IMPORTANCE OF POST PROCESSING FOR IMPROVED BINARIZATION OF TEXT DOCUMENTS

Harjinder Singh Dhanjal[1], Ramandeep Kaur[2]

[1]*Technology Architect, Infosys, Chandigarh, Punjab, India*
[2]*Assistant Professor, Guru Teg Bhadhur Khalsa College of Engineering and Technology, Punjab, India*
harjinder_dhanjal@infosys.com, ramandeep.dhanjal@gmail.com

Abstract

Post processing uses mathematical morphologies to improve the quality of Binary image of a Document. It is observed that when a document is written using a ball point pen or when there is a problem in toner of printer then in scanned copy of such documents there are some minimal stroke gaps which are not easily visible to human eye. But a machine can easily recognise these gaps. These gaps lower the quality of binary image of a document. Post processing very important for improve the accuracy of binarization of degraded documents. In this paper a mathematical technique is discussed to improve the quality of binary image.

Keywords: *Post processing, Binarization, Degraded Documents.*

----------***----------

1. INTRODUCTION

Image binarization is the process of transforming a greyscale image to a black and white image. Representation of documents in binary form is essential for further processing like finding texts, lines, graphics, logos etc. It is the primary step of document image analysis and processing research. Pre-processing step is performed to improve the quality of image before binarization. It removes the noises like skew, ink blobs, Non-Uniform background, Stains and Human's annotations. There are number of filters available in [1-2]. These filters improve the quality of scanned image. Post processing is performed to improve the quality after binarization. It does so by removing ghost objects, filling stroke gaps and breaks. . Some writers while writing does not put equivalent pressure due to which ink could not spread that leaves some stroke gaps and breaks in the text, which are not visible to us, but easily recognized by the machine. Also, if there is a problem in toner of the printer there is discontinuity in printing as highlighted with red color rectangle in Fig.1. Post processing is the prominent area of research because the number of historical documents which are degraded due to aging and lack of preservation are now getting digitized to make easily available. But when these documents are binarized even after applying Pre-processing, these documents still contains some noises, stroke gaps, holes and breaks. These problems degrade the accuracy of Binarizaton. Post processing removes these problems from binary image which further needs to be processed by the Processes like OCR (Optical Character Recognition) as shown in Fig. 2.

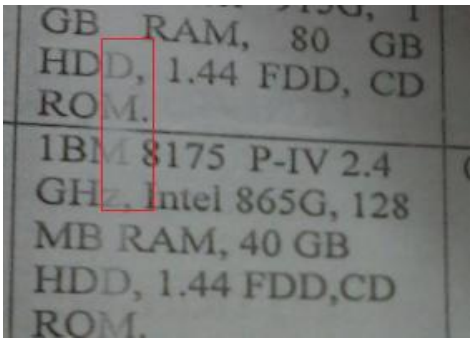

Fig1 Inadequate Printing

If the quality of binary image is poor then the accuracy of these processes gets lowered. So, Post processing is very useful for researchers. There is not much work available in literature. X. Ye et al [3] uses a post processing which improves the quality of binary image by removing ghost objects. The average gradient value at the edge of each printed object is calculated and objects having an average gradient below a threshold are labeled as misclassified and therefore removed. B. Gatos [4] also uses Shrink filter and Windows Swell filter. Y. S. Halabi et al [5] uses a shrink filter and Windows Swell filter to improve the quality of binary image. Y. Zhang and, Lenan WU† [6] uses 'Dilation' and 'Erosion' operators to preserve stroke connectivity and fill possible breaks, gaps, and holes. E. Balamurugan et al [7] implements local thresholding Nilblack method with Post processing to improve the quality of grey scale image.

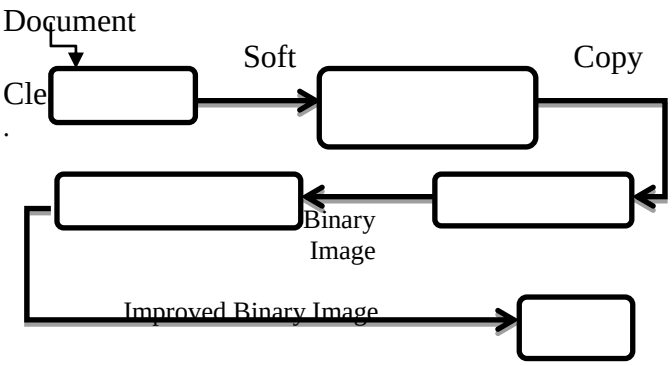

Fig.2 Sequence of various stages

2. METHODOLOGY

The accuracy of Binary image depends upon the type of algorithm used for it. A number of techniques are available for the Binarization. But most commonly used method is Threshold calculation (Global or Local) as defined by J. Sauvola and M. PietikaKinen [8]. In global approach single threshold value is applied for entire image while local thresholding apply different threshold values to different regions of images. A critical component in the binarization process is choosing a correct value for the threshold. If the threshold is set too low, then the resulting binary image will primarily be comprised of white pixels. Conversely, if the threshold is set too high, the resulting image will feature a large number of undesired black pixels. Thus, the threshold must be selected carefully to ensure the data information is preserved.

A. PERFORM BINARIZATION

Initially, computeThreshold() method is used to compute a threshold value. For every pixel it computes an intensity gradient by choosing a maximum of difference of left and right pixel intensity and upper and lower pixel intensity for calculating a threshold value. Then binarization is performed by using this threshold value in Binarize () method. In this method intensity of each pixel is compared with this threshold value. If pixel intensity value is greater than threshold value then pixel is set to 1 else set to 0.For this we have taken a text document shown by Fig. 3 as input and the results of binarization algorithm is shown by Fig.4.

computeThreshold()
[IN] Image as pixelArray[][] [OUT] int intensityThreshold

int totalWeightedIntensity = 0, int totalWeight = 0
1 For each pixel in pixelArray
1.1 Compute the Pixel Intensity gradient as weight = Maximum Of {(Intensity of Left Pixel – Intensity of Right Pixel), (Intensity of Upper Pixel – Intensity of lower pixel)}
1.2 totalWeightedIntensity = totalWeightedIntensity + weight * Intensity of Pixel.
1.3 totalWeight = totalWeight + weight;
2 intensityThreshold = totalWeightedIntensity/totalWeight;

binarize()
[IN] Image as pixelArray[][], [OUT] Binary Image as binaryPixelArray[][]

int intensityThreshold = 0;
1 call computeThreshold() for [IN] pixelArray [OUT] intensityThreshold
2 For each pixel row in pixelArray
2.1 If pixel Intensity is less than intensityThreshold
2.1.1 In binaryPixelArray, Set corresponding pixel as white (0)
2.2 Else
2.2.1 In binaryPixelArray, Set corresponding pixel as Black (1)

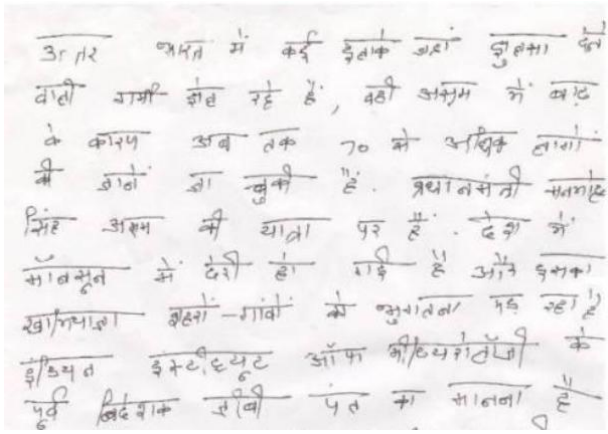

Fig.3 Original text Document

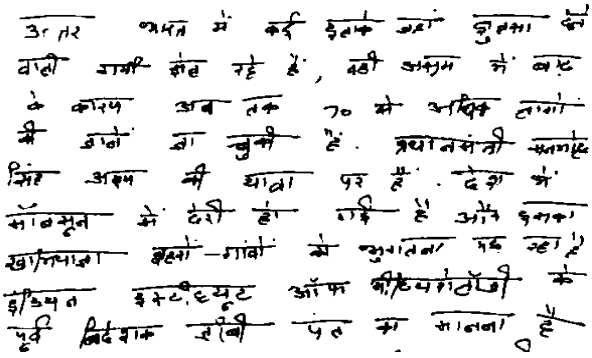

Fig.4 Binary image of text document

B. POST PROCESSING

In Post processing a rangeFilter() and some mathematical morphologies are used to preserve the stroke continuity and fill various gaps and breaks. Basically it is done by increasing the stroke width. In order to achieve this, the borders of all the characters are obtained and then imposed on characters. Fig.5. shows the effect of rangeFilter on the binay image, Fig.6 shows the boundaries of various connected components and Fig.7 shows final output having wider text stroke width.

rangeFilter()
[IN] Image as pixelArray[][], [OUT] Image as pixelArrayOut[][]

int maxValue = *0*, *int* minValue = *0*

1 For each pixel in pixelArray
1.1 Set maxValue = 0, minValue = 0
1.2 Find the maximum Value (RGB) as maxValue, in the 3x3 neighborhood pixels (around pixel.)
1.3 Find the minimum Value (RGB) as minValue, in the 3x3 neighborhood pixels (around pixel).
1.4 In pixelArrayOut, Set corresponding pixel value = maxValue – minValue

postProcessing()
[IN] Binary Image as binaryPixelArray[][]
, [OUT] Binary Image as processedPixelArray[][]

1 call rangeFilter() for for [IN] binaryPixelArray [OUT] processedPixelArray
2 perform logical Not operation for processedPixelArray.
3 perform logical AND of processedPixelArray and binaryPixelArray, and store result in processedPixelArray.

By applying these series of mathematical operations, scanned text document can be enhanced for better stroke width and almost zero stroke gaps, which actually helps in increasing the accuracy of character recognition application/software.

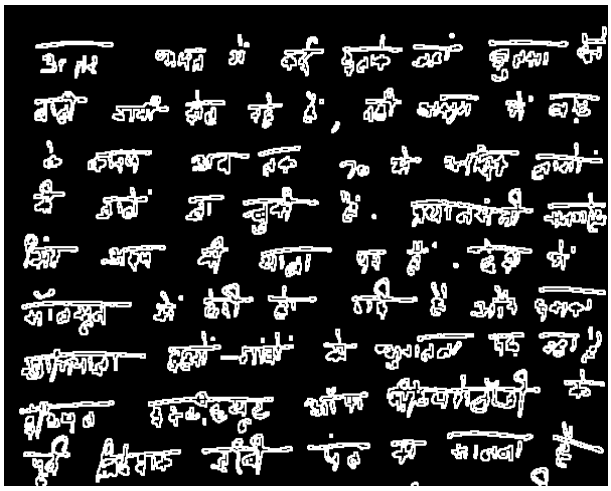

Fig.5 Effect of range filter on binary mage

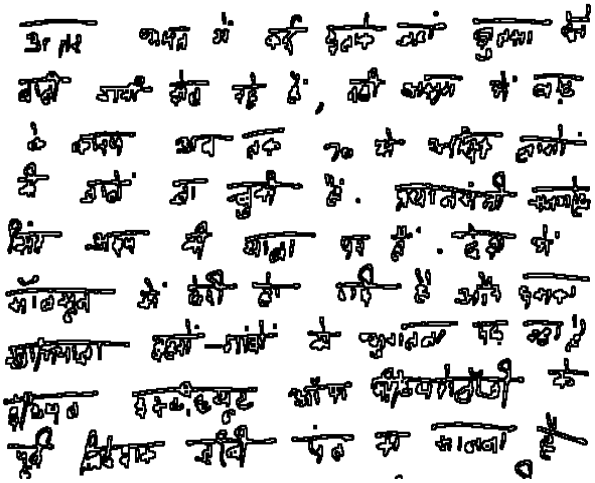

Fig.6 Boundary of text

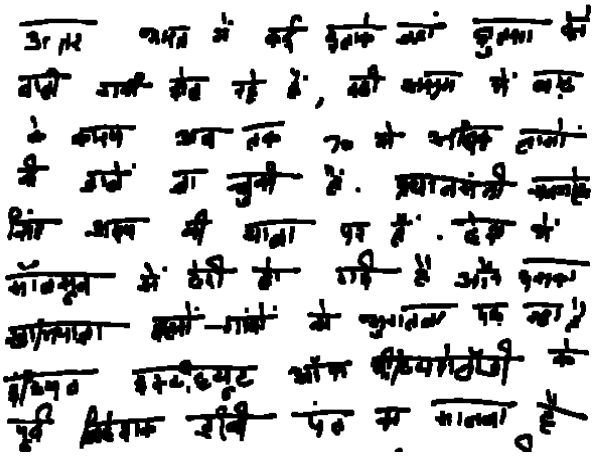

Fig.7 Final output with increased stroke width

CONCLUSIONS

In this paper a Binarization technique and a post processing step is implemented. The Post processing step increases the stroke width in order to fill stroke gaps and breaks to improve the quality of binary image. Further it is very useful to increase the accuracy of text recognition software.

REFERENCES

[1] Mamatha H.R, Sonali Madireddi and Srikanta Murthy K, " Performance analysis of various filters for De-noising of Handwritten Kannada documents ", International Journal of Computer Applications, Vol. 48, No. 12, pp. 29 -37, 2012.

[2] H. Mansara, R.Paladiya, S.Kakadiya, R.Chodvadiya, KrutiDangarwala " A Comparative Study On Various Techniques Of Noise Removal Process ", International Journal of Research in Computer and Communication technology, IJRCCT, Vol 1, Issue 5, PP.118-123, 2012.

[3] X. Ye, M.Cheriet,and C.Y.Suen,"Stroke-Model-Based Character Extraction from Gray-Level Document Images",IEEE Transactions on Image Processing, Vol. 10, No. 8, pp.1152-1161, 2001

[4]B. Gatos, I. Pratikakis, and S. J. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents, S. Marinai and A. Dengel (Eds.): DAS, pp. 102-113, 2004.

[5] Y.S. Halabi,Z.Sa.Sa, F.Hamdan and K.H.Yousef,"Modeling Adaptive Degraded Document Image Binarization and Optical Character System" European Journal of Scientific Research, Vol.28 No.1, pp.14-32,2009.

[6]Y. Zhang and L. Wu, " Fast Document Image Binarization Based on an Improved Adaptive Otsu's Method and Destination Word Accumulation", Journal of Computational Information Systems 7: 6, PP.1886-1892, 2011.

[7] E. Balamurugan,K. Sangeetha and P. Sengottuvelan" Document Image Binarization Using Post Processing Method ", Computer Engineering and Intelligent Systems, Vol.2, No.4. , pp. 14-17, 2011.

[8] J. Sauvola and M. PietikaKinen, "Adaptive document image Binarization", Pattern Recognition 33 (2000), pp. 225-236