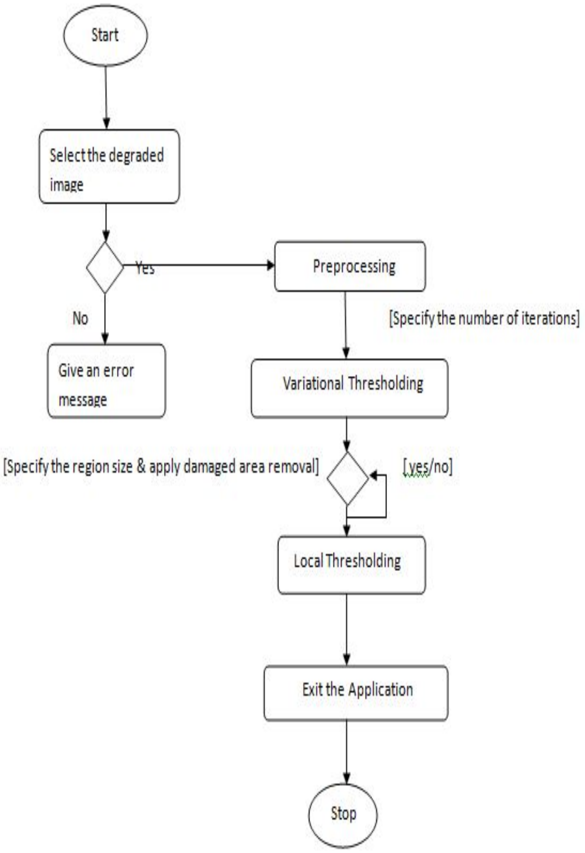
Different thresholding methods were discussed by **Chen Yan & Graham Leedham[22]** and they made comparative studies regarding some of the existing local global algorithms for handwriting extraction. The study of local adaptive analysis method is the foundation where local feature vectors are used. This is compared against some existing global and local algorithms. To get thresholding local area this vector is used and found to be the best. And different algorithms are picked instantly for the documents which are under investigation. Division of main image under test into 10 different sub images recursively done. Further evaluation of algorithm is done after testing 10 ancient images. One of the measuring component i.e. Recall value is chosen for performance of evaluation of program.

Drawback of the existing system

1. Existing system gets failed when degraded documents have varying background.
2. Existing system do not work well with palm leaves.

III. SYSTEM DESIGN

The proposed method is based on simple image processing technique. The document images are taken from the database. Initial enhancement to the documents is made by Retinex technique. In the second phase variational thresholding is performed, which segment the text or object from background and the text information is preserved. The color image is converted into gray scale image. During third phase, the areas which still contain background noise are detected and are removed. Finally gray scale image is converted into binary image



1. preprocessing:
It is the first step of image processing. Different kinds of degraded old documents are taken as input to preprocess. Image preprocessing eliminates some amount of noise and makes image clear. Smoothens the background texture and provides contrast enhancement between background and text. After selecting image, next stage is to preprocess the selected image. The original image is converted into gray scale image. In preprocessing Retinex algorithm followed by Gaussian filter is used to enhance the image.
1. Initially enhancement is done with Multiscale Retinex technique.

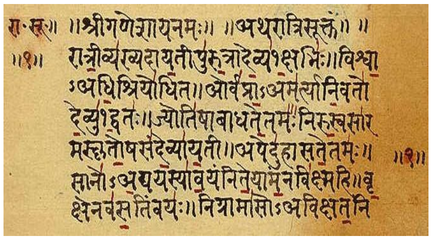


Fig.2: Original ancient document

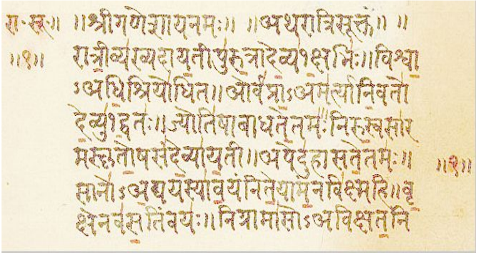


Fig. 3: Output image after MSR

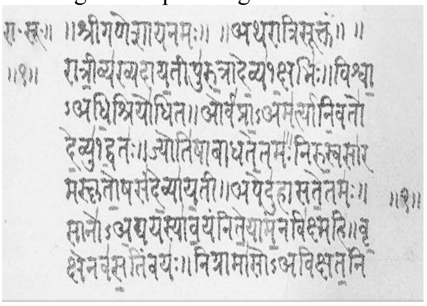


Fig.4: Image after applying Gaussian filter

2. variational thresholding:
This is second module need to be performed. In this extraction of text from document is performed. The improvements to thresholding i.e. border preservations are done by choosing number of iterations. It can be obtained by increasing or decreasing the iteration. In proposed method maximum 1000 iterations are limited.
The estimation of q is selected by doing tests on images and is called as edge sensitive function. The value q is picked as q=50 it appears, better thresholding score on document under test. If the increment is done to value q, with no any bound then, $g(x, y)$ it will tend to zero for most pixel areas (x, y) and thusly the impact of edge sensitive data term on energy functional decreases.

At that point regularization term rules over data term. Hence forth there is need to restrict estimation as 50. To improve intent in the obtained threshold binary image of Fig 5. There is need to acquire biggest and second biggest dark associated segments from center segment of the image. Next morphological operations are utilized to take out any remaining white associated segment inside the image. Further any conceivable superfluous parts associated are additionally dispensed with by morphological operations.

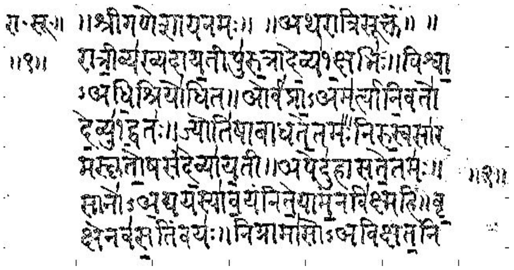


Fig.5: Output image after variational thresholding

3. local thresholding:

Further toning to the document is provided by local thresholding. The choice of region selection is given to user. If there is a still noise remains after variational thresholding then user can go for damaged area removal.

it tries to eliminate any spurs in the image using morphological operation and then it further eliminate unwanted region from the image using region props and the final output is shown in fig.6

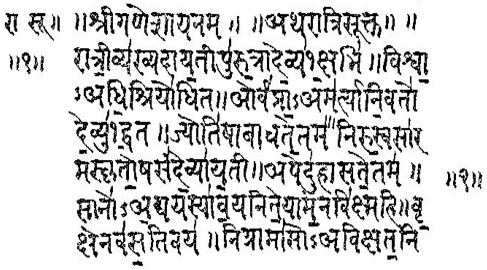


Fig.6: Final binary image after Local thresholding

IV. Results comparisons

Variational thresholding is edge sensitive. It preserves more text compared to global thresholding. In given figure 7 the output of global thresholding loses foreground information, where in case of variational thresholding foreground text information is preserved fig 8

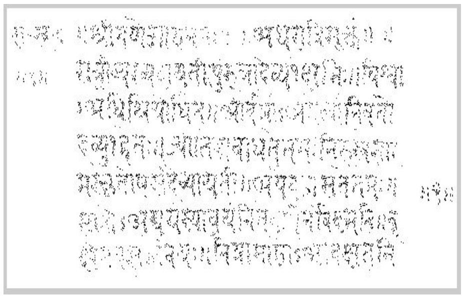


Fig. 7: Result of global thresholding

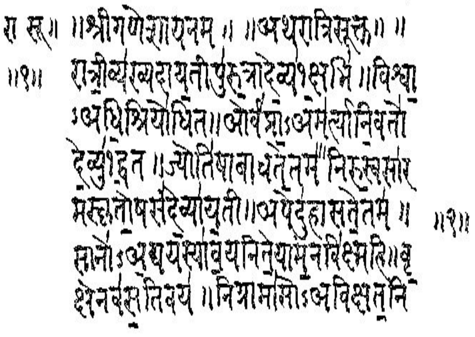


Fig.8: Result of variational thresholding

CONCLUSION

The proposed Method is implemented using MATLAB. In this study Thresholding method based on Retinex theory followed by a histogram processing is achieved aiming at removal of background noise from the historical documents. The proposed method overcomes the drawbacks of the related global threshold techniques. It can deal with degradations which occur due to non-uniform illumination, low contrast, uneven background and stains.

REFERENCES

[1] Kale ,Preeti ,Phade, G.M Gandhe,S.T,Dhulekarpravin "A Enhancement of old images and documents by digital image processing techniques" , International Conference on Communication, Information &Computing Technology (ICCICT),IEEE 2015.

[2] N Venkat Rao , A. V. Srinivasa Rao, S Balaji ,L Pratap Reddy, "Cleaning of Ancient Document Images Using Modified Iterative Global Threshold" , International Journal of Computer Science ISSN: 1694-0814,Issues, Vol. 8, Issue 6, No 2, November 2011.

[3] Rafael C. Gonzalez, University of Tennessee Richard E. Woods, "Digital image processing ",MedData Interactive 3rd edition.

[4] M Preetham Krishna, A Sriram, N B Puhan "Clustering based Image Binarization in Palm Leaf Manuscripts", IEEE ,2014.

[5] B. Gatos, I. Pratikakis, S.J. Perantonis, "Adaptive degraded document image Binarization", The journal of pattern recognition society pp.317-327 Elsevier 2005

[6] Rupinder Kaur, Mr.Naveen Goyal "Document Image Binarization Technique For Degraded Image By Using Morphological Operators."

[7] Prashali Chaudhary, B.S. Saini "An Effective And Robust Technique For The Binarization Of Degraded Document Images",International Journal of Research in Engineering and Technology (IJRET), eISSN: 2319-1163 | pISSN: 2321-7308 , 2014.

[8] Bolan Su, Shijian Lu *Member, IEEE*, Chew Lim Tan *Senior Member, IEEE,"* A Robust Document Image Binarization Technique for Degraded Document Images",IEEE2012.

[9] KaveriJagtap, Chandraprabha. A. Manjare, **"**An Ancient Degraded Images Revamping Using Binarization Technique"International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307,vol-4,Issue -6, January 2015**.**

[10] J. Sauvola, M. PietikaKinen, "Adaptive document image Binarization", Pattern Recognition 33 (2000) 225},1999.

[11] J. He, Q. D. M. Do, A. C. Downton and J. H. Kim, **"**A Comparison of Binarization Methods for Historical Archive Documents"**,** International Conference on Document Analysis and Recognition (ICDAR'05),IEEE 2005.

[12] VavilisSokratis, ErginaKavallieratou, Roberto Paredes, and Kostas Sotiropoulos "A Hybrid Binarization Technique for DocumentImages"SCI 375, pp. 165–179.Springer-Verlag Berlin Heidelberg 2011

[13] ToufikSari,AbderrahmaneKefali, and Halima Bahi. "Text Extraction from Historical Document Images by the Combination of Several Thresholding Techniques", Volume 2014, Research Article ID 934656, 10 pages.

[14] Zia-ur Rahman, Daniel J. Jobson,Glenn A. Woodell,"Retinex processing for automatic image enhancement" NASA Research centre **,** Journal of Electronic Imaging 13(1), 100–110,january 2004.

[15] Daniel J. Jobson, *Member, IEEE,* Zia-ur Rahman, *Member, IEEE,* and Glenn A. Woodell, "A MultiscaleRetinex for Bridging the Gap Between Color Images and the Human Observation of Scenes", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 6, NO. 7, JULY 1997

[16] Marian Wagdy, IbrahimaFaye andDayangRohaya "Document Image Binarization Using Retinex and Global Thresholding", ELCVIA ISSN: 1577-5097 Published by Computer Vision Center / UniversitatAutonoma de Barcelona, Barcelona, Spain, Jun 2015

[17] XijiangKe, Rai Jin*, Xia Xie, Jie Cao, "A Distributed SVM Method based on the IterativeMapReduce", IEEE ICSC 2015, Anaheim, California, USA, February 7-9, 2015,

[18] Sonia saini, Ms. RichaDogra," Robust Document Image Binarization Technique For Degraded Document Images**",** Global Journal of Computers & Technology, Vol. 3, No. 2, September 23, 2015.

[19] Ergina Kavallieratou and Hera Antonopoulou, **"**Cleaning and Enhancing Historical Document Images", springer 2005

[20] Ms. R.Girija, "Exploring Image Enhancement And Optimization Techniques For Fabric Defect Identification In Hosiery Industry", 2014-2015.

[21] Ling Tang, ShunlingChen,Weijun Liu, Yonghong Li, "Improved Retinex Image Enhancement Algorithm", Published by Elsevier,2011.

[22] ChenYan & Graham Leedham "Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document", Images Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition IEEE 2004.

[23] Nilanjan Ray BaidyaNathSaha Alberta, Canada Edmonton, "Edge Sensitive Variational Image Thresholding".

[24] Akshay Gujar, DarshanChatur, Ellora Bhattacharya, Jaspreet Singh," Intensification of Old Documents and Photos by Digital Image Processing Techniques",International Journal of Advanced Research in Science,Engineering and TechnologyVol. 2, Issue 10 , October 2015

[25] J.Ramya,B.Parvathavarthini "An Efficientbinarization Technique For Historical Document Images" , Journal of Theoretical and Applied Information Technology,july 2014

[26] lekarpravin "A Enhancement of old images and documents by digital image processing techniques" , International Conference on Communication, Information &Computing Technology (IE 2015.