Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 9, September - 2015

Enhanced Binarization Technique And Recognising Characters From Historical Degraded Documents

Bency Jacob
Department of Computer Engineering
Sinhgad Institute of Technology
Lonavla,India
bencyjac@gmail.com

Mr. S.B. Waykar
Department of Computer Engineering
Sinhgad Institute of Technology
Lonavla,India
sbwaykar@gmail

***Abstract*— Degradations starts appearing in historical document with time. It is very difficult to understand and retrieve contents from very badly degraded documents as there is variation between the document foreground and background. Thresholding of such document images either result in distorted characters and also detecting texts falsely. Numerous algorithms exist that can separate foreground text and background efficiently in the textual regions of the documents but portions of background are mistaken as text in areas that do not contain any text. This paper presents a way to overcome these problems by an improved binarization technique in which the entire image is divided into user defined grids and on each gird using variance method Otsu algorithm is applied that recovers the text from a severely degraded document images and thereby increases the accuracy of character recognition. The proposed document recovery algorithm efficiently removes degradations from document images. The results are evaluated based on PSNR, F-measure and Precision and Recall. The document that is retrieved is the thinned and each character is segmented from the text and the characters are trained and recognized by the proposed training method by score calculation.**

Keywords— binarization, denoising, global thresholding, local thresholding, thresholding

I. INTRODUCTION

The old handwritten documents, manuscripts, printed books which included various historical books, old papers which were written by our ancestors. All these documents are very important for us today .But due to the less proper management these documents are no longer in readable form.
Now a days these historical documents are presented in digital form for various purposes[1].
Digital era has made paper documentation not so useful process for everything today is done with the help of computers.

The goal of this project is the secure data Academic libraries, institutions and historical museums pile-up or preserve documents in storage areas. The work in this paper contributes to documents safe and efficient preservation in its original state throughout the years and their unconditional exploitation to researchers, a major issue for historical documents collections that are not properly preserved and are prone to degradation processes, see Documents digitalization, allows access to wider public, while various cultural institutions and heritage organizations create local or national digital libraries which are accessed through the internet. Our proposed work concentrates on basic techniques used for enhancing image and restoration, denoising and binarization. The document is recovered and characters are recognized.

To analyze the document, the degraded document image is binarized before processing it. It is nothing but segmenting the document background & the foreground visible text. For the confirmation of document image processing task a more accurate document image binarization technique is a must. After years of research in document image binarization, even today thresholding of degraded document images is still found to be a very challenging task because of the high inter/intra variation between the document background and the text stroke and across various degraded document images. The stroke width, stroke brightness, stroke connection, and document background change in the handwritten text within the degraded documents. Moreover, bleed through degradation is seen in historical documents by variety of imaging outputs. For most of the existing techniques many kinds of document degradations, it is still an unsolved problem of degraded document image binarization due to the document thresholding error. A document image technique for binarization presented in this paper is an extended version of an existing local maximum minimum method [5].

II. RELATED WORK

Many thresholding techniques are available to convert the image into its binary format. As many degraded documents, lot of variation in image pattern, sometimes global thresholding cannot be a better approach for the degraded document binarization and thus adaptive thresholding is better approach.

Generally Otsu's thresholding is used as global thresholding. But cannot be used for image with large number of variation, but if windowing is used and then for each window Otsu's is applied and thresholding can be enhanced and new array can be evaluated for further processing to get a clear segmentation of text from the background.

[3] Other approaches have also been reported, including background subtraction recursive method ,texture analysis, decomposition method, contour completion Markov Random Field, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance and combination of binarization techniques But these method are often complex for analyses.

OCR mainly deals with improving the efficiency and accuracy. All efforts in OCR technology concentrate on this property. The OCR technology for Roman scripts or Indian, follow the same basic methodology of pre-processing, segmentation, feature detection and extraction, and classification as referred by [8].

In this approach vertical and horizontal projections are used for line, word and character segmentation which obtain a performance result of 93%. The [9] proposes an OCR system including pre-processing by binarization and size normalization by trial and error methods. They perform segmentation using projection profile and report a recognition rate of 87%. Much of the exploration on OCR system's efficiency improvement, in Indian context, has revolved around the exploration of using Otsu's method for preprocessing. The Otsu's thresholding algorithm is the basic thresholding technique used popularly for binarization in most works.

The [10] discusses Otsu thresholding method gives a better binarization result in degraded documents
Global thresholding algorithms are usually faster as they use a single threshold based on the global histogram of the gray-value pixels of the image. This method is an improvement to existing methods to create a novel and effective way to binarize historical documents.

The [11] proposes a method for binarization of image using its texture . They use Otsu thresholding iteratively to produce a threshold values, and texture feature associated with each threshold are retrieved by run length algorithm. They report an improvement of 8.1% over the original Otsu's algorithm. A new method for the binarization of the image other than Otsu's is proposed by [12] in which binarization is done using Morphological operators. Morphological operators are very efficient in that it performs image binarization effectively by taking care of complement color combination patches in the image.

The Morphological operator serves better in case of noisy corpus, especially when border noise is present in abundant. Statistics show that thresholding method fails completely in case of fore mentioned cases. This is done by using two filters based on two operations dilation and erosion [12]. It gives a much better result than Otsu's binarization method by removing of the background noises much effectively.

III. PROPOSED METHOD

There are two important parts of system are :
A. Binarization
B. Character Recognition and Generation

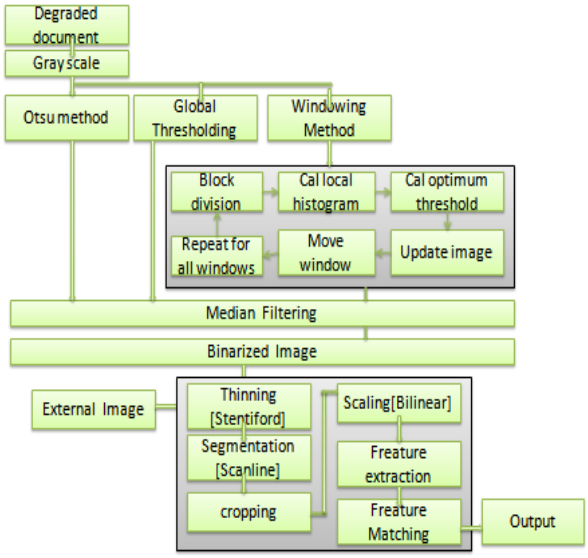

Fig1. System Flow

Here we are first converting the image to grayscale, then dividing the image into small windows and on each window we calculate the local histogram, then calculate the optimum threshold this is done on each window and then using the media filter the noise is cleared and the binarized image so obtained is thinned, segmented, cropped, scaled, then feature extracted and using the training and recognizing the characters in the text are recognized.

For a given a degraded document image, (R,G,B) colors are separated from a given colored image and then each color is ANDed with 0xff to obtained the 8 bit binary value of each color(R,G,B).

Then After separating each colour gray scaling which is 8 bit binary value is obtained. Thresholding is applied on the gray scale image. In this paper basically window based thresholding is applied and then Ostu's is applied over which window to obtain.

Threshold value for each window. This is done in the preprocessing stage. And finally filtering is done using median filter. Proposed method is simple and requires minimum parameter tuning. To enhance the quality performance of the technique Characters are recognized by training and recognition of characters.

Formula for Grey Scaling is (GS) = (R+G+B)/3.
Thresholding will be applied to gray scale image value i.e. only two values will be generated either Black Or white I.e. if gray scale value is greater than threshold then the pixel will turned to white & If gray scale value lesser than threshold then the pixel will turned black

1.Windowing technique

- After getting the gray scale image window function is applied to the grayscale image.
- For image blur Window can be of size 3 by 3, 5 by 5or 9 by 9 Less will be the window size less blur vice versa Windows width and height (means how many pixels in X and how many pixels in Y)
- Eg- (100*100)*(window size) 100 * 100 * (3*3) W * H * (size of the window)
- While traversing through each window Ostu's is applied to each window so that threshold is obtained for each window. This is done because the image has large variation.
- Otsu's alone cannot be used for the complete image as it gives the single global threshold value, so if the image has variation information will be lost.

2. Median Filter

- After creating foreground pixel map, some morphological post processing operations such as erosion, dilation and closing are performed to reduce the effects of noise and enhance the detected regions. Noise is also removed by using median filter.
- While going through the text in the window if neighboring pixel does not have any overlap edge of the text then that will be treated as noise and converted to white(i.e. 1) So after applying the filter we will get the text in readable form
- Then the image is then thinned using Stentiford algorithm. Then segmented to get each character and then cropped accordingly. Later the Scaling technique is used to scale the individual character and then the characters are trained and features are extracted and features are matched and the characters are recognized by Score calculation.

IV. MATHEMATICAL MODEL

A. Proposed Technique

Load the image
Set radius to divide into block default 10 or specify the radius.

```
ostusLocal(radius)
{
        {  scan vertically from y0 to h
        { scan horizontally from x0 to w
        }}
                Total =((radius*2)+1)^2;
                Sum=histogram of all pixels in local
frame.
                Wb=E of all sum histogram
                Wf= Total-wb;
                sumB=E(all(i)*histogram);
                mB=sumB/wF;
                Variance=wB*wF*(mB=mF)*(mB-mF);
                if(variance>maxvariation)
                then maxthreshold<-maxvariation
}
.
```

In image processing, Otsu's method is used to perform clustering-based image thresholding,[1] or, the reducing a graylevel image to a binary image. The algorithm assumes that the image contains two classes of pixels following bi-modal histogram (i.e foreground pixels and background pixels), it then calculates the optimum threshold separating the two classes so that their combined spread (intra-class variance) is minimal.[2]

B. STENTIFORD ALGORITHM

Stentiford algorithm has 3 steps

1. Template Matching

It checks whether any of the templates in fig 2. Match with the current template
Stentiford Thinning Algorithm[7]
It uses a set of four 3 x 3 templates to scan the image shows these four templates.

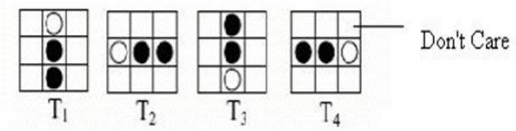

Fig.2 Stentiford Template

1.1. Find a pixel location (i, j) where the pixels in the image match those in template T1. All pixels along the top of the image are removed with this template from left to right and from top to bottom.

1.2. If the central pixel is not an endpoint, and has connectivity number = 1, then mark this pixel for deletion. Endpoint pixel: if A pixel is connected to just one other pixel it is considered an endpoint . That is, if a black pixel has only one black neighbour out of the eight possible neighbours.

1.3. Repeat steps 1 and 2 for all pixel locations matching T1.

1.4. Repeat steps 1-3 for the rest of the templates: T2, T3, and T4. T2 will match pixels on the left side of the object, moving from bottom to top and from
left to right. T3 will select pixels along the bottom of the image and move from right to left and from bottom to top. T4 locates pixels on the right side of the object, moving from top to bottom and right to left.

1.5. Set to white the pixels marked for deletion.

2. Check whether it is an endpoint

   This step checks whether there is any other pixel close by if there is no pixel then the current pixel is removed

3. Get Connectivity count

   Based on the following

$$\sum (k - (k * k + 1 * k + 2))$$

Where k=1,3,5,7.

The values of k are at the following positions in the 3X3 matrix below.

4	3	2
5	0	1
6	7	8

Fig3. Matrix for connectivity count

C. Median Filter [3]

Median filter is a non linear filter.
For
$A\{a1,a2,a3,....an\}$, and
$a1 \le a2 \le a3 \le ..... \le an \in R$ the new value of intensity of a pixel (i, j) of an image I is given by:

$$median(A) = \begin{cases} a_{\frac{n+1}{2}}, \text{if n is odd} \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{\frac{n}{2}+1}} \right), \text{if n is even} \end{cases} \quad \ldots.1$$

Character Recognition and Generation.

A 10X10 grid is formed for each character and is stored in the dataset .The characters generated from the degraded image is then scaled to a 10X10 grid with each pixel either black or white. Now the two grids are compared and a score is generated .Based on the grid which has the maximum value the index from the dataset is chosen. The character which best matches the character template that is already in the dataset.

Score is calculated for each block in the 10X10 grid if it matches that is if both the blocks have same value the score is added with 3 and is the values differ then 5 is subtracted . The score of all the characters are compared and the one which has the maximum score is then generated as the most relevant character. And is returned as recognized character.

V. Results And Discussion

To make the text in the degraded document understandable and to train and recognize the characters from the degraded document results are calculated by precision and recall calculations . The precision and recall values are as shown in the results below. Which shows results of how many characters from the degraded documents are properly binarized and how many characters are recognized.

		Precision	Recall
actual		0.958333333	0.92
wrong data set		0.95	0.826086957
Total		0.954166667	0.873043478
Accuracy Percentage		0.873043478	

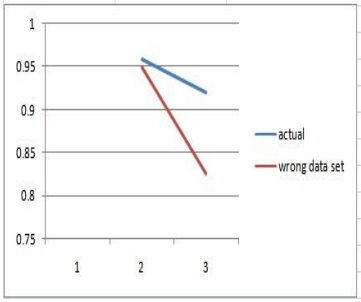

Fig 4. Binarization Result

		Precision	Recall
actual		0.980392157	0.961538462
wrong data set		0.958333333	0.92
Total		0.969362745	0.940769231
Accuracy Percentage		0.940769231	

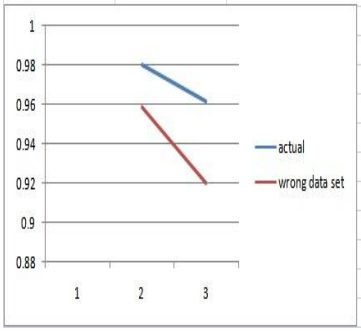

Fig 5. Character Recognition Result

TESTING OUTCOME

Fig 6.1.

Fig 6.2

Fig. 6.3

The outcome of the proposed method which is the binarization result Fig 6.1 showing the input image Fig 6.2 the output of the base paper, Fig 6.3 the outcome of the proposed method.

Fig. 7.1

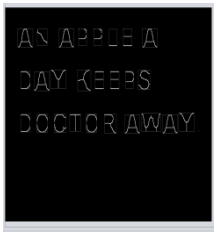

Fig. 7.2

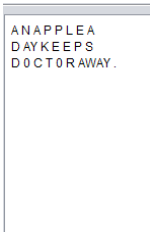

Fig. 7.3

Outcome of the character recognition method Fig.7.1 showing the input image, Fig. 7.2 the trained image and Fig 7.3 showing the recognized characters.

Based on the various techniques the PSNR value calculated are as follows.

Method	PSNR	F-Measure
Global Thresholding	9.81	77.2
OTSU	15.34	78.72
OTSU Thresholding [1]	19.65	93.5
Proposed Method	8.02	94.3

FIG 8 EVALUATION RESULTS

VII.CONCLUSION

In this paper, an improved approach is proposed for removing degradations and recovering text from printed and handwritten scanned document images by binarization and training and recognizing the characters. Most of the document analysis and recognition works reported are on better-quality documents. But still it remains a highly challenging task to implement a character recognition that works under all possible conditions and gives highly accurate results. Elaborate research on poor-quality documents are not much undertaken by the scientists in the development of script independent OCR. The Proposed technique gives 87.3% Accuracy for Binarization and for character Generation it gives 93.1 % Experiments should be made to observe the effect of poor quality paper as well as noise of various types, and take corrective measures.

ACKNOWLEDGMENT

I submit my gratitude and sincere thanks to my guide Prof.S.B.Waykar, Head of Computer Department Dr. S.D. Babar, who has been very concerned and has aided for all the material essential for the dissertation work and preparation of this thesis report, He helped me to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research oriented venture. I am thankful to our ME Co-ordinator Prof. M. S. Chaudhari, for his unwavering moral support and motivation during the entire course of the project. I would also like to thank our Principal Dr. M. S. Gaikwad who encouraged us and created a healthy environment for all of us to learn in best possible way. I would like to thank all the staff members of our college and technicians for their help.

REFERENCES

[1]. Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, IEEE 'Robust Document Image Binarization Technique for Degraded Document Images',APRIL 2013

[2] NTOGAS, NIKOLAOS and VENTZAS, DIMITRIOS, 'A BINARIZATION ALGORITHM FOR HISTORICAL MANUSCRIPTS,' in Proc. 28th Int. Conf. VLDB, Hong Kong, China, 2002,pp. 155ˆa16612th WSEAS International Conference on COMMUNICATIONS, Heraklion, Greece, July 23-25, 2008

[3] Brij Mohan Singh ˆa Mridula, 'Efficient binarization technique for severely degraded document images,' CSIT DOI 10.1007/s40012-014-0045-5

[4] Arie Shaus, Eli Turkel and Eli Piasetzky, 'Binarization of First Temple Period Inscriptions ˆa Performance of Existing Algorithms and a New Registration Based Scheme,' 2012 International Conference on Frontiers in Handwriting Recognition

[5] Maya R. Guptaˆa, Nathaniel P. Jacobson, Eric K. Garcia, 'OCR binarization and image pre-processing for searching historical documents,' Elsevier Received 28 October 2005; received in revised form 27 February 2006; accepted 28 April 2006

[6] Sayali Shukla, Ashwini Sonawane, Vrushali Topale, Pooja Tiwari, 'Improving Degraded Document Images Using Binarization Technique(ISSN : 2277-1581)Volume No.1', NTERNATIONAL JOURNAL OF SCIENTIFIC TECHNOLOGY RESEARCH VOLUME 3, ISSUE 5, May 2014.

[7] wikipedia,www,google.com'.

[8] Bansal V, Sinha RMK, "A complete OCR for printed Hindi text in Devanagari script", Sixth International Conference on Document Analysis and Recognition, 2011, pp. 800-804.

[9] Desai A, Malik L, Welekar R, "A New Methodology for Devanagari Character Recognition", International Journal of IT, Vol. 1, 2011, pp. 626-632.

[10] Liu Y, Srihari SN, "A recursive Otsu thresholding method for scanned document binarization", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 191,1997, pp. 540-544

[11] Brodic D, Milivojevic DR, Tasic V, "Preprocessing of Binary Document Images By Morphological Operators", In MIPRO 2011 Proceedings of the 34th International Convention ,2011, pp. 883-887.

[12] Shafait F, Breuel TM, "A simple and effective approach for border noise removal from document images", IEEE 13th International Multitopic Conference INMIC, 2009, pp. 1-5