Document Recovery from Degraded Images

[1] Jyothis T S, [2]Sreelakshmi G, [3]Poornima John, [4]Simpson Joseph Stanley,
[5]Snithin P R, [6]Tara Elizabeth Paul

[1]AP, CSE Department, Jyothi Engineering College, Kerala, India

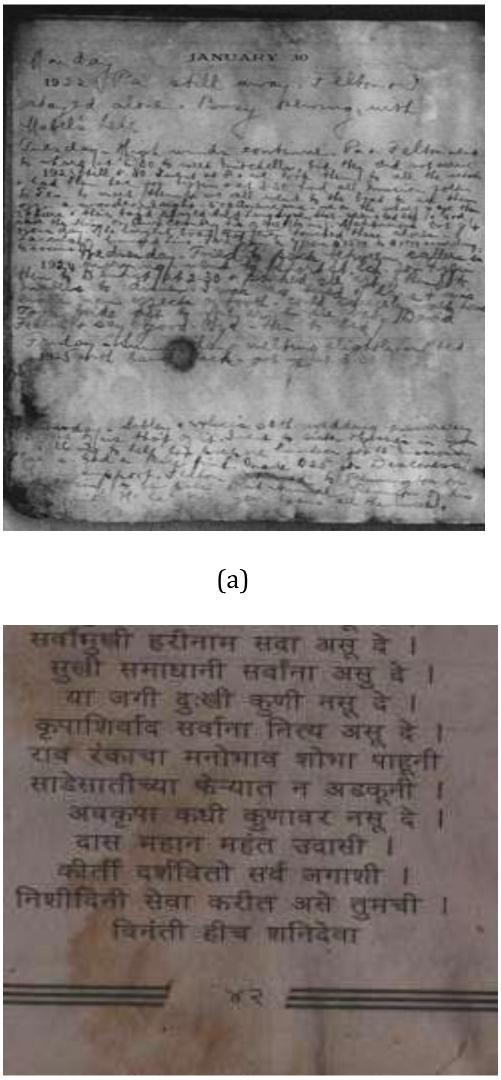
[2 3 4 5 6] Students, CSE Department, Jyothi Engineering College, Kerala, India

Abstract - *Recovery of document from its damaged fragments plays an important role in the field of forensics and archival study. Also, now-a-days, there are many activities which depend upon the internet.. Many a times it happens that institutes and organizations have to maintain the books for a longer time span. Books being a physical object, so it will definitely have the issues of wear and tear. The pages definitely get degraded and so does the text on the pages. Due to this degradation many of the document images are not in readable. So, there is a need to separate out text from those degraded images and preserve them for future reference. This paper introduces a method for accomplishing the task of recovering the contents from the degraded papers. The image is converted to contrast image, whose difference in luminance makes an object clear. The edges are detected which is then binarized. The segmentation of document text is carried out by a local Threshold which is estimated based on the intensities of detected edge strokes. Experiments are carried out on several challenging bad quality document images which show the best performance of the proposed system within a shorter period of time.*

Key Words*: *Image contrast, Binarization, Edge Detection, Pixel classification.

1. INTRODUCTION

Recovery of degraded documents has always been a challenge to people. There are many situations where paper documents become a crucial part. Recovering the paper documents plays an important role in forensics and archival studies. Such situation needs an efficient solution to get the exact contents of the paper documents. Now-a-days everything being digitized it is really hard to convert old paper works to computerized one's. It happens many a times that many organizations and instituted store their record works in paper books and with time it would have been severely spoiled. There also Exists situations where people try it hard to read the contents being written on the old paper works. In such cases there is an essentiality for a system that can help read all these degraded documents.



(a)

(b)

Fig.1 Degraded document image example.

An optimal solution for eliminating these problems is to use binarization technique which converts grayscale document images to binary document image. The image is initially converted to contrast image which helps distinguish the contents. Prior to local threshold estimation the contrast image is converted to grayscale so as to clearly identify the text stroke from background and foreground pixels. After segmentation using local threshold method which is estimated based on the intensities of the detected text stroke edge pixel it is converted to binary form. The quality of the image is improved using the post processing method.

1.1 Literature Survey

There are many techniques which have been developed for document image binarization. The problem with the existing technique is its complexity and the cost to recover data and also it is slow for large images. It does not accurately detect the background depth due to non uniform illumination, shadow, smear or smudge. Global thresholding [10] cannot be considered as a suitable approach for degraded document binarization as many documents do not have a clear bi-modal pattern. Local threshold estimation [14] is a better way to deal with variations in the documents. There are other methods too like recursive method [15], decomposition method [16], texture analysis, matched wavelet, background subtraction [4] for thresholding. The methods combine much image information and are also very complex.

In [2] Sauvola has proposed a method where the contrast values of text background and text are focused. The threshold is found using two methods Soft Decision method (SDM) and Text Binarization method (TBM). SDM is used to remove noise and separate text components from background. TBM is used in cases of uneven illumination.

The paper [4] explains the fusion of two well known binarization methods: Gatos et al. and Niblack, using dilation and logical AND operations. Artificial Neural Network combined with fuzzy algorithm [5] can be used to map different degrading factors. A Back propagation neural network is used to train N samples and the output is compared with the desired output of the sample.

To segment text from document background, local image contrast and local image gradient features can be used because the document text has certain image contrast to its neighboring image background. In paper [3] the local contrast is defined as:

$$C(i,j) = I_{max}(i,j) – I_{min}(i,j) \qquad (1)$$

where C(i,j) is the contrast of image pixel (i,j) , $I_{max}(i,j)$ and $I_{min}(i,j)$ are maximum and minimum intensities within a local neighborhood window. This method is simple but cannot be applied to complex documents.

We here use a local image contrast method which is based on paper [1] and it is evaluated as follows:

$$C(i,j) = \frac{I_{max}(i,j) - I_{min}(i,j)}{I_{max}(I,j) + I_{min}(I,j) + €} \qquad (2)$$

Where € is positive but very small. The equation 2 introduces a normalization factor in order to compensate the image variation.

2. PROPOSED SYSTEM

There are five modules in our proposed system. They are: Contrast image construction, Text stroke edge pixel detection, Local threshold estimation, Binary conversion, Post processing. Given a degraded document, initially the contrast image is constructed which then determines the edge strokes of the text document. Text is segmented based on the local threshold which is estimated from the detected text stroke pixels. It is further converted to binary form. Finally post processing is done in order to improve the efficiency of the resultant image. The system architecture can be shown as:

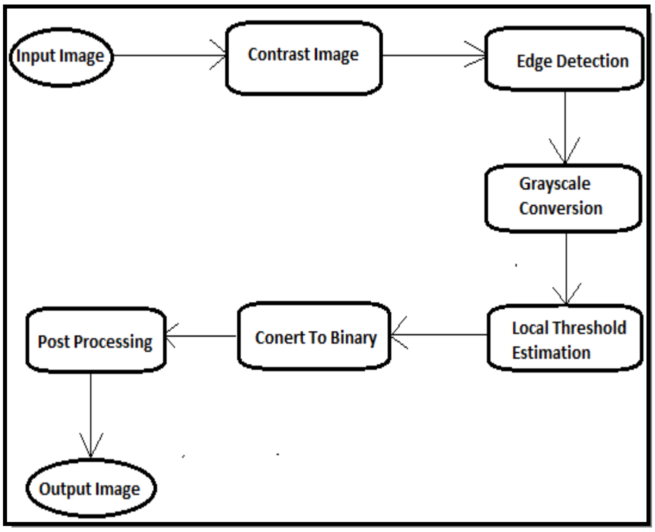


Fig.2. System Architecture

2.1 Contrast Image

Usually contrast is the difference in the luminance or color of the image which makes the object clear. It can also be thought as the variant in the color and intensities of the objects .Image gradient is used to detect the text stroke edges of the degraded document In order to detect only the stroke edges it is necessary that the gradient is normalized.

The equation 2 shows (as in [1]) the local contrast calculation where the numerator captures the image gradient and the denominator is a normalization factor which suppresses the image variation. For image pixels within bright regions the image contrast will be less and for those with darker regions the image contrast will be high. A combination of local image contrast and local image gradient will be helpful in handling bright text properly. So the Adaptive local image contrast is as follows:

$$C_\alpha(i,j) = \alpha C(i,j) + (1-\alpha)(I_{max}(i,j) - I_{min}(i,j)) \qquad (3)$$

Here C(i,j) is the local contrast as in equation 2 and $(I_{max}(i,j) - I_{min}(i,j))$ is the local image gradient whose value is normalized to [0,1].A local window is required for local image contrast an the window size is set to 3, α is the weight between local contrast and image gradient. The value of α will assigned large for image contrast when there occurs a high variation in image intensity. Else image gradient will be assigned with large α value. The weight α can be calculated as:

$$\alpha = (Std/128)^{\gamma} \qquad (4)$$

Where Std is the Standard deviation of the document image and γ is the pre-defined parameter.

2.2 Edge Detection

The contrast image construction is a n important phase whose purpose is to detect the stroke edges pixels of the document. This is used to produce a border around the foreground text pixels thereby differentiating the foreground and background pixels. The contrast image which is constructed has a clear bi-modal pattern. Here we calculate the text stroke edge pixels candidate by using Otsu's thresholding method. Since the contrast image has a bi-modal pattern it can be combined with edges from Canny's edge detector as it has a good localization property i.e. it can mark the edges close to its real edge location. Before performing Otsu's thresholding the contrast image is converted to grayscale image. It is done in order to sharpen the edges of text stroke thereby increasing the efficiency.

The most generally used grayscale method is the averaging method. But in our system we use Luminance grayscale [6] method as it is much more suitable for enhancing the text strokes. The luminance grayscale method is as follows:

$$Gray = (Red*0.21 + Green*0.71 + Blue*0.072) \qquad (5)$$

2.3 Local Threshold Estimation

There are mainly two characteristics that can be observed from document images; one is that the text pixels will be very much close to the detected text pixel. The other one is that there is a distinct difference in the intensities of high contrast text stroke edge pixels and the surrounding background pixels. The detected text stroke edge pixels can thus be used to extract the document text image. It is as follows:

$$R(x,y) = \begin{cases} 1, & I(x,y) \le E_{mean} + E_{Std}/2 \\ 0, & \text{Otherwise} \end{cases} \qquad (5)$$

Where E_{mean} and E_{Std} are the mean and standard deviation of intensities of detected text stroke edge pixels. The edge width is calculated by using the edge width estimation algorithm.

2.4 Convert to Binary

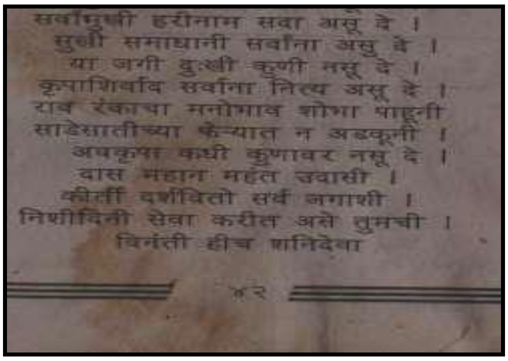
The image obtained after threshold estimation is converted to binary format i.e. 0 and 1. The image pixels at background are assigned value 0 and those of foreground are assigned the value 1 which has highest intensity.

2.5 Post Processing

There are chances that still there occurs some background pixels in the recovered image due to variation in background intensities and irregular luminance. These unwanted pixels are to be removed and this is done by post processing. It returns a clear image which consists of the actual image. In the post processing procedure, first, the pixels which do not connect with the foreground pixels are removed out to set the edge pixel precisely. Next, if the neighborhood pixels lie in the same class then one among the pair is labeled to another category.

3. RESULTS AND ANALYSIS

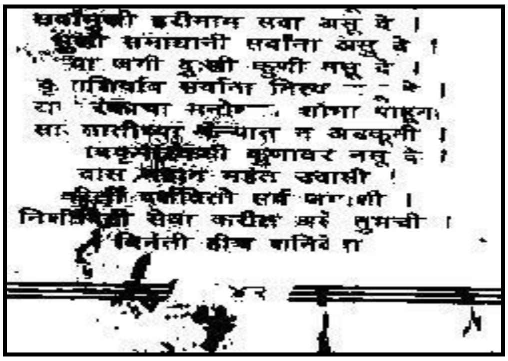
The input to our proposed system is a degraded image. Suppose it is the image as shown below:



सर्वांमुखी हरीनाम सदा असू दे ।
सुखी समाधानी सर्वांना असु दे ।
या जगी दुःखी कुणी नसू दे ।
कृपाशिर्वाद सर्वांना नित्य असू दे ।
राव रंकाचा मनोभाव शोभा पाहूनी
साडेसातीच्या फेऱ्यात न अडकुनी ।
अवकृपा कधी कुणावर नसू दे ।
दास महान महंत उदासी ।
कीर्ती पसरविली सर्व जगाशी ।
निशीदिनी सेवा करीत असे तुमची ।
विनंती हीच शनिदेवा
४२

Fig.3.Original Input

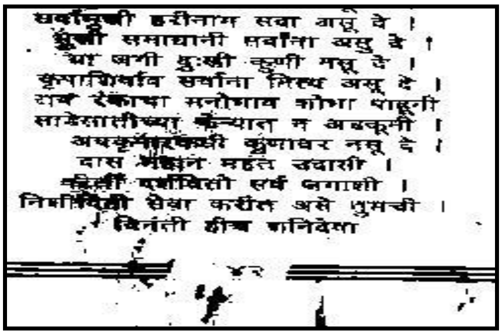
The first operation performed is the contrast construction. Here both local contrast and local image gradient are applied on the image. Then the edge detection is done It is shown in fig 4.



सर्वांमुखी हरीनाम सदा असू दे ।
सुखी समाधानी सर्वांना असु दे ।
या जगी दुःखी कुणी नसू दे ।
कृपाशिर्वाद सर्वांना नित्य असू दे ।
राव रंकाचा मनोभाव शोभा पाहूनी
साडेसातीच्या फेऱ्यात न अडकुनी ।
अवकृपा कधी कुणावर नसू दे ।
दास महान महंत उदासी ।
कीर्ती पसरविली सर्व जगाशी ।
निशीदिनी सेवा करीत असे तुमची ।
विनंती हीच शनिदेवा
४२

Fig.4. Edge Detected image

The final resultant after the entire process can retrieve all the text contents without any significant content loss. The resultant output image is as in fig 5.



सर्वांमुखी हरीनाम सदा असू दे ।
सुखी समाधानी सर्वांना असु दे ।
या जगी दुःखी कुणी नसू दे ।
कृपाशिर्वाद सर्वांना नित्य असू दे ।
राव रंकाचा मनोभाव शोभा पाहूनी
साडेसातीच्या फेऱ्यात न अडकुनी ।
अवकृपा कधी कुणावर नसू दे ।
दास महान महंत उदासी ।
कीर्ती पसरविली सर्व जगाशी ।
निशीदिनी सेवा करीत असे तुमची ।
विनंती हीच शनिदेवा
४२

Fig.5.Output Image

4. CONCLUSION

Our project is based on recovering the degraded document contents. The usage of binarization technique has made the system more efficient in task. Our system is an adaptive method to recover the contents from any document set. This is a very simple and fast technique and also efficient with any sort of document. The main highlight is that it can be used for any language .Our project is irrespective of language and can recover any language contents. The application is useful in many fields like forensics, historical department etc. With the digitization of the world everything has turned out to computer so our system also focuses on digitizing the old paper documents which are highly confidential and important.

ACKNOWLEDGMENT

First and foremost, we express our thanks to The Lord Almighty for guiding us in this endeavour and making it a success. We take this opportunity to express our heartfelt gratitude to all respected personalities who is guided, inspired and helped us in the completion of this Main Project.

REFERENCES

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, “Robust image binarization technique for degraded document images”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.

[2] J. Sauvola and M. Pietikainen, “Adaptive Document Image Binarization”

[3] S. Lu, B. Su, and C. L. Tan, “Document image binarization using background estimation and stroke edges,” *Int. J. Document Anal. Recognit.*, vol. 13, no. 4, pp. 303–314, Dec. 2010.

[4] Brij Mohan Singh Mridula “Efficient binarization technique for severely degraded document images”, CSIT (November 2014) 2(3):153–161

[5] Harshmani, Nancy Gupta*, Gurpreet Kaur, “Neuro-Fuzzy Approach: A Robust Way to RestoreDegraded Documents”, International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181 Vol. 5 Issue 05, May-2016

[6] Yogita Kakad1, Dr. Savita R. Bhosale, “An Advanced document binarization for Degraded document recovery”International journal of Advanced technology in Engineering and science, Volume No 03, Special Issue No. 01, April 2015

[7] B. Gatos, K. Ntirogiannis, and I. Pratikakis, “ICDAR 2009 document image binarization contest (DIBCO 2009),” in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382

[8] Jyotirmoy Banerjee, Anoop M. Namboodiri, and C.V. Jawahar “Contextual Restoration of Severely Degraded Document Images”,Proceedings of 3rd IRF International Conference, 10th May-2014, Goa, India.

[9] G. Bala, G. Agama, O. Friedera, G. Frieder “Interactive degraded document enhancement and ground truth generation”, 2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC).

[10] A. Brink, “Thresholding of digital images using two-dimensional entropies,” *Pattern Recognition.*, vol. 25, no. 8, pp. 803–808, 1992.

[11] Manoj S Ishi, Lokesh Singh, Manish Agrawal “Reconstruction Of Images With Exemplar Based Image Inpainting And Patch Propagation”, Icices2014 - S.A.Engineering College, Chennai, Tamil Nadu, India, 2014

[12] Brij Mohan Singh Mridula “Efficient binarization technique for severely degraded document images”, CSIT (November 2014) 2(3):153–161

[13] S.Tamilselvan, M.E., S.G.Sowmya, M.E.,“Content Retrieval From Degraded Document Images Using BinarizationTechnique.”,international conference on computation of power, energy, information and communication(ICCPEIC),2014.

[14] J. Bernsen, “Dynamic thresholding of gray-level images,” in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.

[15] Y. Liu and S. Srihari, “Document image binarization based on texture features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5,pp. 540–544, May 1997

[16] Y. Chen and G. Leedham, “Decompose algorithm for thresholding degraded historical document images,” *IEE Proc. Vis., Image Signal Process.*, vol. 152, no. 6, pp. 702–714, Dec. 2005.