Enhancement of Highly Degraded and Distorted Text Images using Image Processing

Ravinder Kaur Saini
M.Tech. – CSE, GNE, Ludhiana

ABSTRACT
This paper presents the algorithm to extract the text from the degraded and distorted images based on pattern recognition techniques used in image processing. Degraded images are quite often observed in old manuscripts, where it is strongly needed to enhance the quality of the manuscript for reproduction efficiently. The degradation of text may arise due to oxidation of ink used for printing/writing of the manuscript or extra ink spots or somewhere inkless printing. Degradation may also arise from the image grabbing techniques from a good quality image. This may arise due to poor illumination or low resolution of the camera being used. However, distortion in text may arise on account of incomplete character printing or deletion of some part of a character due to aging or any other reason. Therefore, to reproduce the text from the old manuscripts, it is required to apply some non-destructive techniques (NDT) to extract the useful text. Pattern recognition in image processing provides the solution for this task.

General Terms
Pattern Recognition, Non-Destructive Testing, Image Processing

Keywords
NDT, COG

INTRODUCTION

The objective of document image analysis is to recognize the text and graphics components in images and extract the intended information as a human would. Images of paper documents are almost inevitably degraded in the course of printing, photocopying and scanning and this loss of quality is responsible for an abrupt decline in accuracy by the current generation of text recognition systems[4]. A higher loss of quality can be observed for documents issued from a digital camera. This work tries to find some solutions to increase the recognition rate of degraded and distorted texts. The proposed system consists of the following steps:

1. Image Grabbing
2. Image Enhancement using thresholding algorithm
3. Segmentation
4. Character Recognition
5. Distorted Character Estimation
6. Result
7. Final Text presentation

1. IMAGE GRABBING
A CCD or digital camera is used to grab the image of the text to be extracted. Following image is presented here that is highly degraded text. Fig. 1 shows five characters and out of these five, only four are clearly identifiable. However, the first character is distorted. It can be judged to be the character 'E'. But this judgment capability is with human intelligence only. However, in the presented paper, efforts are put in to empower the computer system to have some sort of intelligence in this regard.

Fig. 1

2. IMAGE ENHANCEMENT AND THRESHOLDING
One well-known and very popular approach is Otsu's method [1]. An important problem with a global Otsu method is due to a non-uniform illumination which introduces most of the noise when using only the Otsu method [3]. This bad illumination appears as wide noisy areas, so we assumed that the illumination noise has a lower frequency spectrum than the character one. But, sometimes, global thresholds do not work so well because of a foreground with gray levels very close to the background or because of a non-uniform illumination.

For, we present here a local histogram based adaptive thresholding method for manuscripts and textual documents. Here, a window size of 7x7 is chooses. Say, N_{max} is the max. number pixels of gray value I_{max} and N_{min} is the min. no. of pixels of gray value I_{min}. in the current window.

Thresholding for the local current window is given by:

$$Ti = (N_{max} + N_{min}) / 2$$

Scan the whole image using the window of size 7x7 and apply the above thresholding value to get the binary image[2]. This gives a good quality binary image of text documents and old manuscripts.

Salt and Paper noise is removed by the following algorithm:

If (P_0 = BLACK) & $P_1 = P_2 = P_3 = P_4 = P_5\ P_6 = P_7 = P_8$ = WHITE

Then P_0 is the Background Pixel.

If (P_0 = WHITE) & $P_1 = P_2 = P_3 = P_4 = P_5\ P_6 = P_7 = P_8$ = BLACK

Then P_0 is the Object Pixel.

P_8	P_1	P_2
P_7	P_0	P_3
P_6	P_5	P_4

Fig. 2 Original Image

Fig.. 3 After Thresholding

Fig. 4 After removing upper and lower thick lines

3. SEGMENTATION

Once, the image is binarized. It is subjected to segmentation process. Here, the characters are segmented and grouped in one to one block with known access pointer. It is well known understanding that the distance between characters is smaller than the one between words. This criterion is used to club characters into words. In the presented paper, the characters are segmented based on their connectivity with their neighborhood. We have used the bwlabel function in matlab for segmenting of the characters.

4. CHARACTER RECOGNITION

To identify a character, we propose statistical parameters based algorithm. For, a character is confined in a boundary box proportional to its aspect ratio. Now, the image confined in a boundary box is our target, where all white pixels belong to background and all black pixels belong to the object of interest.

Following statistical features are computed from the analysis of the character with respect to centre of gravity for categorization:

[a] Normalised Maximum Radii in each Quadrant represented by R1, R2, R3, and R4. See fig. (5)

[b] Intercepts on each axis represented by X1, X2, Y1 and Y2 with respect to centre of gravity of object. See fig. (5).

[c] Mean Radius (R_M)

[d] Figure Aspect i.e. length to width ratio (FA)

FA = (X1 + X2) / (Y1 + Y2)

[e] Normalised Perimeter (Np)

Np = Total no. of pixels at the contour of object / R_M

[f] Normalised Standard deviation of radii taken from centre of gravity of object (NSD).

$SD = \sqrt{[(Ri - R_M)^2 / Np]}$

NSD = SD/ RM

Where RM, and Ri are the mean radius and i^{th} radius i.e. distance of ith pixel on contour of the pattern from its centre of gravity.

[g] Normalized Area (NA) of the pattern.

Normalised area = Total pixels on objects / R_M^2

[h] Euler No.: It is define as the no. of holes per object/character.

All features are normalized with respect to mean radius of the character [9]. It makes all the statistical features independent of size of the character. The set of described statistical features may be termed as figures of merit to classify a character.

The character confined in the boundary box is subjected to orthogonal transformation of Cartesian coordinates, so that if the character is oriented at some angle α from its original X-axis, then new coordinates can be found from the following orthogonal transformation[1]:

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \times \begin{pmatrix} X \\ Y \end{pmatrix}$$

Where (X', Y') are the new coordinates when X-axis is oriented at angle of α from the normal horizontal position and (X,Y) are the coordinates when X-axis is in its normal horizontal position.

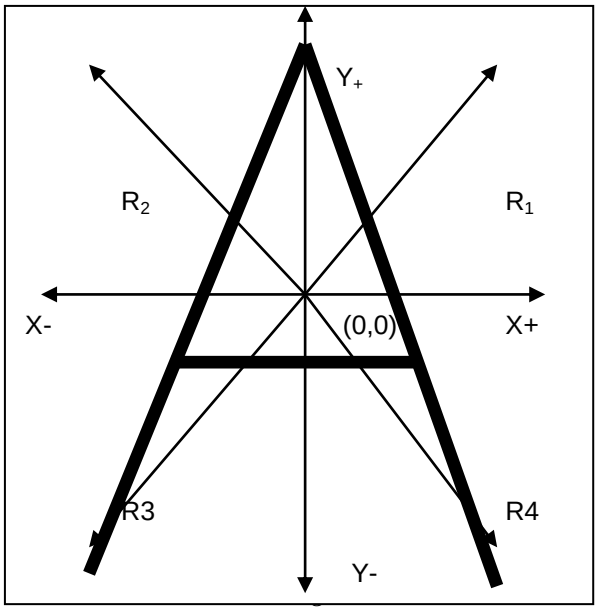

Fig. 5

5. DISTORTED CHARACTER RECOGNITION

The statistical parameters as discussed in section–4 possess very strong property of invariance. Size invariance has been taken care off by normalizing the statistical parameters with respect to mean radius [2]. Orientation invariance has been taken care of by orthogonal transformation of coordinates. Therefore, the character has been fully arrested by the all possible statistical attributes and does not pose any threat if the character is enlarged, bold or tilted at some angle. The orientation of the text characters is taken care of by the orthogonal coordinates system as discussed above.

One more parameter has been computed i.e. Euler number. Euler no. is defined as the no. of holes in an object/character.

Even if any of the character is distorted with respect to missing part, extended part or discontinuity in the character, then the statistical parameters will not vary much for a character under test from the normal set of statistical parameters computed when the character is OK. And that is the objective of this paper work.

6. RESULTS

Tables 1 and 2 show the results as obtained from the presented algorithm when implemented in matlab software. Parameters in table 2 for the character E* distorted in Fig. 3 are in tune with the standard character E parameters as in table 1. A look up table for all standard characters of English alphabets can be generated for identification of the text characters either normal or distorted to some extent.

7. FINAL TEXT PRESENTATION

After recognition of character sets present in the text image, the characters are presented in normal text language just like in Notepad document and presented in .TXT document form.

8. REFERENCES

[1] I.Guyon, A scaling law for the validation-set training-set size ratio, AT&T Bell Laboratories, 1997.

[2] A.Souza and al., Automatic filter selection using image quality assessment, 7th ICDAR Conference, 2003.

[3] Hybrid Image Thresholding Method using Edge Detection, Febriliyan Samopa† and Akira Asano, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[4] Digital Image Processing - Gonzalez,Woods, Edition 1997

[5] A Versatile Machine Vision System for Complex Industrial Parts (IEEE Trans. On Computers, Sep. 1977, No. 9, Vol. C-27).

[6] " Fuzzy Sets, Uncertainty and Information " A book by George J. Klir & Tina A. Folger.

[7] NI Vision Based Automatic Optical Inspection (AOI), Proceedings of 2009 IEEE International Conference on Applied Superconductivity and Electromagnetic Devices Chengdu, China, September 25-27, 2009.

[8] "Neural network And Fuzzy Systems – A Dynamical System Approach To Machine Vision Intelligence" Edition-1990, A book by Bart Kosko.

[9] "Fuzzy Logic And Control: Software And Hardware Applications" A book by Mohammad Jamshidi, Nader Vadice and Timothy j. Ross.

[10] "Edge Moments in Pattern Recognition" Thesis by Dr. H.K. Sardana.

Character	R1 Max	R2 Max	R3 Max	R4 Max	R1 Min	R2 Min	R3 Min	R4 Min	X1	X2	Y1	Y2	NP	NSD	FA	NA	Euler No.
V	1.78	1.78	1.46	1.46	0.42	0.33	0.25	0.33	0.75	0.67	0.07	1.45	11.27	0.78	1.06	2.15	0
A	1.39	1.39	1.62	1.59	0.67	0.68	0.58	0.51	0.70	0.73	1.38	0.45	8.53	0.57	1.27	2.01	1
T	1.44	1.44	1.99	1.99	0.22	0.24	0.26	0.23	0.15	0.16	0.97	1.99	10.09	0.41	9.70	1.86	0
E	1.66	1.43	1.44	1.70	0.22	0.24	0.16	0.15	0.81	0.52	0.17	1.36	12.36	0.39	1.16	2.09	0

Table 1:- Standard Characters Statistical Features Computed from the presented algorithm

Character	R1 Max	R2 Max	R3 Max	R4 Max	R1 Min	R2 Min	R3 Min	R4 Min	X1	X2	Y1	Y2	NP	NSD	FA	NA	Euler No.
V^*	1.70	1.70	1.46	1.46	0.40	0.37	0.30	0.32	0.75	0.67	0.07	1.45	11.27	0.78	1.06	2.15	0
A^*	1.45	1.56	1.52	1.59	0.62	0.60	0.68	0.45	0.70	0.73	1.38	0.45	8.53	0.57	1.27	2.01	1
T^*	1.49	1.40	1.90	1.99	0.12	0.20	0.29	0.29	0.15	0.16	0.97	1.99	10.09	0.41	9.70	1.86	0
E^*	1.60	1.36	1.34	1.60	0.20	0.27	0.26	0.19	0.71	0.62	0.19	1.30	8.36	0.34	1.15	1.99	0

Table 2:- Test Characters Statistical Features Computed from the presented algorithm

(V^*, A^*, T^*, E^* are the characters extracted segmented from the test image Fig. 3)