Global Scientific JOURNALS

GSJ: Volume 7, Issue 12, December 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

Degraded Documents Data Preservation Using Digital Image Processing

Miss. Maheshwari H. Ukirde
Department of Information Technology
Siddhant College of Engineering,
Sudumbare, Pune
Email- maheshwariukirde@gmail.com

Prof. Sonali Rangdale
Department of Information Technology
Siddhant College of Engineering,
Sudumbare, Pune
Email- sonali_rangdale@rediffmail.com

Abstract*—Documents can be a valuable source of information but often they suffer degradation problems, especially in the case of historical documents, such as strains, background of big variations and uneven illumination, ink seepage, etc. Binarization techniques should be applied to remove the noise and improve the quality of the documents. Collections of historical and old document images care commonly provided to public through digital libraries. Specialized processing is required to these document images for removing background noise in order to become more legible. A hybrid binarization approach is proposed in this paper for improving the quality for the old documents. Combination of global and local thresholding techniques are used for the same .Initially, a technique named global thresholding is applied to the whole image. The image area that still has background noise are detected and the technique is again re-applied to each area separately. Therefore, a better adaptability is achieved for the algorithm where various kinds of noise re exist in different areas of same image. Advantage of applying global thresholding, is that it avoids the computational and time cost of applying a local thresholding in the entire image. Hence it is indicated that this technique is pretty effective in removing background noise and improving the quality of degraded images.

Index Terms — Image enhancement, Iterative global thresholding, Local thresholding, hybrid binarization, Noise

I. INTRODUCTION

There are libraries in the world which has collections of historical & ancient documents which are of great scientific & Cultural importance. To maintain the quality of the originals it is essential that the documents are transformed into digital form, by doing these scholars are provided to have full access to the information. Degradation problems are quiet common in these documents. Few factors that impede (in many cases may disable) the legibility of the documents are strains, big variations, uneven illumination, presence of smear, seepage of ink, etc.

Before libraries expose them to public view it is important to remove noise from historical document images and improve their quality and appropriate filtering methods should be developed as well. Noise is considered anything that is irrelevant with the textual information (i.e., foreground) of the document image. Binarization is used as a standard procedure to convert a grey-scale image to binary form in image analysis systems. An ideal binarization algorithm would be able to perfectly discriminate foreground from background, helping in removing any kind of noise that obstructs the legibility of the document image. Binary image is ideal for pre-processing like recognition of the contents by applying OCR techniques etc., discrimination of printed from handwritten text, etc. In the framework of a library collection of historical and ancient documents, the document images in many cases do not need further processing apart from removing the background noise and leave some “traces of time” behind, which is intended to be exposed to public view. In such cases it’s possible for the document images to remain in greyscale form.

Figure 1: Sample Degraded image

The figure1 shows various binarization images at various degraded levels.

Traditional binarization approaches can be divided into two main categories:

1. Global thresholding methods: According to global threshold, the pixels of the image are classified into background or text. Generally, such methods are simple and fast. If the background noise is unevenly distributed in the entire image, then they cannot be easily adapted. (e.g., smear or strains).

2. Local thresholding methods: According to local threshold determined by their neighbouring pixels, the pixels of the images are classified into background or text.

Adapting these methods are of more help and also can deal with different kind of noise existing in one image. On the other side it's more time consuming and computationally expensive.

From another point of view, binarization approaches can be divided as follows:

- General-purpose methods: Image can be dealt with this method. Hence, specific characteristics of document images are not taken into account.
- Document image-specific methods: Advantage of document image characteristics (e.g., background pixels is the majority, foreground pixels are in similar grey-scale tones etc.) are taken.

In many cases, such methods are variations of general-purpose approaches.When dealing with historical document images; the latter approaches should be more effective. Recent results show that general-purpose methods can be more reliable under certain conditions .

Previously, we have presented an Iterative Global Thresholding (IGT) approach that is specifically designed for document images .

This method has the additional advantage of providing the option to maintain the image in grey-scale after the removal of background noise, apart from efficiency inherent in any global thresholding approach, a more familiar form for human readers.

A hybrid approach is proposed in this paper to combine the advantages of local and global thresholding. After the application of IGT to the document image, the areas that are more likely to still include significant amount of noise are selected and, then, IGT is re-applied to these areas separately. This is the main idea. Therefore, document images which have different kind of background noise, and which are unevenly distributed in the entire image can be processed more effectively. Additionally, in comparison to original local thresholding techniques, since only a limited number of areas (instead of the entire image) need to be processed separately, the time -cost of the approach remains on low level. In order to evaluate the proposed approach, a degree is represented in which the legibility of image has been improved.

II. LITERATURE 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. The existing system is not able to produce accurate and clear output. This output may include the contents of some background degradations.

In this system the input image goes through different methods. These methods include contrast inversion, threshold estimation and binarization. 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.

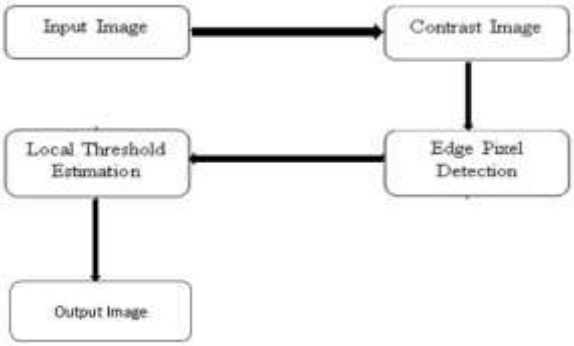

Figure 2: Existing System

There are many methods proposed in global and local thresholding. Based on the variance of pixel intensity, Otsu proposed one of the earlier methods in image binarization. Bernsen calculates local thresholds using neighbours. Standard deviation and local mean is used by Niblack. A method is presented by Sauvola which is specialized on document images that applies two algorithms in order to calculate a different threshold for each pixel. Leedham compares some of the traditional methods on degraded document images, as far as the problem of historical documents is concerned, while a method using a combination of existing techniques is proposed by Gatos. Shi and Yan is also applied to some historical documents from the US library of Congress.

A serialization of k-means algorithm is implemented by Leydier [16] who works with colored document images. These above methods have also used per-processing or post-processing filters for improving the quality.

III. HYBRID BINARIZATION

A. The Proposed Approach:

The Combination of global and local thresholding which is a hybrid approach for improving the quality of historical document images is proposed. A document image is applied a global thresholding approach (IGT) firstly. After that the areas which still contain noise are detected and re-processed separately. If we go more deep, the proposed algorithm consists of the following steps:

i. IGT to the document image is applied.
ii. Detect those areas in which we can still find background noise.
iii. Then re-apply IGT to each detected area separately.

We avoid the cost of applying local thresholding to the entire image, by selecting only specific areas of the image for processing based on local thresholding.

B. Detection of Areas with Remaining Noise:

A simple method is used for detection of areas that needs further processing. The areas that still contain background noise will include more black pixels on average in comparison with the other areas.

The image is divided into segments of fixed size nxn. The frequency of black pixel is calculated in each segment. The segments that satisfy the following criteria are, then, selected as:

$$f(S)>m+ks$$

Where f(S) is the frequency of the black pixels in the segment S while m and s are the mean and the standard deviation of the black pixel frequency considering the segments of the entire page, respectively. The selected segments form areas by connecting neighbouring segments in respect to their original position in the image. The row-by-row labelling algorithm is used.

C. Local Thresholding on Selected Areas:

Based on local thresholding, the areas detected by the previously described procedure are separately re-processed. The IGT method is applied to the corresponding area of the original image, which is for a given area. The iterations stop when either the criterion of formula is satisfied or the number of iterations exceeds the corresponding number of iterations required for the global thresholding on the entire image (from the first step of the proposed approach).

IGT removes a lot of pixels during the first iterations, given that the selected areas have relatively high average density of black pixels. The background noise in the selected areas is more likely to be removed since the area is likely to be more homogeneous than the entire image, in comparison to the application of IGT to the entire page. Generally, this procedure tends to move more pixels of the selected areas to the background in comparison with the previous application of

IGT to the entire image. The foreground is not attenuated unless it consists of different grey-scale tones (e.g. the presence of both printed and handwritten in the same area), if in case a selected area does not contain considerable amount of background noise.

An important factor for the successful re-application of IGT in selected areas is the size of the window nxn used in the procedure of selecting the appropriate areas (described in the previous subsection). A small window size forms more but smaller areas, which can be seen. There is advantage of adapting area in more detail at the part of the image that still has noise. In many cases they do not provide enough information for successfully re-apply the IGT algorithm on them and also the resulting areas are small. Fewer but bigger areas are detected, in case of large window size.

In order to effectively remove background noise, they provide enough information to the IGT algorithm. These areas cannot be easily adapted to a specific part of the image that still contains noise. The final image may contain neighbouring areas that have dissimilar amount of background noise, as a consequence.

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 3 shows the architecture 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):

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

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

Where C(i, j) denotes the local contrast in Equation 1 and(Imax(i, j) – Imin(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 adjusting 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 into two 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.

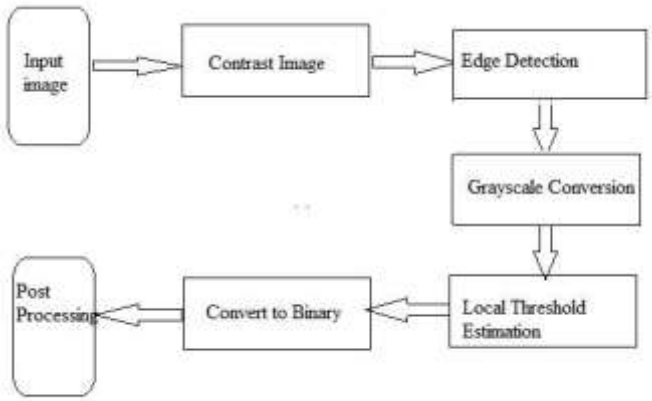

Figure 3: Proposed System Architecture

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

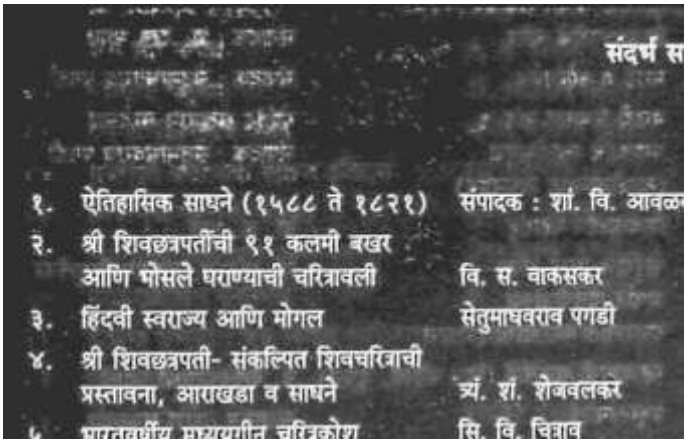

Figure 4: Original Image

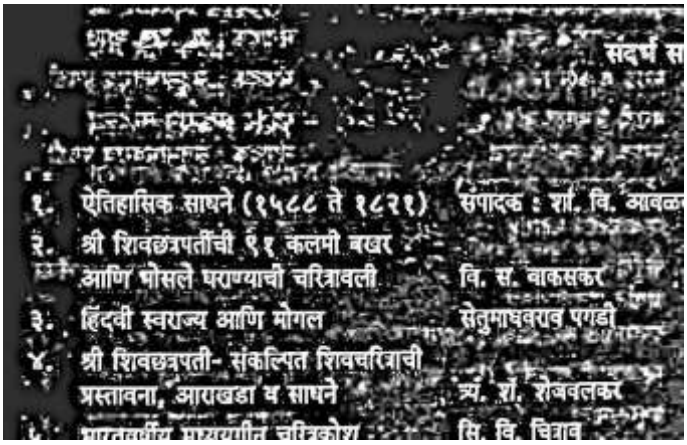

Figure 5: Contrast Image

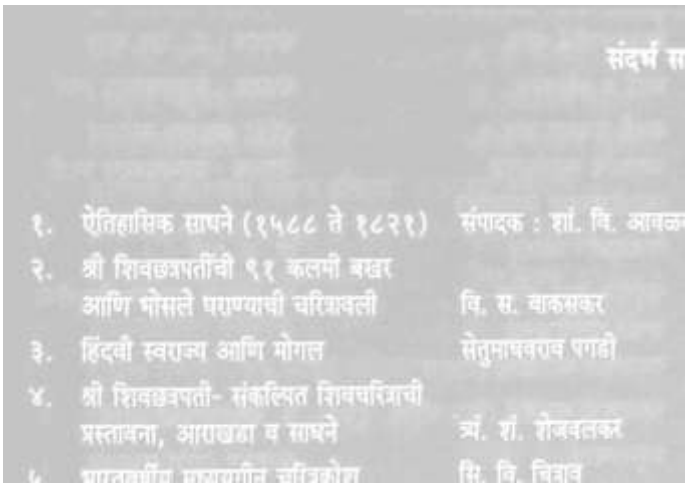

Figure 6: Edge Detected image

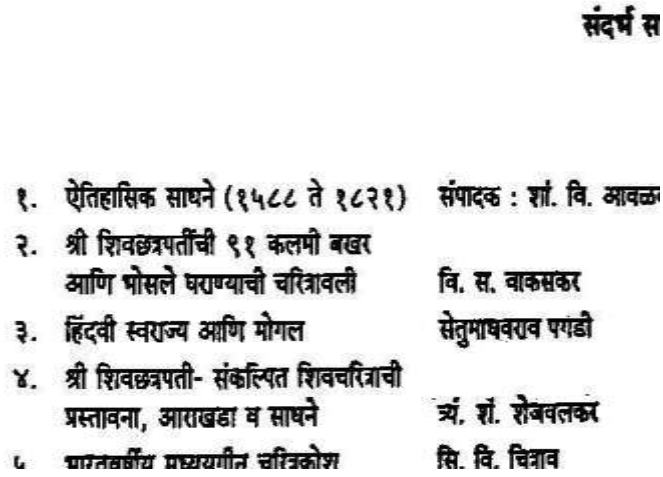

Figure 7: Final Image

VI. 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 grey 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 system.

VII. 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 guidance was an extremely appreciative experience for me.

VIII. REFERENCES

[1] Mrs. Preeti Kale, Dr.S.T.Gandhe, Prof.G.M.Phade, Prof.Pravin Dhulekar "Enhancement of old Images and documents by Digital Image Processing Techniques" in 2015 INTERNATIONAL CONFERENCE ON COMMUNICATION, INFORMATON AND COMPUTING TECHNOLOGY(ICCICT)

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 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..

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

[9] 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.

[10] 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.

[11] 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.