A Foreground Text Extraction Mechanism For Degraded Document Images Using Adaptive Binarization

Mr. Yogeshwar L. Ade
Information Technology Department
Siddhant College of Engineering,Sudumbare
Pune, India
E-mail: yogeshwar.ade@gmail.com

Prof. Sonali Rangdale
Information Technology Department
Siddhant College of Engineering,Sudumbare
Pune, India
E-mail: sonali_rangdale@rediffmail.com

***Abstract*— Now-a-days, there are many activities which depend upon the internet. And there is a great need to shift all the activities which are performed by user towards the digitization of world. Many a times it happens that institutes and organizations have to maintain the books or novels for a longer time span and there arises a new challenge for the institutes. 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. The data on the pages can be confidential and sensitive and there should be very robust and dynamic mechanism for preserving the data on the same. 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 gives a great reason for developing a foreground text extraction mechanism that will aid in preserving the documents or in other words, the text on those documents. The proposed system includes such a mechanism that not only helps to detect the textual matter on the documents but also preserve the text on the other image. Previously, many such algorithms have been proposed for this purpose, but as seen by the research done for years, Optical character recognition, Handwritten text recognition such algorithms were developed but there are still few areas which were yet to be worked on. The proposed system focuses on improving the text extraction efficiency and therefore eradicates the use of Canny's edge map and makes use of simple Otsu thresholding and edge detection and luminance Grayscale method for improving the detected edge sharpness. Also the very important aspect on text extraction is clarity of text being extracted. In this paper, Post processing algorithm works on the same task for smoothening the extracted text and also removing the unwanted pixels from the image. These algorithms include image contrast inversion, edge estimation, image binarization and post processing of binary image. We can able to separate out the foreground text from back ground degradations after applying these all methods.**

Keywords- Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.

I. INTRODUCTION

Image processing is a famous and most interested area for researchers. Use of computer based algorithms on digital images to perform image related operations is known as image processing. The textual matter has become simple to see and easy to understand due to imaging technology. All of our general activities are connected with the image and its processing. Historical documents such as novels or scripts or confidential deeds and documents are being preserved by storing them into an image format. So that our next generation is capable of witnessing these old documents.

Bifurcation of text and background from poorly degraded document images is a difficult task between the document background as well as the foreground text of various document images due to the higher background variation. Due to low quality papers, documents fail to preserve the text written on it and gradually the text becomes unreadable. Sometimes the documents get degraded due to some natural problems. There should be an efficient technique to recover these degraded document images so that it can be converted into the readable format. This paper presents a new image binarization technique for the better and accurate recovery of such document images. The Binarization of image is performed in the four stage of document analysis and to separate the foreground text from the document background is its main function. An appropriate document image binarization method is important for recovering document image with the help of processing tasks such as Contrast Enhancement. This technique avoids using canny's, [1] edge algorithm for edge detection , instead canny's edge algorithm is not at all included as the intention of using canny's edge algorithm was to increase the efficiency of the text stroke. But after a long research and analysis, grayscale method is used to sharpen the edges found by Otsu thresholding and edge detection method.

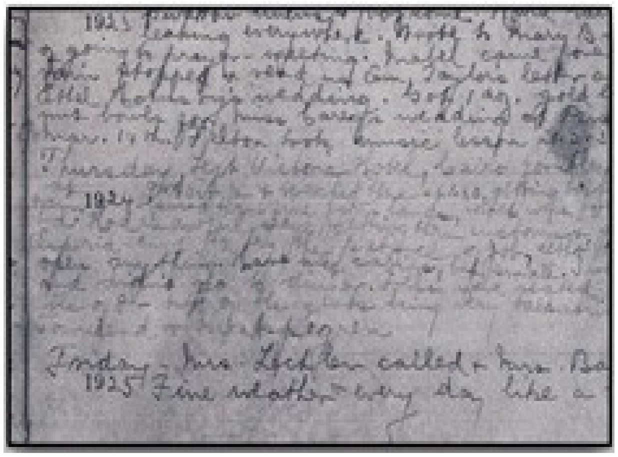


Figure 1: Shows example of degraded image

II. LITEERATURE SURVEY

For document image binarization, many techniques have been developed. Complexity of the existing method is more and consequently the cost to recover the data. The resulting binarization process is slow for large images. Caused by non-uniform illumination, shadow, smear or smudge it does not accurately detect background depth and due to very low contrast without obvious loss of useful information as in Figure 1.The existing system is not able to produce accurate and clear output [8] . This output may include the contents of some background degradations. The Table 1 shown below shows the comparison of the existing systems and their various values.

Table 1. Comparison of various methods

Methods	PSNR	NRM	MPM
OTSU	17.51	9.77	1.35
SAUV	15.96	16.31	1.96
NIBL	15.73	19.06	1.06
BERN	8.57	21.18	115.98
GATO	15.12	21.89	0.41
LMM	17.83	11.46	0.37
BE	18.14	9.06	1.11
PROPOSED METHOD	20.12	6.14	0.25

A. Bolan Su, Shijian Lu,Chew Lim Tan, "Robust Document Image Binarization technique for Degraded Document Images"

In this system the input image goes through different methods. These methods include contrast inversion, threshold estimation and binarization, [1]. Even though it passes through these all techniques, it is not producing efficient output. The edge detection done by the canny's method is not much efficient to detect all the text strokes. The produced output still contains some background pixels. The flow followed for recovering text from degraded documents is shown in Figure 2.

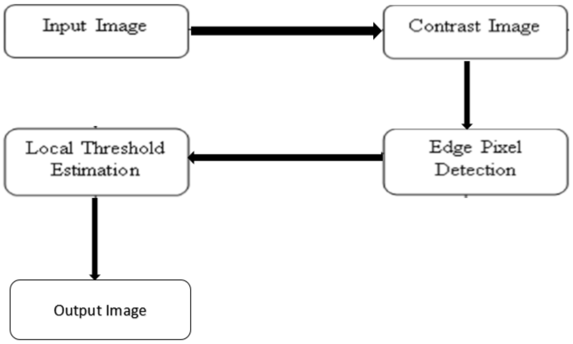


Figure 2: Architecture of existing system.

B. G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Thresholding algorithms for text/background segmentation in difficult document images"

The entropy of the distribution of the gray levels in a scene describes this class of algorithms [9]. As indication of maximum information transfer the maximization of the entropy of the threshold image is interpreted. Entropy system works as shown in Figure 3 below.

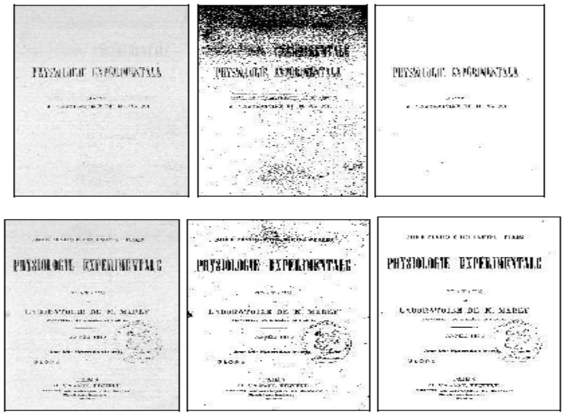


Figure 3: Flow of Entropy system.

C. Rosenfeld and P. De la Torre, ''Histogram concavity analysis as an aid in threshold selection":

Based on the shape properties of the histogram, this category of methods achieves thresholding. In different forms the shape properties are available [6] . From the convex hull, the distance of the histogram is investigated as shown in Figure 4 below, histogram analysis helps us in finding out the hull of the pixels found at foreground.

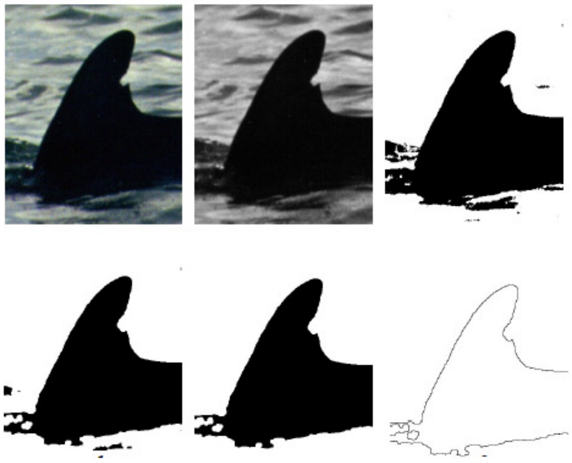


Figure 4. Histogram analysis.

D. I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model":

Normally on the basis of neighborhood pixel values, the threshold values of each pixel are used by the binarization method. If the surface value is lower than the threshold value [11] , those pixels are considered as the background pixels and the other pixels are considered as the foreground pixels.

E. J. Sauvola and M. Pietikainen, "Adaptive document image binarization,":

The contrast value of the text background and the text are focused here. There are two different approaches to find the threshold [4] which are soft decision method (SDM) and text binarization method (TBM). The capabilities of SDMhas noise filtering and tracking of signal, To separate text components from background of the image the TBM is used, due to uneven illumination or noise which is in bad conditions format. At last, the output of these two algorithms is combined together. Proper ways to benchmark utilizes the results against ground and truth measures are important for the algorithm selection process and directions that future research should take. A well-defined performance evaluation shows which capabilities of the algorithm still need refinement and which capabilities are sufficient for a given situation. Figure 5 shows various binarized images at various degraded levels.

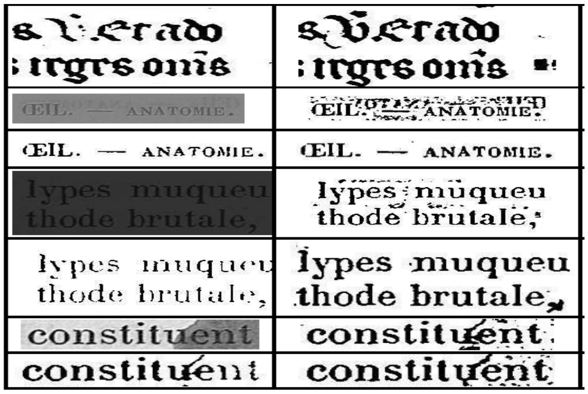


Figure. 5. Sample degraded image

F. L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images,"

This method is evolved by multi resolution approximation; with considerably lower computational complexity, the threshold surface is constructed and is smooth, yielding faster image binarization and better visual performance. By interpolating the image gray levels at the points this method calculates a threshold surface where the image gradient is high. Though high image gradients indicates the probable object edges, where the image gray levels are between the object and the background levels.

G. O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarizationmethods,":

This technique follows a methodology for evaluation of low-level image analysis techniques, using binarization (two-level thresholding) for an instance. Binarization [11] of scanned or photo copied grey scale images is the first step in most document image processing and analysis systems. Appropriate choice of suitable binarization method for an input image domain is a difficult problem. Usually, a human expert viewer evaluates the binarized images according to his/her visual criteria. However, define the objective evaluation, one needs to determine how well the subsequent image analysis steps will perform on the binarized image. This approach is termed as goal-directed evaluation, and it can be used to evaluate other low-level image processing methods as well.

III. Algorithm

A. Gray Scale method:

The gray scale method is the most convoluted, so let's first address it. The most common grayscale conversion routine is "Averaging", and it works like this:

$$Gray = (R+ G + B) / 3$$

Where, R is Red, G is Green, B is Blue. Equivalent to grayscale, this formula generates a reasonably nice, and it is very simple to propose and optimize due to its simplicity. However, this formula is not without defect while fast and simple, relative to the way humans grasp radiance (brightness),for representing shades of grayit does a poor job. For that, we required something a bit more convoluted.

The proposed system avoids using averaging grayscale method. Luminance grayscale method is much more suitable for enhancing the text strokes. Luminance grayscale method is as shown below:

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

B. Edge Width Estimation Algorithm

Requirements: The Image I is the Input Document Image and Edg is the corresponding Binary Text Stroke Edge Image.
Ensure: EW is the Estimated Text Stroke Edge Width
1: Store the width and height of Image I
2: Then for Each Row i in Image I = 1 to height in Ed_g do3: to find edge pixels scan the Image from left to right that meet the following criteria:
a) if its label is 0 (background);
b) if the next pixel is labeled as 1(edge).
4: pixels which are selected in Step 3, Check the intensities in I of those pixels, and the pixels that have a minimum concentration than the coming pixel cut out that next within the same row of I.
5: Then the remaining adjacent pixels are matched into pairs in the same row, and then distance between two pixels in pair will find.

```
6: end for
7: A histogram for those calculated distances is then calculated.
8: Then as the estimated stroke edge width EW use the most frequently occurring distance.
```

C. Post Processing Algorithm

```
Require: I is the Input Document Image, B is the initial Binary Result B and Edg is the Corresponding Binary Text Stroke Edge Image.
Ensure: B f  which is the Final Binary Result
1: all the stroke edge pixel's connect components in Edg are find out.
2: The pixels which do not connect with other pixels remove that pixels.
3: for every remaining pixels (i, j) of edge: do
4: After that its surroundings pairs are taken: (i, j – 1) and (i, j + 1)
   (i – 1, j) and (i + 1, j);
5: if the pixels in the identical pairs belong to the identical class(Both text or background) then.
6: Then to forefront Class (text) Assign the pixel with lower intensity, and the other to background class.
7: end if
8: end for
9: along the text stroke Boundaries after the document thresholding, remove single-pixel artifacts [4].
10: Then new binary result Store to B f
```

IV. Proposed System Modules

As we discussed, the existing techniques have some limitations. To overcome these limitations our system uses new binarization technique. System having five modules. Figure 6 shows the architrecture and flow of the proposed system.

A. Module of Contrast Image:

Contrast is the difference in luminance and/or color that makes an object clear. In visual approach of the real world, within the same field of view, contrast is the variant in the color and intensities of the object and other objects. The adaptive contrast is computed as shown in Equation (2):

$$C(i,j) = \frac{\left(I_{\max(1,j)} - I_{\min(1,j)}\right)}{\left(\left(I_{\max(1,j)} + I_{\min(1,j)}\right) + \epsilon\right)} \quad (1)$$

$$C_a(i,j) = \alpha C_{\square}(i,j) + (1-\alpha)\left(I_{\max(1,j)} - I_{\min(1,j)}\right) \quad (2)$$

Where $C(i, j)$ denotes the local contrast in Equation 1 and$(I_{max}(i, j) - I_{min}(i, j))$ refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 empirically. α is the weight between local contrast and local gradient that is controlled based on the document image statistical information

Here we are going to use adaptive contrast which is contribution of the two methods. First one is the local image contrast, it is nothing but the inversion of the actual image contrast. It only create an opposite contrast image. Second one is local image gradient. In that we are adjesting gradient level of background pixels. Gradient of image is a variation in the contrast level.

B. Module to find the edges

For detection of the edges of each pixel we are using otsu edge detection algorithm. The contrasted image which is further processed for edge detection is an important phase in the project. This will produce the border of the pixel around the foreground text. Pixels are classified intotwo parts, background pixels and foreground pixels. A foreground pixel is the area included within text stroke. And a background pixel is the degraded pixel. From text stroke image construction we obtain the stroke edge of the predicted text patterns found on the degraded document. The constructed contrast image consist a clear bi-modal pattern. For performing clustering based image thresholding or gray level image reduction the Otsu's method is very useful. This algorithm consist of two classes of pixels following bi-modal histogram, then separating the two classes it calculates the optimum threshold so that there is minimal combined spread.

C. Grayscale Conversion:

The Edge Stroke Image obtained from the second module is then transformed to image that are grayscale so as to sharpen the edges of the text stroke detected and thereby increase the efficiency of the further modules.

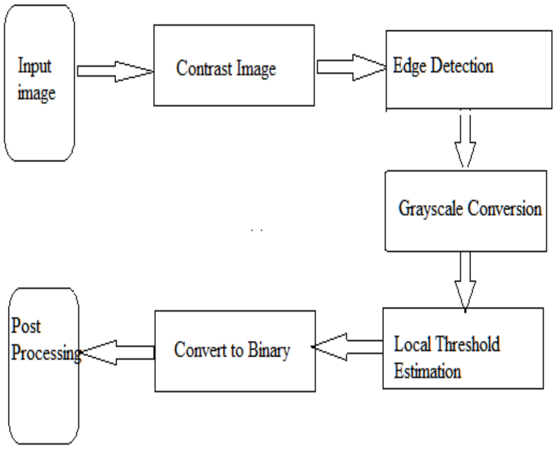


Figure 6. Proposed System Architecture

D. Local threshold Estimation:

The detected text stroke from edge text detection method is evaluated in this method. Here we are creating separation of pixels into two types. We are deciding one threshold value. Depending on that threshold value and pixel value comparison, pixels are categorized as foreground pixels or background pixels.

E. Module to convert into binary:

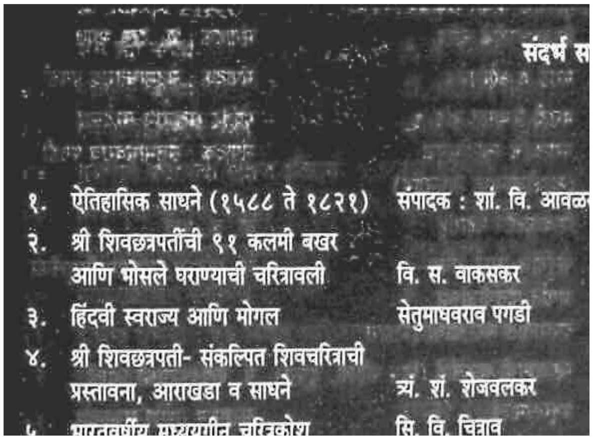
The threshold estimated image is then converted into binary format i.e. 1 and 0.The image pixels at background are marked as 0 and image pixels at foreground are marked as highest intensity i.e. 255 in this case and then combining both to form a bimodal clear image.

F. *Post Processing Module:*

Binarization creates bifurcation of foreground and background pixels in image. But due to variation in background intensities and irregular luminance, it still shows some background pixels on the recovered document image. So we use post processing to avoid such pixels being displayed on the recovered image. And it returns a clear image which consists of actual text. We can easily observe the changes in output image and input image. Output image contain clean and efficient text.

V. RESULTS AND ANALISYS

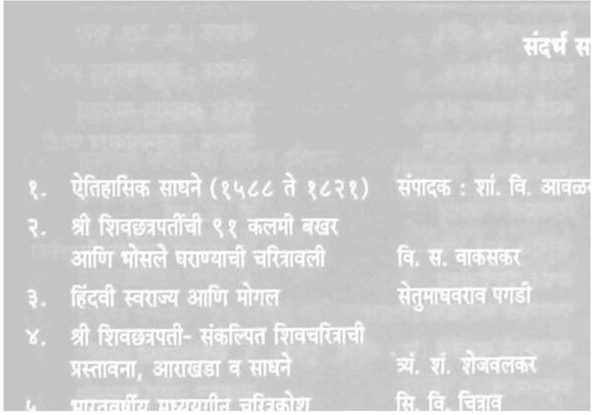
In this project we are accepting degraded image as an input to our system. Suppose we have used following image shown in Figure 7 as an input image.



संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकल्पित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर
५. भारतवर्षीय प्राचीन चरित्रकोश सि. वि. चित्राव

Figure 7. Original Image

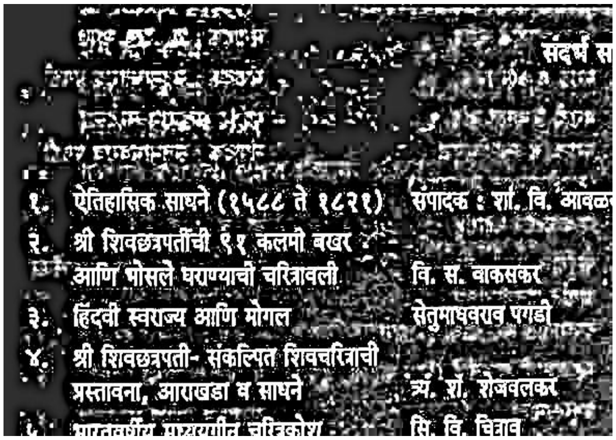
Very first operation on this image will be contrast enhancement. Figure 8 shows the output of the first module (Contrast Module). Here we are applying both local image contrast and local image gradient on same image.



संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकल्पित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर
५. भारतवर्षीय प्राचीन चरित्रकोश सि. वि. चित्राव

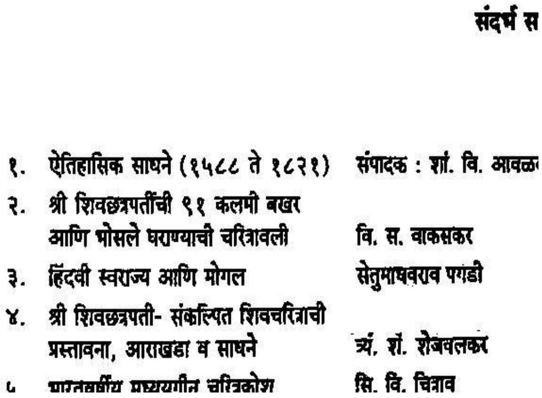
Figure 8. Contrast Image

After the contrast image, the output image undergoes process of next module i.e. edge text detection. Figure 9 shows the output of the Edge detection module. Here we are applying Otsu's thresholding and text stroke edge detection method to effectively detect edges.



संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकल्पित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर
५. भारतवर्षीय प्राचीन चरित्रकोश सि. वि. चित्राव

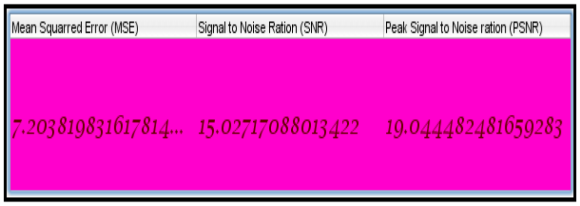
Figure 9. Edge detected Image



संदर्भ स
१. ऐतिहासिक साधने (१५८८ ते १८२१) संपादक : शां. वि. आवळ
२. श्री शिवछत्रपतींची ९१ कलमी बखर
आणि भोसले घराण्याची चरित्रावली वि. स. वाकसकर
३. हिंदवी स्वराज्य आणि मोगल सेतुमाधवराव पगडी
४. श्री शिवछत्रपती- संकल्पित शिवचरित्राची
प्रस्तावना, आराखडा व साधने त्र्यं. शं. शेजवलकर
५ भारतवर्षीय प्राचीन चरित्रकोश सि. वि. चित्राव

Figure 10. Final Image

Figure 10 shows the final image which is produced by our system. The final image is very clear easily displays the text extracted from the degraded documents.



Mean Squarred Error (MSE)	Signal to Noise Ration (SNR)	Peak Signal to Noise ration (PSNR)
7.203819831617814...	15.02717088013422	19.044482481659283

Figure 11. System evaluations Parameters

The Figure 11 shows that, evaluation of system parameter, which should be produced by the system. The PSNR value is nothing but Peak Signal to Noise Ratio which can be computed as shown in Equation 3:

$$\text{PSNR} = 10 \text{ Log } (C^2/\text{MSE}) \quad (3)$$

Where C is a constant and can be defined as 1 and MSE is Mean Square ERROR.

VI. Expected Results

Contrast operation must produce equivalent output so that it will become easy to detect text strokes in Otsu's method. System should produce an image which contains only foreground text. Text on the Output file or a recovered file produced by system will be in readable format. And it should not contain any background degradations. System should produce efficient evaluation parameters which much be better than existing mechanisms.

VII. Conclusion

Thus we can conclude that this method can create more efficient output than other existing techniques. This can become very useful to retrieve original data from degraded documents. This paper uses gray scale method to sharpen the edge strokes which not only increases the efficiency of the proposed system but by removal of canny's edge detection algorithm, accuracy also increases to a higher extent and complexity of the system reduces. Finally system produces image containing only foreground text. At the end we evaluate the efficiency parameter of our system. The evaluation parameters show that the entire system works with great efficiency and produces much more efficient output as compared to existing systems.

Acknowledgment

Author would like to take this convenience to explicit our deep obligation and deep view to Prof. Sonali Rangdale, for her excellent advice, beneficial feedback and constant confidence throughout the duration of the project. Her beneficial suggestions were of enormous help throughout my project work. Her discreet criticism kept me working to make this project in a much better way. Working under her was an extremely appreciative experience for me.

References

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, “Robust Document Image Binarization Technique for Degraded Document Images ” ieee transactions on image processing, vol. 22, no. 4, april 2013

[2] 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.

[3] B. Su, S. Lu, and C. L. Tan, “Binarization of historical handwritten document images using local maximum and minimum filter,” in Proc. Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159–166.

[4] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “H-DIBCO 2010 handwritten document image binarization competition,” in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. 2010, pp. 727–732.

[5] 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

[6] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR 2011 document image binarization contest (DIBCO 2011),” in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 1506–1510.

[7] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” J. Electron. Imag., vol. 13, no. 1, pp. 146–165, Jan. 2004.

[8] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, “Comparisonof some thresholding algorithms for text/background segmentation in difficult document images,” in Proc. Int. Conf. Document Anal.Recognit., vol. 13. 2003, pp. 859–864..

[9] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, “Thresholding algorithms for text/background segmentation in difficult document images” Informatica 38 (2014) 329–338 329.

[10] I.K. Kim, D.W. Jung, and R.H. Park, “Document Image Binarization Based on Topographic Analysis Using a Water Flow Model,” Pattern Recognition, vol. 35, pp. 265-277, 2002.

[11] O. D. Trier and A. K. Jain, “Goal-directed evaluation of binarizationmethods,”: IEEECS Log Number P95138.