International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 2, February 2014)

Visual Up gradation of Degraded Document Images

Swapna Halder[1], Sudipta Das[2]
[1,2]*Assistant Professor, CSE Department, Mallabhum Institute of Technology, INDIA*

Abstract- All over the World millions of historical documents are stored in Government record offices, museums, libraries and many other places. Now, these documents, originally, are frequently found in degraded conditions due to poor storage over large spans of time. Thus the documents are often difficult to human eyes to decipher. In order to emphasis preservation of those delicate documents as well as to provide easier access to scholars and researchers, those are now-a-days being scanned and preserved as document images. So, these documents are processed specially to remove damages and degradation and thus to make the document visually sound. This paper uses the image processing techniques which includes filtering, thresholding and image refinement in order to solve the problem of visual up gradation of degraded document images.

***Keywords*- Binarization, Filtering, Image, Threshold**.

I. INTRODUCTION

Visual up gradation of degraded document images is one of the key research areas in image processing. Lots of efforts have been provided in this field. Most of the existing visual up gradation techniques can be categorized into three distinct classes.

i. Binarization / Thresholding method
ii. Combination of Binarization / Thresholding method and other image processing techniques
iii. Non threshold based method

Binarization / Thresholding method involves converting a gray scale image into a binary form. It is used to separate out the region of the image corresponding to objects, from the regions of the image that corresponding to background. Normally black pixels correspond to background and white pixels correspond to foreground (or vice versa). In simple implementations, the segmentation is determined by a single parameter known as the threshold. In a single pass, each pixel in the image is compared with this threshold. If the pixel's intensity is higher than the threshold, the pixel is set to, say, white in the output.

If it is less than the threshold, it is set to black. Thresholding can be done in two ways: a) Global Threshold Method- In this method the pixels of the image are classified into background pixels and foreground pixels according to a global threshold value. b) Local Threshold Method- In this method the pixels of the image are classified into background pixels and foreground pixels according to a local threshold value by their neighboring pixels. Global thresholding methods are useful when the background is uniform. But in cases where different parts of a document have different backgrounds or foregrounds, local thresholding methods can be very useful.

Otsu's medthod [1] is old but very popular global threshold based algorithm. It is based on the variance of pixel intensity. The algorithm assumed that the image to be threshold contains two classes of pixels namely foreground and background, then calculate the optimum threshold, separating those two classes so that their intra class variance is minimal but this method does not work well with variable illumination. Cheriet [2] presented a recursive approach of Otsu's algorithm, for a specified class of document images of bank checks. In this approach, in each recursion the object with the lowest intensity from the input image is segmented first and it continues until there is one object left in the image. Kittler and Illingworth [3] suggested an algorithm that is applicable in multi threshold selection. It is assumed that the object and background class conditional probability density function are normal distribution. The histogram is used to classify the error for a mixture of two Gaussians. By locating the optimum threshold this algorithm minimize the probability of classification error. Don [4] proposed noise attribute threshold (NAT) method for document image binarization. In this method the noise attribute features are extracted from the image to make the selection of threshold value for image thresholding. This method works well when some objects in the image are too weak to show a prominent peak in the gray level histogram of the image. Solihin and Leedham [5] proposed QIR (Quadratic Integral Ratio) based method for image thresholding.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 2, February 2014)

In this approach the image is divided into three classes of pixels: foreground, background and a fuzzy class, where it is hard to determine whether a pixel actually belongs to the foreground or the background. A threshold value is chosen in the third class i.e. in the fuzzy region.

One of the most popular local threshold binarization method is proposed by Ni-Black [6]. Ni-Black calculated the threshold according to the local mean and the local standard deviation over a specific window size around each pixel location. But Ni-Black fails to adapt large variation in illumination, especially in the document image. Zang and Tan [7] proposed an improved version of Ni-Black method to overcome the distortion occurred during scanning thick bound documents. Souvola [8] also proposed another improved version of Ni-Black's method. Souvola computed the threshold value by using the dynamic range of image gray value standard deviation. Sauvola's algorithm is not so efficient when the gray values of the foreground and background pixels are close to each other. Bernses [9] compute the local minimum and maximum for a neighborhood around each pixel and uses the median of the two as the threshold for the pixels in consideration.

II. PROPOSED METHODOLOGY

Historical documents present information distortions that are visible in the form of poor quality, shadows, non uniform illumination, low contrast, large signal dependent noise, smear and strain spurious point noise and broken edges etc. Since document images vary in characteristics with respect to degradation types, it is needed to classify the documents image types before applying any visual up gradation methods on them. The document image degradation types have been grouped to eight distinct categories.

Image Degradation Type	Comments
Broken Character	Documents with thin strokes of pen, typing
Poor quality paper	Documents with oily page, ink seeking from other side of page
Quite non visible characters	Document with Poor contrast between foreground and background
Wrinkles to the paper	High humidity cause wrinkles effects.
Transparent Page	Document with Ink wet character visible at both side of the document
Uneven illumination	Documents effected due to image acquisition, causes shadowing effect.
Aging	The brightness of paper colors effected due to aging
Spot, Stains, Smears, Smudges	Dirty document with spots, stains, smears or smudges with less or more background noise.

The visual up gradation of degraded document images is performed by five stages: Image acquisition, image preprocessing, image filtering, image thresholding and finally image refinement. The document images are generally acquired by a digital camera with high resolution ratio, stored in computer and converted to compressed file formats with low storage requirements. JPEG files were used due to lower size and computation time. The document images can also be acquired by scanner. Image preprocessing can be involved by the process of cropping an image means, selecting the part of an image that one want to display and deleting the rest.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 2, February 2014)

But images captured often may be influenced by noise, however, the resulting images may not provide desired images for analysis. In addition, in images with acceptable quality, certain regions may need to be emphasized or highlighted. For that purpose Image filtering is needed. It can be broadly classified into two broad categories: Image filtering in spatial domain and Image filtering in frequency domain.

Thresholding are applied by global (General, Otsu's) and local (Local adaptive thresholding, Niblack,) thresholding techniques on previous stage resulting filtered images. A refinement procedure, based on erosion and dilation, is also applied on the binarized image, such that the obtained image has its characteristics further clarified in the texture and foreground compared with the background area.

Figure 1: Categories of different degraded document image

The methodology for visual up gradation of degraded document image is applied over several document images, which include pure textual, pure photographic, tabular, and compound images. All types of images are collected and converted into gray scale image in advance. Then the image is cropped to obtain the particular portion of the image for visual up gradation.

The cropping is not applied for all the images because sometimes whole the document image is needed for visual up gradation. Then filtering, Thresholding and refinement

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 2, Issue 2, February 2014)

Document Degradation Type	Global Threshold (Spatial Filter)	Global Threshold (Frequency Filter)	Otsu Threshold (Spatial Filter)	Otsu Threshold (Frequency Filter)	Local Threshold (Spatial Filter)	Local Threshold (Frequency Filter)	Niblack Threshold (Spatial Filter)	Niblack Threshold (Frequency Filter)
Broken Characters	Bad	Bad	Average	Average	Good	Best	Average	Worst
Poor Quality Paper	Average	Good	Good	Best	Bad	Bad	Bad	Worst
Poor Contrast	Bad	Average	Bad	Average	Good	Best	Fair	Fair
Wrinkle page	Best	Good	Good	Good	Average	Average	Bad	Worst
Ink wet characters visible both side	Good	Good	Good	Best	Average	Average	Bad	Worst
Uneven illumination	Bad	Bad	Bad	Worst	Best	Good	Average	Average
Aging Problems	Best	Good	Good	Good	Fair	Good	Bad	Worst
Dirty Documents	Good	Best	Good	Good	Bad	Bad	Worst	Worst

stages are performed step by step. The result of the applied methodology is listed below.

Figure 2: Summary of Results.

III. CONCLUSION

Historical documents contain important and interesting information. After years of storage, the documents are frequently degraded because of the different storage skills, environments, and the different quality of paper and ink etc. These documents are needed to be preserved digitally. In this paper a methodology is implemented which consist of five discrete stages: image acquisition, image preprocessing, image filtering, image thresholding and finally image refinement.

According to the image degradation types the proposed methodology is applied over several degraded document images. It is founded that, no method works well for all types of degraded document images.

Degraded document images with broken characters give best result by Local Threshold method with frequency domain filtering. Degraded document images with poor quality paper give best result by Otsu's Threshold method with frequency domain filtering. Degraded document images with poor contrast gives best result by Local Threshold method with frequency domain filtering. Degraded document images with Wrinkle Page gives best result by Global Threshold method with spatial domain filtering. Degraded document image with Ink wet characters visible at both side gives best result by Otsu's

Threshold method with frequency domain filtering. Degraded document image with Un even Illusion gives best result by Local Threshold method with spatial domain filtering.

Degraded document image with Aging Problems gives best result by Global Threshold method with spatial domain filtering. Degraded document image with dirty document gives best result by Global threshold Method with frequency domain filtering. So it is suggested that improved performance can only be obtained by selecting or combining appropriate method for the type of document images under the degradation type.

REFERENCES

[1] Otsu, N.: 'A threshold selection method from grey level histogram', IEEE Trans. Syst. Man Cybern., 1979, 9, (1), pp. 62–66

[2] Cheriet, M.: 'Extraction of handwritten data from noisy grey-level images using a multi-scale approach', Int. J. Pattern Recognit. Artif. Intell., 1999, 13, (5), pp. 665–685

[3] Kitter, J.; and Illingworth, J. 1986. Minimum error thresholding. Pattern Recognition 19 (1): 41-7, January – February.

[4] Don, H.S.: 'A noise attribute thresholding method for document image binarization'. Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), 1995, pp. 231–234

[5] Solihin, Y., and Leedham, C.G.: 'Integral ratio: a new class of global thresholding techniques for handwriting images', IEEE Trans. Pattern Anal. Mach. Intell., 1999, 21, (8), pp. 761–768

[6] Niblack, W.: 'An introduction to digital image processing' (Prentice Hall, Englewood Cliffs, NJ, USA, 1986), pp. 115–116

[7] Zhang, Z., and Tan, C.L.: 'Restoration of images scanned from thick bound documents'. Proc. Int. Conf. on Image Processing (ICIP), 2001, Vol. 1, pp. 1074–1077

[8] Sauvola, M. Pietikäinen, Adaptive document image binarization, Pattern Recognition, Volume 33, Issue 2, February 2000, Pages 225-236, ISSN 0031-3203

[9] Bernsen, J.: 'Dynamic thresholding of grey level images'. Proc. Int. Conf. on Pattern Recognition (ICPR), 1986, pp. 1251–1255