Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

Document Image Binarization Technique For Enhancement of Degraded Historical Document Images

Manish Deelipkumar Wagh[1], Mayur Yashwant Bachhav[2] and Vijay Balasaheb Gare[3]
[1,2,3]*Department of Information Technology, MVPs Karmaveer Adv. Baburao Ganpatrao Thakare College of Engineering Nashik*

Abstract— Segmenting text from image of rough, degraded historical document is difficult task because of higher inter variation and intra variation between foreground text and document background of different document images. This can be solved with the help of image binarization technique. We can make use of this technique to clear the historical document, making them usable for further processing. As the image of old historical documents are often in degrade form it is hard to retrieve and understand what is written on them. So it is necessary to find solution to this problem, by taking picture and binarizing it with suitable technique we can able to understand the text so here we developed a image binarization technique. Here we provide new segmentation algorithm in which each pixel has its own threshold value. In old system the contrast image is used as preprocessing step but in this technique gray scale image is used. The pixels are categorized as foreground or background pixels depending upon their comparison with threshold valued. Pixel value is compared to the threshold value. The technique is known as window thresholding .We are doing work on window of size p*q and from each window text stroke of each pixel are extracted .Again local threshold is used to segment the document text that is estimated based on intensity of detected text stroke of edge pixel in local window.

Keywords- Image Processing, Image Binarization, Image Segmentation.

I. INTRODUCTION

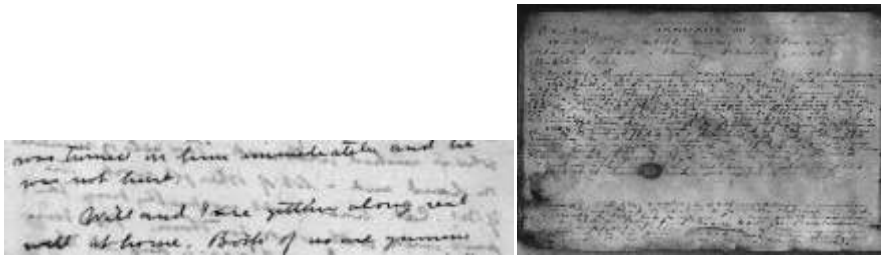
Binarization of Document Image is required in the essential stage for investigation of report and the undertaking of Document Image Binarization is to isolate the frontal area information from the archive foundation information. A brisk and right archive picture binarization strategy is critical for following handling of record picture undertakings for instance optical character acknowledgment [4].

Binarization is a dynamic research zone in the field of Document Image Processing. Binarization changes over dark picture into binarized picture. Report picture binarization is the most imperative stride in pre-preparing of examined archives to spare all or greatest subcomponents such us content, foundation and picture [1]. Binarization registers the edge esteem that separate protest and foundation pixels. Shading and dark level picture handling devours heaps of execution powers. Be that as it may, binarized pictures diminish the computational load and expand effectiveness of the given frameworks [7] [10]. Binarization has numerous points of interest, for example, medicinal picture preparing, archive picture investigation, face acknowledgment and so on. Binarization can be ordered into two classifications: worldwide and versatile. Worldwide techniques depend on the finding a solitary edge esteem for the whole picture, and versatile strategies depend on the nearby data acquired from the hopeful pixel and is required for the estimation of limit esteem for each pixel. On the off chance that elucidation of info picture is not comparative (uniformly lit up), neighborhood strategies may perform better [2]. On the off chance that picture has break even with light then worldwide strategies can work better. In any case, worldwide techniques can't deal with any of the picture debasement and not ready to uproot commotion. Nearby techniques are essentially additional tedious and computationally costly. Quick and precise calculations are essential for Document Image Binarization Using Image Segmentation frameworks to perform operations on report pictures [8]. To accelerate the preparing,

parallel execution of a calculation should be possible utilizing Graphics Processing Unit (GPU) as broadly useful calculation equipment; programmability and ease make it profitable [1].

Record Image Binarization is performed in the preprocessing arrange for archive examination and it intends to section the forefront message from the foundation of report picture. A speedy and appropriate archive picture binarization strategy is imperative for the guaranteeing record picture preparing errands, for example, Document Image Binarization Using Image Segmentation Though report picture Binarization has been contemplated for a long time, the thresholding of debased report pictures is still an unsolved issue because of the high entomb/intravariation between the content stroke and the report foundation crosswise over various report pictures [6][7]. The manually written content inside of the corrupted records frequently indicates different sorts of issues regarding the stroke splendor, stroke association, stroke width, and report foundation. Moreover, old archives are frequently corrupted by the discharge through where the ink of the other side leaks from end to end [9]. Moreover, old records are regularly corrupted by various sorts of imaging antiquities. These distinctive sorts of Document corruptions are prone to affect the record thresholding mistake and make debased archive picture binarization a major test to best in class procedures [3].

The Thresholding of debased archive pictures is an unsolved issue from parcel of days because of impact of buries variety or `intra variety between the record foundation and content stroke crosswise over various report pictures [11]. The content which is composed by submit the harsh reports demonstrates a specific measure of distinction regarding the edge width, brilliance of stroke, association in the middle of line, and foundation designs [14].



(a) (b)

Fig 1: Two degraded document image examples (a) and (b) are taken from Internet randomly as an example

Notwithstanding it, old column verifiable records are some of the time debased by the maturing and leaking of ink, for the situation of ink leaking the opposite side of report ink leaks through the front. Some environmental variables like temperature, dampness are outstanding focuses [5]. The essential point of this framework is that the framework ought to clear all the corrupted foundation so we will get clear binarized picture organization of the archive.

II. RELATED WORK

Picture binarization is executed as preprocessing venture in picture record examination. Pointing division of content picture and whatever other example display on specific report picture. Division frontal area from foundation and speaking to it in a required shape. The procedure ought to be quick and precise [17, 15]. There are a few methods which are proposed for picture binarization as a preprocessing step [18].

Numerous thresholding strategies have been accounted for archive picture binarization. The same number of debased reports doesn't have a reasonable bimodal example; worldwide thresholding is generally not a suitable approach for the corrupted archive binarization. Versatile thresholding, which assesses a neighborhood edge for every archive picture pixel, is frequently a superior way to deal with

manage varieties inside debased record pictures []16. The nearby picture differentiate and the neighborhood picture slope are exceptionally valuable components for portioning the content from the record foundation in light of the fact that the report message more often than not has certain picture complexity to the neighboring archive foundation [1]. They are extremely powerful and have been utilized as a part of numerous archive picture binarization procedures.

Bolan Su, Shijian Lu, and Chew Lim Tan, Who are the Senior Member of IEEE show the methodology for image binarization in their research Robust Document Image Binarization Technique for Degraded Document Images in that they use a techniques such as Canny Edge Detection for Text Stroke, Otsu Algorithm, Contras mage creation. The Canny Technique involves lots of parameter setting; sometimes result into miss segmentation, contrast images are hard to understand

III. PRPOSED METHOD

Image binarization method for corrupted report pictures is the approach for the making the vigorous archive coherent, usable and clear them for future utilize which is accomplished by procedure of picture binarization, we are attempting to improving the binarization handle, framework work in three stages First of all the info picture i changed over to grayscale picture, then in second stage picture is divided to discrete frontal area message from foundation to acquire clear binarized picture, then in third stage some post preparing is done to make it more exact. binarization procedure for corrupted record pictures is the approach for the making the vigorous verifiable archive intelligible, usable and clear them for future utilize which is accomplished by procedure of picture binarization, we are attempting to improving the binarization prepare, framework work in three stages. Most importantly the information picture i changed over to dark scale picture, then in second stage picture is divided to partitioned frontal area message from foundation to get clear binarized picture, then in third stage some post preparing is done to make it more exact

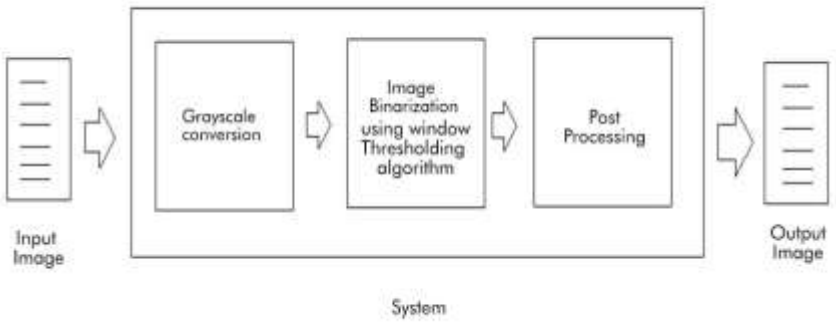


Fig 2: Block Diagram for Proposed System

The diagram shows the architecture of proposed system. In this architecture input is taken in the form of form of image file by using scanner or camera or mobile phone. After taking the input to our system, system first converts it into gray scale image which is used as preprocessing stage for image processing. The gray scale images go through binarization so that the unwanted degraded part will get clear. After that post processing need to apply, this increases the intensity and joins the disconnected edges. Finally we will get clear binarized image as an output. Let us see all of them one by one.

Conversion of image into Gray image

Gray scale pictures are unmistakable from one-piece bi-tonal high contrast pictures, which with regards to PC imaging are pictures with just the two hues, dark, and white. Dim scale pictures have

numerous shades of dark in the middle. Report picture binarization alludes to the transformation of a dark scale picture into a twofold picture. The Algorithm is utilized to change over the Degraded Image into the dim scale by serial approach. The calculation is executed on each of the window of the picture serially and the picture is changed over to dark scale.

Algorithm:

Input: document image
Output: grayscale image

Process:

For each rows and column Ri, Cj respectively
Color C = getPixel (Ri, Cj)
R = c.getRed ()
G = c.getGreen ()
B = c.getBlue ()

Avg = (R+G+B) / 3

NewColor Cg = (avg, avg, avg)
SetPixel (Ri, Cj, Cg)
end for;

The Image Binarization (Window Thresholding Method)

Image segmentation is a key innovation in picture preparing, and limit division is one of the strategies utilized regularly. Gone for that stand out edge or a few limits is set in conventional edge based division calculation, it is hard to separate the mind boggling data in a picture; another division calculation that every pixel in the picture has its own edge is proposed. In this calculation, the edge of a pixel in a picture is evaluated by figuring the mean of the dim scale estimations of its neighbor pixels, and the square difference of the dark scale estimations of the neighbor pixels are likewise ascertained as an extra judge condition, so that the aftereffect of the proposed calculation. Truth is told the proposed calculation is equivalent to an edge indicator in picture handling. Exploratory results show that the proposed calculation could deliver exact picture edge, while it is sensible to evaluate the edge of a pixel through the factual data

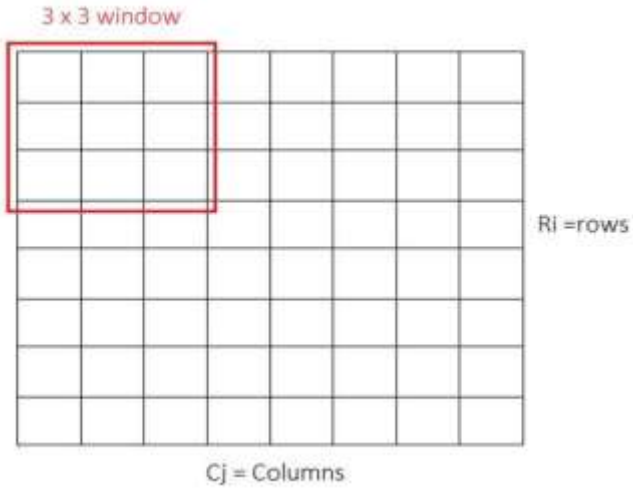


Fig 3: Figure shows how Image is get divided to get a window of required size

For segmentation purpose the gray scale image is divided into set of pixels that we call here as window. Window size can be 30*30 pixels or 40*40 pixels.

Algorithm:

```
Input:      Grayscale image Gi;
            Window size W;
            Threshold value: Th;

Process:  for each row Ri and column Cj;
          for each window W ;

          Color C = getPixel( Ri, Cj );

                calculate sum of pixels;
              avg = sum / W
                        for each window
                        if ( currentPixel - avg ) <Th

                         background // 0
                         else
                         foreground // 1
              end for;
          end for;
          end for:
```

Post Processing

The post processing of digital images using parallel computing, particularly for gray scale, brightening, darkening, thresholding and contrast change. The point to point technique applies a transformation to each pixel on image concurrently rather than sequentially.

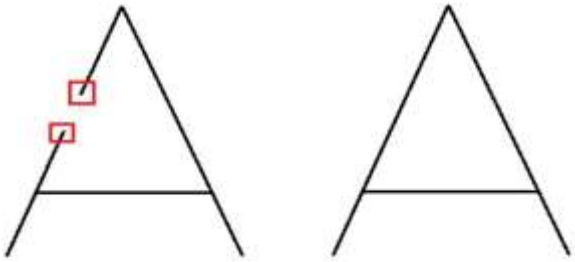


Fig 4: Post processing to improve result.

Algorithm:

Input: Binarized Image
Output: Clear Formatted Image
Process:

- Find out all the connect components of the stroke edge pixels in Edg

- Remove those pixels that do not connect with other pixels.

for each pixel
for remaining edge pixels(i,j)

get(i-1,j) &&get(i+1,j);
get(i,j-1)&& get(i,j+1);

if the pixels in the same pairs belong to the same class (both text or background)

then Assign the pixel with lower intensity to foreground class (text), and the other to background class.

end if
end for

IV. SYSTEM RESULTS

The proposed framework utilizes different calculations. The corrupted record picture will go as the information to the framework the picture will be changed over to dim scale first by utilizing dim scale calculation. At that point the dark scale archive picture will be go to ascertain the force of the picture then the picture division calculation will be apply on the report picture. The archive picture will be separated into different fragments for producing yield. Finally the post handling calculation will be apply to distinguish the feed edges of the words naturally and the unmistakable binarized archive picture will be produced. The information to the framework will be the debased archive picture. We can likewise give the data to the framework by passing different corrupted pictures. For the info to the framework we are utilizing the information set gave by the DIBCO (2009) and DIBCO (2011). The DIBCO gives different information sets we are utilizing the information sets as data to the framework.

Gray Scale Conversion: - Passing the degraded image to the for getting clear binarized image we will first convert the image into the gray scale form by using serial and parallel approach.

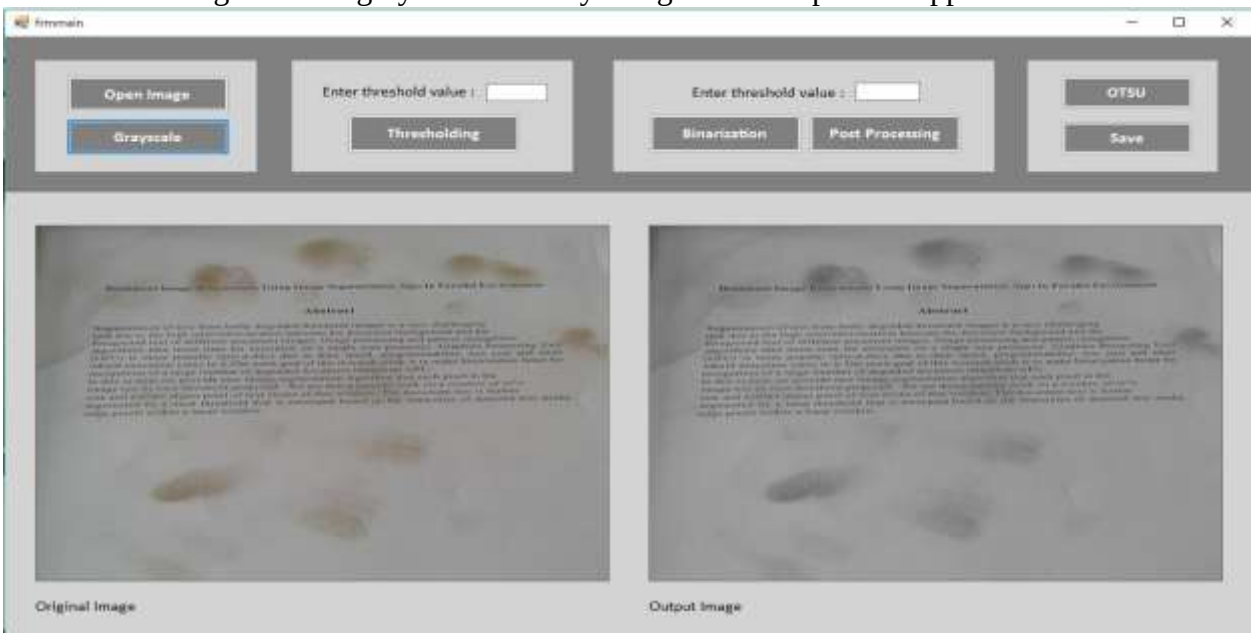


Fig: Gray Scale Conversion

Output: - The final output of the framework will be clear binarized picture. For that we will apply different calculation first we will change over the picture to the dark scale and after that force estimation will be done of every window of the picture. After that we will apply the picture division calculation for producing clear binarized picture. The preparing will be done in two methodologies serial and parallel and the count will be finished.

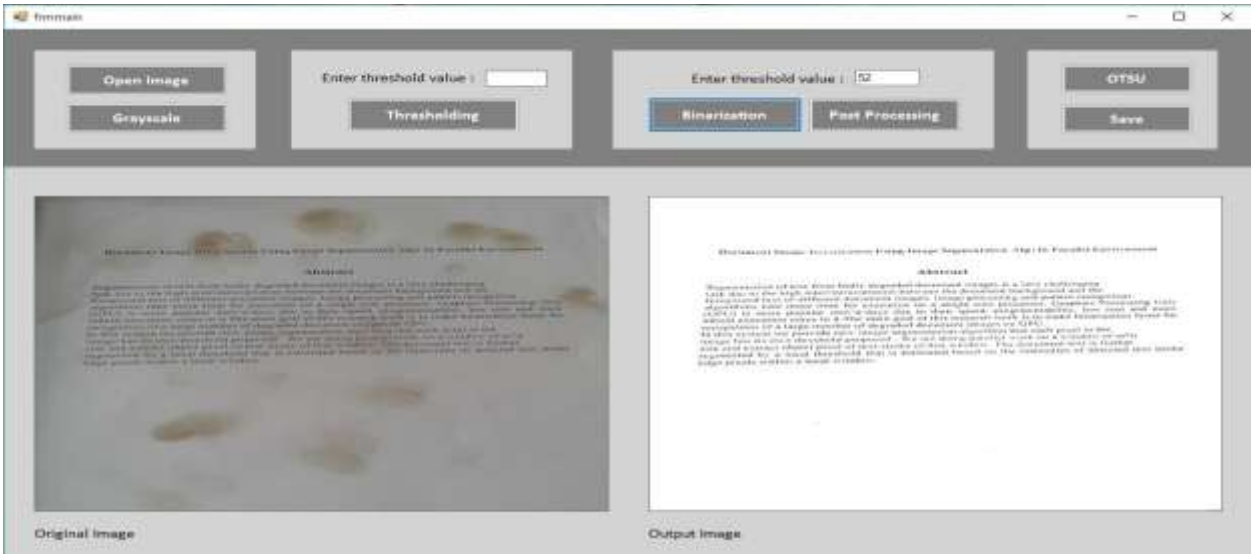


Fig: Output of the system.

V. SYSTEM APPLICATION

All things considered, there are loads of reports that are debased due maturing, leaking of ink. This makes the archive ambiguous configuration. We can't comprehend the content which is there in the report. So to recoup the content from the muddled foundation we have built up this framework. This framework performs the division of content from picture of harsh, debased recorded archive. This binarization of record picture is required in the essential stage for investigation of archive. The Technique can be connected where the picture investigation is required, in picture chronicles, to digitally save reports, it can be utilized as a sub framework for another frameworks.

VI. CONCLUSION

In this project we are developing this system for image binarization intended for document images. As the image of old historical documents are often in degraded form it is hard to retrieve and understand what is written on them. So it is necessary to find solution to this problem, by taking picture and binarizing it with suitable technique we can able to understand the text so here we developed a image binarization technique, called window thresholding based on threshold segmentation in this algorithm each pixel in an image has its own threshold, which is estimated by calculating the statistical information of its neighborhood pixels. We also trying to improve the binarization result by doing some post processing techniques.

VII. ACKNOWLEDGMENT

With deep sense of gratitude we would like to thanks all the people who have lit our path with their kind guidance. We are very grateful to these intellectuals who did their best to help during our project work. The special gratitude goes to Prof. J. R. Suryawanshi for her precious Guidance in completion of this work.

REFERENCES

[1] Robust Document Image Binarization Technique for Degraded Document Images by Bolan Su, Shijian Lu and Chew Lim Tan, Senior Member, IEEE, IEEE Transactions on Image Processing, VOL 22, No 4, April 2013.
[2] 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. 303314, Dec. 2010.
[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 Documen Analysis. Syst., Jun. 2010, pp. 159166
[4] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, Comparison of some thresholding algorithms for text/background segmentation in difficult document images, in Proc. Int. Conf. Document Analysis. Recognit., vol 13 2003, pp. 859864.
[5] A. Brink, Thresholding of digital images using two-dimensional entropies, Int Jr Pattern Recognit., vol. 25, no. 8, pp. 803808, 1992.
[6]] O. D. Trier and A. K. Jain, Goal-directed evaluation of binarization methods, IEEE Trans. Pattern Analysis. Mach. Intell., vol. 17, no. 12, pp. 11911201, Dec. 1995
[7] O. D. Trier and T. Taxt, Evaluation of binarization methods for documentimages,IEEE Trans. Pattern Analysis. Mach. Intell., vol. 17, no. 3, pp. 312315, Mar. 1995.
[8] A. Brink, Thresholding of digital images using two-dimensional entropies, Int JrPattern Recognit., vol. 25, no. 8, pp. 803808, 1992.
TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013
[9] J. Bernsen, Dynamic thresholding of gray-level images, in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 12511255
[10] I.-K. Kim, D.-W. Jung, and R.-H. Park, Document image binarization based on topographic analysis using a water flow model, Pattern Recognit., vol. 35, no. 1, pp.265277, 2002.K. Elissa, "Title of paper if known,"
[11] W. Niblack, An Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1986
[12] J. Parker, C. Jennings, and A. Salkauskas, Thresholding using an illumination model, in Proc. Int. Conf. Doc. Anal. Recognit., Oct. 1993, pp. 270273.
[13] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, Text extraction and document image segmentation using matched wavelets and MRF model, IEEE Trans. Image Process., vol. 16, no. 8, pp. 21172128, Aug. 2007.
[14] E. Badekas and N. Papamarkos, Optimal combination of document binarization techniques using a selforganizing map neural network.Eng. Appl. Artif. Intell., vol. 20, no. 1, pp. 1124, Feb. 2007.
[15] B. Gatos, I. Pratikakis, and S. Perantonis, Improved document image binarization by using a combination of multiple binarization techniques.and adapted edge information, in Proc. Int. Conf. Pattern Recognit., Dec. 2008, pp. 14.
[16] Q. Chen, Q. Sun, H. Pheng Ann, and A double-threshold image binarization method based on edge detector, Pattern Recognit., vol. 41, no. 4, pp. 12541267, 2008
[17] An Image Segmentation Algorithm in Image Processing Based on Threshold Segmentation, Shiping Zhu,Qingrong Zhang, Kame Belloulata, Third International IEEE Conference on Signal-Image technologies and Internet-Based System 2011