ROBUST DOCUMENT IMAGE BINARIZATION TECHNIQUE FOR DEGRADED DOCUMENT IMAGES

Mr. Yogeshwar Ade [1] , Prof. Sonali Rangdale [2]

[1] Department of Information Technology, Siddhant College of Engineering, Maval, Pune

[2] Department of Information Technology, Savitribai Phule Pune University, Maharashtra, India

ABSTRACT:

Now a days, the Technology is connecting the whole world together by the means of internet. Every segment of our data is in the form of digital document. We can save, copy, and backup our data in digital form. But what about old data which is in the form of traditional paper. Sometimes the old data plays important role in a major tasks. Many of the paper data is being degraded due to lack of attention. Many of these degraded documents have mix up with their front and rear data. To make this frontend separate from backend we have proposed binarized documentation technique. In this we firstly applying the invert contrast mechanism on degraded document. Then we are going to compare that with canny'sedge graph and then we are applying the binarization method on that degraded image. The output of this all technique will produce a clear and binarized image.

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

[1] INTRODUCTION

Imaging plays a key role in many diverse areas, such as astronomy, remote sensing, microscopy or tomography, just to name few. Due to imperfections of measuring devices and instability of observed scene, captured images are blurred, noisy and of insufficient spatial or temporal resolution. Image restoration methods try to improve their quality. The old images has become an unreadable document due to non-maintenance. Sometimes due to some natural problems the images get degraded. Images can be degraded manually to degrade the quality of image.

The image Binarization is performed in the four stage of document analysis and it aims to separate the foreground text from the document background. The accurate document image binarization technique is important for the recovering document image by the help of processing tasks such as Contrast Enhancement. Though document image binarization the thresholding of degraded document images is solved now. It was due to the high inter/intra-variation between the text stroke and the document background across different document images. And we have solve this problem using the collaboration of canny's edge map and binarization method

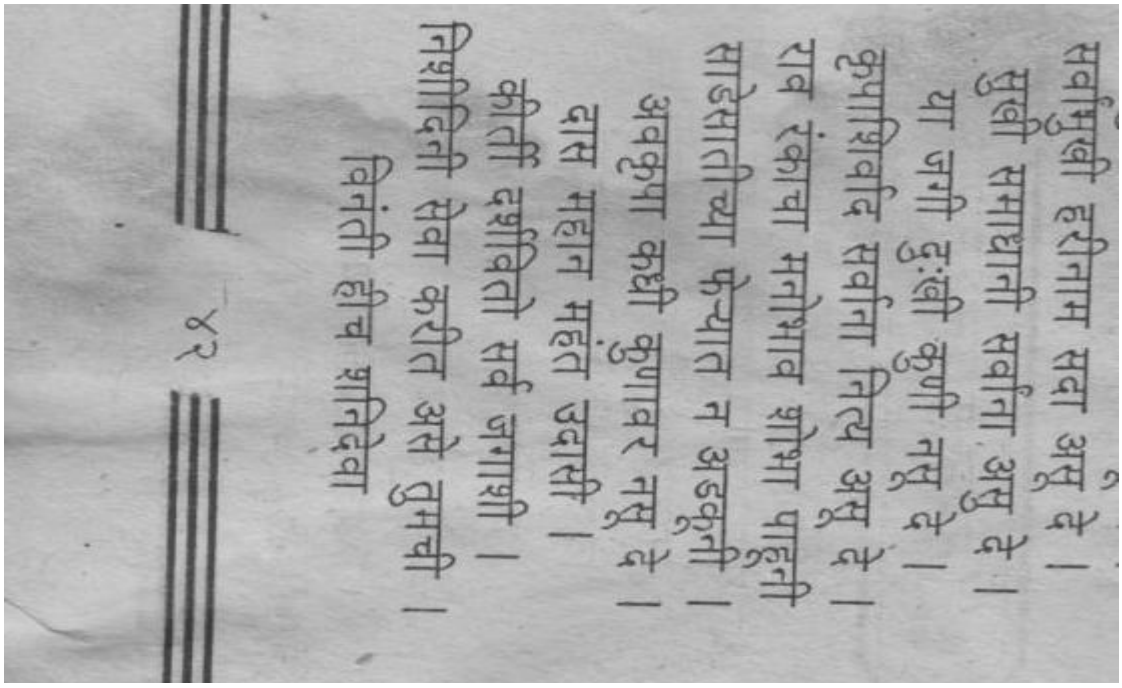

Figure: 1 Sample of a degraded documentimage.

[2] LITERATURE SURVEY

Many techniques have been developed for document image binarization. As many degraded documents do not have a clear pattern and it may be in a bad condition. Thresholding alone is not a good approach for the degraded document binarization. Adaptive thresholding, which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images.

[2.1] RECOVERY METHOD FOR DEGRADED IMAGES

In this paper the degraded document image is reconstructed by using inversion gray scale image contrast. It consist of two algorithms:

- **Edge Width Estimation:**

Require: The Input Document Image *I* and Corresponding Binary Text Stroke Edge Image *Edge*
Ensure: The Estimated Text Stroke Edge Width *EW*

1: Get the width and height of I
2: for Each Row x = 1 to height in Edge do
3: Scan from left to right to find edge pixels that meet the following criteria:
a) its label is 0 (background);
b) the next pixel is labelled as 1(edge).
4: Examine the intensities in I of those pixels selected in Step 3, and remove those pixels that have a lower intensity than the following pixel next to it in the same row of I .

5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
6: end for
7: Construct a histogram of those calculated distances.
8: Use the most frequently occurring distance as the estimatedstroke edge width EW.

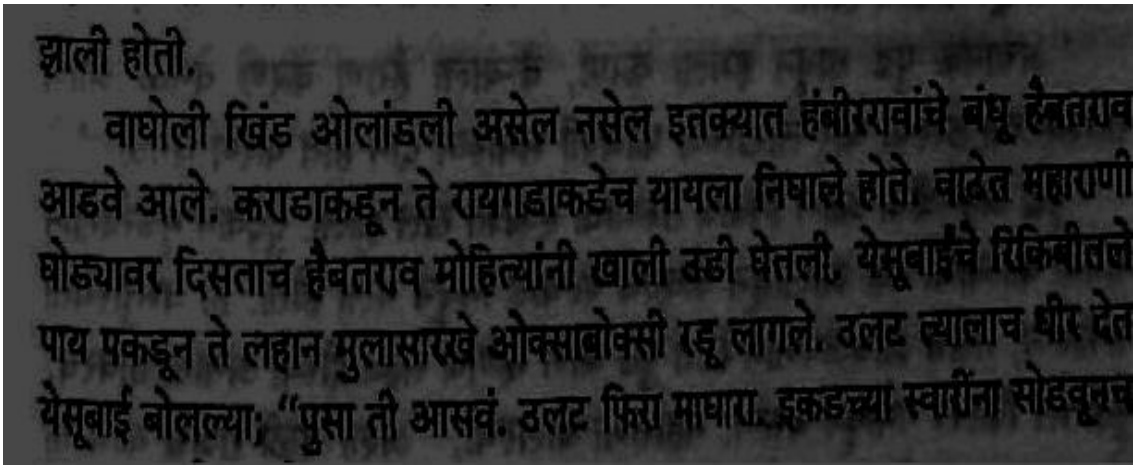
झाली होती.
वाघोली खिंड ओलांडली असेल नसेल इतक्यात हंबीररावांचे बंधू हैबतराव
आडवे आले. कराडाकडून ते रायगडाकडेच यायला निघाले होते. वाटेत महाराणी
घोड्यावर दिसताच हैबतराव मोहित्यांनी खाली उडी घेतली. येसूबाईचे रिकिबीतले
पाय पकडून ते लहान मुलासारखे ओक्साबोक्शी रडू लागले. उलट त्यालाच धीर देत
येसूबाई बोलल्या; "पुसा तो आसवं. उलट फिरा माघारा. इकडच्या स्वारींना सोडवून

Figure: 2 Sample of a degraded document image

- **Post-Processing Procedure:**

Require: The Input Document Image I , Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edg
Ensure**:** The Final Binary Result B*f*

1: Find out all the connect components of the stroke edge pixels in Edg.
2: Remove those pixels that do not connect with other pixels.
3: for Each remaining edge pixels (x,y): do
4: Get its neighbourhood pairs: (x − 1, y) and (x + 1, y);
(x, y − 1) and (x, y + 1)
5: if the pixels in the same pairs belong to the same class (both text or background) then
6: Assign the pixel with lower intensity to foreground
class (text), and the other to background class.
7: end if
8: end for
9: Remove single-pixel artefacts' [4] along the text stroke boundaries after the document thresholding.
10: Store the new binary result to B*f* .

[2.2] AN INVARIANT APPROACH: ALGORITHM

In this paper it converts the definitions of image moments into a discrete domain. In this there are two classes of such features: the first one in the spatialdomain and the second one in the frequency domain.

Algorithm Invar_Match

1) Inputs:
g – blurred image of the size *N* ´ *N*,
f – template of the size *L* ´ *L*, *N* @ *L*,
r – maximal order of the invariants used.

2) Calculate **C**(*r*)(*f*).
3) for*i*= 1 to *N* - *L* + 1
for*j* = 1 to *N* - *L* + 1
t = *g*(*i*: *i*+ *L* - 1, *j* : *j* + *L* - 1);
Calculate **C**(*r*)(*t*);
D*ij*= *dr*(*f*, *t*);
End;
end;

4) Find (*i*0, *j*0) such that
D (*i*0 *j*0)= min .
5) Output:
(*i*0, *j*0)–position of the template in the scene (upper-leftcorner).

[2.3] BINARIZATION USING NEURAL NETWORK

This technique uses Kohonen Self-Organizing Map neuralnetwork.The main stages of the proposed PEA, for two parameters (*P*1, *P*2) , are as follows:

Stage 1 Define the initial range of the PS values. Consider as 1 1 [*s* ,*e*] the range for the firstparameter and 2 2 [*s* ,*e*] the range for the second one.
Stage 2 Define the number of steps that will be used in each iteration. For the two parameterscase, let 1 *St* and 2 *St* be the numbers of steps for the ranges 1 1 [*s* ,*e*] and 2 2 [*s* ,*e*],respectively. In most cases 1 2 *St* = *St* = 3.
Stage 3 Calculate the lengths 1 *L* and 2 *L* of each step, according to the following relations:

$$L_1 = \frac{e_1 - s_1}{St_1 - 1}, \qquad L_2 = \frac{e_2 - s_2}{St_2 - 1}$$

Stage 4 In each step, the values of parameters 1 2 *P*, *P* are updated according to the relations:
P(*i*) = *s* + *i*·*L* , *i*= 0,..., *St* −1
P (*i*) = *s* + *i*·*L* ,*i*= 0,..., *St* −1
Stage 5 Apply the binarization technique to the processing document image using all thepossible combinations of 1 2 (*P*, *P*) . Thus, *N* binary images , 1,..., *j D j* = *N* are produced,where *N* is equal to 1 2 *N* = *St* ·*St* .
Stage 6 Examine the *N* binary document results, using the algorithm described in Section 3 andthe chi-square histogram. The best value of a parameter, for example parameter 1 *P* ,

isequal to the value (from the set of all possible values) that give the maximum sum oflevel values corresponding to this specific parameter value. For example, in the chisquarehistogram of Figure 4, we have nine levels produced by the combination of twoparameters having each one three values. As it can be observed, each parameter valueappears in three levels. In order to determine the best value of a parameter, we calculatethe sums of all level values that correspond to each parameter value and the maximumsum indicates the best value of the specific parameter. In our example, the levels (1, 2,and 3) are summed and they compared with the sum of the levels (4, 5, and 6) and levels (7, 8,and 9). The maximum value defines that the best value of the *W* parameter is equal to five

Stage 7 Redefine the lengths of the steps for the two parameters that will be usedduring the next iteration of the method:

$$L_1^{'} = \frac{L_1}{2} \text{ and } L_2^{'} = \frac{L_2}{2}$$

Stage 8 Redefine the ranges for the two that will be used duringthe next iteration of the method, according to the relations:

$$s_1^{'} = P_{1B} - (StNo_{1B} - 1) \cdot L_1^{'} \text{ and } e_1^{'} = P_{1B} + (St_1 - StNo_{1B}) \cdot L_1^{'}$$

$$s_2^{'} = P_{2B} - (StNo_{2B} - 1) \cdot L_2^{'} \text{ and } e_2^{'} = P_{2B} + (St_2 - StNo_{2B}) \cdot L_2^{'}$$

Stage 9 Redefine for the ranges that will be used in the next iteration accordingto the relations:

$$St_1^{'} = \begin{cases} \text{if } \; St_1 > e_1^{'} - s_1^{'} + 1 \text{ then } St_1^{'} = e_1^{'} - s_1^{'} + 1 \\ \text{else } \; St_1^{'} = St_1 \end{cases}$$

Stage 10 If *St* > 3 go to Stage 4 and repeat all the stages. The iterations are terminated whenthe calculated numbers of the new steps have a product less or equal to 3. The best PS values are those estimated during the Stage 6 of the last iteration.

[3] PROPOSED SYSTEM

As we saw above, the proposed techniques have some limitations. To overcome these limitations our system uses new binarization technique along with canny's edge map. There four modules in our system.

[3.1] MODULE OF CONTRAST IMAGE

To detect the exact text stroke it is very necessary to adjust the level of contrast in the image. In this module we are keeping the image contrast at min or max level. It is depend upon how much the foreground text is mixed with background paper noise. Here we invert the current contrast level i.e. we are reversing the color of the image.

[3.2] MODULE TO FIND THE EDGES

The contrasted image is further match with canny's edge detection graph. Which will produce the outline of the pixel around the foreground text. These pixels then divided into two groups. First one is connected pixels and non-connected pixels. Connected pixels occupies the area around text stroke. And non-connectedpixels shows the other noisy area present in the image.

[3.3] MODULE TO CONVERT INTO BINARY

The edge detected image is then converted into binary format of 0's and 1's. 0 indicates that the image pixels are non-connected pixels and 1 indicate that image pixels are connected pixels and the represents the text strokes. The 0's are removed from the image because they are part of background image.

[3.4] POST PROCESSING MODULE

Output of the binarization method creates separation in the image. So post processing eliminates the non-strokes image from binary image. And it returns a clear image which consist of only text strokes.

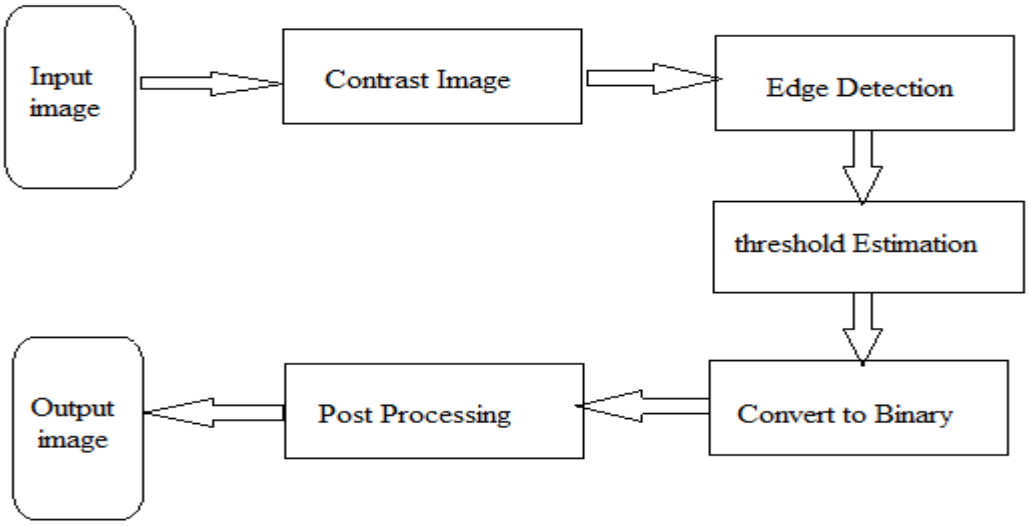

Figure: 2 System Architecture.

[4] CONCLUSION

The proposed method is simple binarization method, which produces more clear output. It can be work on many degraded images. It users contrast enhancement along with threshold estimation. Here we have used canny's edge map to create outlined map around the text. The output of this system produces separated foreground text from collided background degradation. For that we have maintain the contrast level at min and max level. Which will help to make more clear and readable output.

REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 documentimage binarization contest (DIBCO 2009)," in *Proc. Int. Conf. DocumentAnal. Recognit.*, Jul. 2009, pp. 1375–1382.

[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 documentimage binarization contest (DIBCO 2011)," in *Proc. Int. Conf. DocumentAnal. Recognit.*, Sep. 2011, pp. 1506–1510.

[3] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwrittendocument image binarization competition," in *Proc. Int. Conf.Frontiers Handwrit. Recognit.*, Nov. 2010, pp. 727–732.

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

[5] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwrittendocument images using local maximum and minimum filter," in *Proc.Int. Workshop Document Anal. Syst.*, Jun. 2010, pp. 159–166.

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

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

[8] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarizationmethods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12,pp. 1191–1201, Dec. 1995.

[9] O. D. Trier and T. Taxt, "Evaluation of binarization methods fordocument images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17,no. 3, pp. 312–315, Mar. 1995.

[10] A. Brink, "Thresholding of digital images using two-dimensionalentropies," *Pattern Recognit.*, vol. 25, no. 8, pp. 803–808, 1992.

[11] J. Kittler and J. Illingworth, "On threshold selection using clusteringcriteria," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 5,pp. 652–655, Sep.–Oct. 1985.

[12] N. Otsu, "A threshold selection method from gray level histogram," *IEEETrans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62–66, Jan. 1979.

[13] N. Papamarkos and B. Gatos, "A new approach for multithresholdselection," *Comput. Vis. Graph. Image Process.*, vol. 56, no. 5,pp. 357–370, 1994.

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

[15] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarizationof document images," in *Proc. Int. Conf. Document Anal. Recognit.*,Sep. 1991, pp. 435–443.

[16] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarizationbased on topographic analysis using a water flow model," *PatternRecognit.*, vol. 35, no. 1, pp. 265–277, 2002.

[17] J. Parker, C. Jennings, and A. Salkauskas, "Thresholding using anillumination model," in *Proc. Int. Conf. Doc. Anal. Recognit.*, Oct. 1993,pp. 270–273.

[18] J. Sauvola and M. Pietikainen, "Adaptive document image binarization,"*Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.