ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Salvaging Old Manuscripts and Images by Implementing Image Processing Methods

Saba P. Shaikh, Pranil J. Makhecha, Priyanka P. Sarda, Raveena R. Sarda

B. E Student, Dept. of Computer Engineering, Trinity College of Engineering & Research, Pune, India

ABSTRACT: Over the past years, many of our ancient historical documents, legal government documents are corroded by some means intentionally or unintentionally. So to recover these important documents, we are going to implement various techniques in order to save these documents digitally. Firstly, we need to understand the nature of the documents, on which it is written, and it's condition to apply various processing techniques .The image pre-processing and image enhancement is done by applying Image Processing and Hybrid Binarization methods. In this paper, we are going to mainly focus on text detection and OCR(Optical Character Recognition) technique. In some scenario, If the image is distorted or contain lot of background noise, it leads to unreliable OCR digitization as the accuracy is not provided. Our approach is to recover deformation of the entire image by image enhancement. We are maintaining the data set in order to obtain more accuracy. For irregular image illumination, we are going to use various adjustment methods such as brightness, contrast, saturation level etc for local adjustment of the image. Following Steps are carried out for image enhancement:
1)Local adjustment .2)converting image in grey scale.3)IGT is applied to OCR recognition.

KEYWORDS: Hybrid Binarization, OCR(Optical Character Recognition), Tesseract, Thresholding, Weiner Filter.

I. INTRODUCTION

There are documents of ancient times those have been passed on to us from our ancestors as well as there are many archaic documents in the libraries of the world digital or physical. These documents can be in the form of books or scrolls or papers that have been corroded owing to reasons like ink seepage, dust, fire, background noise, etc.In order to preserve these documents it is important that their integrity is kept absolute so that they are available for the world to see, navigate and fathom according to their preference or need, for this we need to enhance these documents by applying certain image processing techniques and filtering techniques. Background noise can be considered as a weed like abnormality which can be in textual or in image form. Also, to convert the grayscale image to binary form binarization technique is used which makes the formatting and the correction of image easier. OCR techniques can also be used on the image when in grayscale form for detection of stains and understanding of the pallid text vividly. On the contrary, when the background noise is to be left as it is to maintain the image in its grayscale so as to indicate its life or to portray it as if it has been marked by the traces of time binarization need not be done.

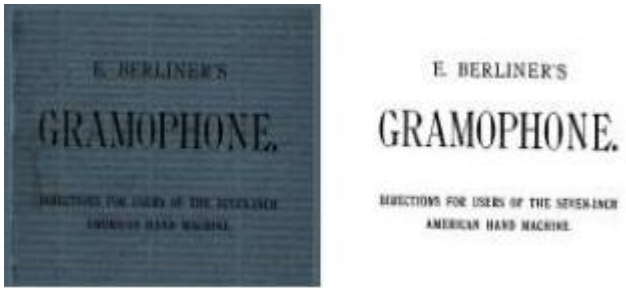


Figure1: (a) Input (b) Output (Texture Feature) [2]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

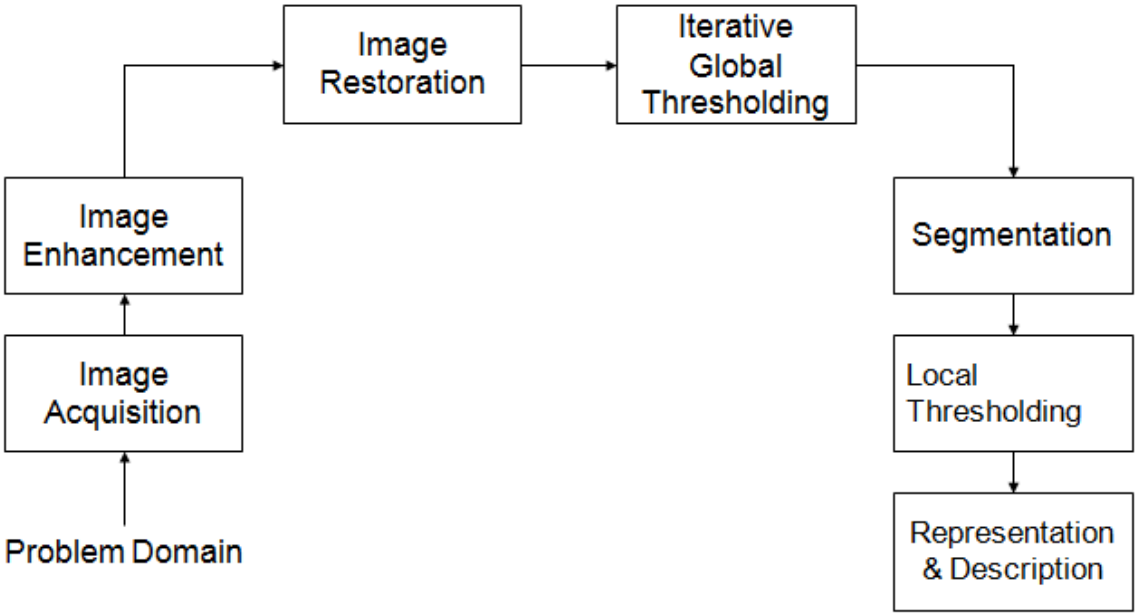
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

II. RELATED WORK

From past many years,our ancient Indian manuscripts and documents have been damaged intentionally or non-intentionally.These documents need to be preserved.Some of the manual techniques are carried out for conservation of ancient scripts,images.The other technique is used to store the image in digital format.In order to re store the information and images correctly, it is most essential to understand the nature of the calligraphy and the material on which it was written.Some of the processing techniques are used depending upon the condition of the manuscripts.In [8], we have managed to learn some of the techniques which need to be applied for digital restoration of the images depending upon various factor of the manuscripts which include Gaussian bandpass filter with thresholding. Paper[9] provides use the guidance to judge the methods which can be applied depending upon various factors of the manuscripts.This is basically used for enhancement of old documents images with damaged background and save our time and make it feasible to decide which method is to be used.We have identified three types of enhancement methods which are 1)using Binarization or thresholding method2)using Hybrid of binarization/thresholding and other procedures and 3)using non-thresholds methods.Finally,we have come at conclusion that second method is used most popularly and is in progress for future related work.

I. SYSTEM ARCHITECTURE



The flow of the system has been explained as below:

Problem Domain: The first step is identifying the document which has been compromised. This is the document that contains noise (ink seepage, dust, background dirt, etc.) or irregularities and which has to cleaned or de-noised.

Image Acquisition: In this step, the action of retrieving an image from some source is done. The image which is obtained after this process is the unprocessed image and this step is required for preserving the original quality of the image.

Image enhancement: This step encompasses two steps taken to improve the quality of the image- contrast stretching and de-blurring. Contrast stretching endeavours to better the contrast of the image by stretching the range of pixel intensity values up to a desired range. De-blurring attempts to de-noise the image by increasing the sharpness by removing anomalies such as defocus aberration, motion blur, etc.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Image Restoration: In this step, the image is made aesthetically pleasing to the observer by increasing the resolution, brightness or sharpness; these are not realistic changes. This basically compensates for the defects which degrade the image.

Hybrid Binarisation: This step incorporates the following three steps:

i. Iterative global thresholding: This is used to create binary images from greyscale images and separate the foreground and background. This method replaces each pixel in an image with a black or white pixel if its intensity is less or greater than the desired fixed constant. After the application of this step, if till some noise exists in the image then segmentation is applied.[1]

ii. Segmentation: This step includes the division of the image into smaller segments or parts for better analysis in order to detect noise more accurately.

iii. Local thresholding: This is similar to global thresholding, but is applied on the segmented parts of the image rather than the whole image unlike global thresholding.[1]

Representation and Description: Here the final result is produced in text format which is practically free of noise and in readable format.

III. IMPLEMENTATION

The following functionalities have been adopted by us in our proposed implementation:

Weiner filter: For cleaning of the image, we use Weiner filter which is a linear time invariant(LTI) filter. The Wiener filter minimizes the mean square error between the estimated random process and the desired process. The Wiener filtering executes an optimal trade-off between inverse filtering and noise smoothing. It removes the additive noise and inverts the blurring simultaneously. The Wiener filtering is a linear estimation of the original image. The approach is based on a stochastic framework. The orthogonality principle implies that the Wiener filter in Fourier domain can be expressed as follows:

$$W(f_1, f_2) = \frac{H^*(f_1, f_2) S_{xx}(f_1, f_2)}{|H(f_1, f_2)|^2 S_{xx}(f_1, f_2) + S_{\eta\eta}(f_1, f_2)},$$

where $S_{xx}(f_1, f_2)$, $S_{\eta\eta}(f_1, f_2)$ are respectively power spectra of the original image and the additive noise, and $H(f_1, f_2)$ is the blurring filter. It is easy to see that the Wiener filter has two separate part, an inverse filtering part and a noise smoothing part. It not only performs the deconvolution by inverse filtering (highpass filtering) but also removes the noise with a compression operation (lowpass filtering).[3]

OCR(Optical Character Recognition): OCR is the electronic conversion of the handwritten or manual files into machine encoded text. This technique scans the document from top left to bottom right corner and works in a sequential manner. The characters within a word are mapped in a two-dimensional format individually so that we can recognize each character by calculating the pixel value at different positions or points in a character. In this way, all the characters in a word are matched by using OCR. OCR works on scalar quantities which consists of magnitude only hence the image needs to be deskewed before applying OCR. It is a widely used technique for digitizing texts so that it can be electronically available in more compact form. We can even edit or search any text in the digitized files.[7]

Tesseract: Tesseract is certainly the best OCR library available so far and has been adapted in our proposed method. Tesseract accurately recognizes texts in more than 60 languages, supports multi-language texts and can be trained to work with previously unknown languages. A lexicon(database) of words is maintained and is used to compare and detect the words within the document. Basically, Tesseract is a powerful engine to carry out OCR functionalities. Thus, it incorporates all the methodologies of OCR whilst offering some additional features.[6]

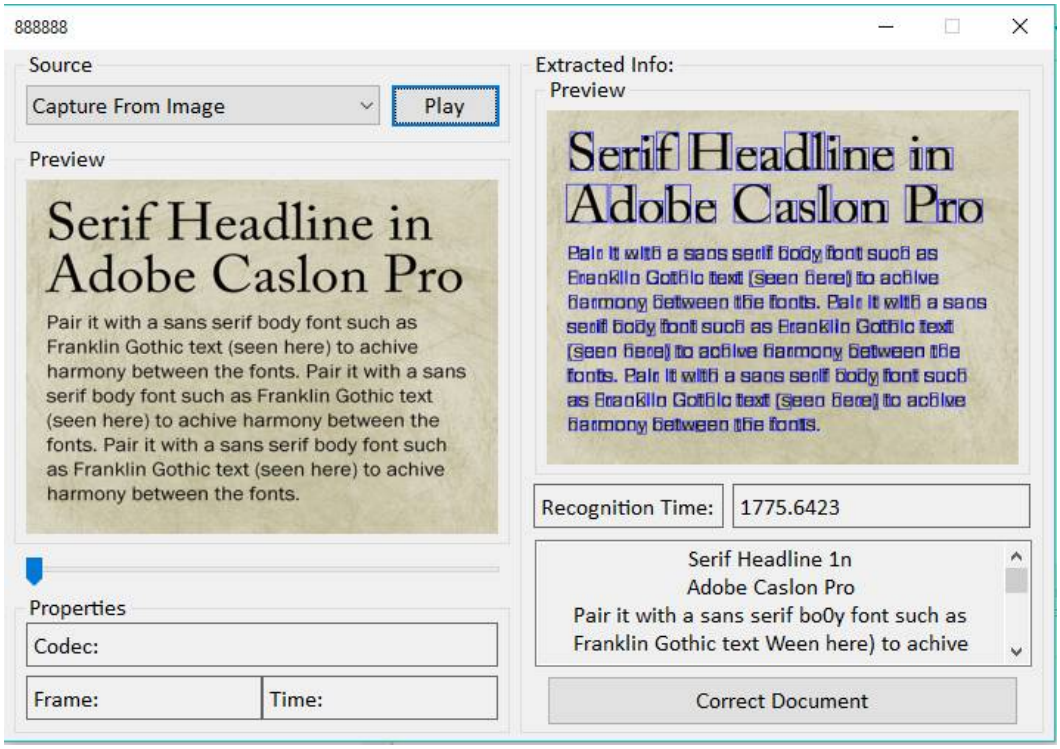
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

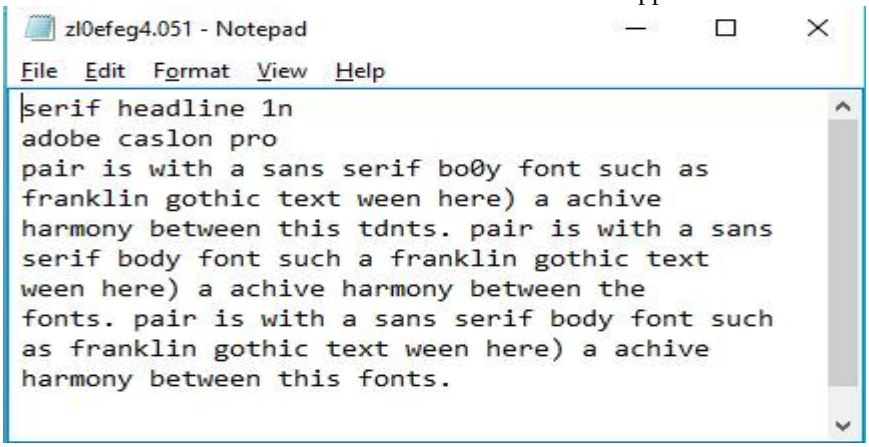
IV. Result And Discussion



(a)

(a) As depicted in the above image, the document is accepted as an image form from the source and then is processed by applying OCR and hybrid binarisation to denoise the image and obtain its result in text format as shown under 'Extracted Info' part. As shown above, each character is individually detected and analysed for comparison with words maintained in the lexicon. The dataset which is maintained for detection of each word under OCR reads and anatomizes each character individually by forming the frames around them as depicted above(blue boxes).

(b) The final output is a text file which holds the text within the document in approximate form.



(b)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

V. CONCLUSION

Thus, by using the aforementioned methods and functionalities the document is transformed from its image to text format, whilst attempting to denoise the document and convert it into readable form. Also as seen, the readability of the output depends upon the degree of degradation of the document. Good quality documents yield good quality text.

REFERENCES

1. Enhancement of old images and documents by Digital Image Processing TechniquesMrs.Preeti.Kale, Prof.G.M.Phade, Dr.S.T.Gandhe, Prof.Pravin.A.Dhulekar.
2. A Review of Degraded Document Image Binarization Techniques Jagroop Kaur, Dr.Rajiv Mahajan.
3. Historical Document Preservation using Image Processing Technique, Prof. D.N. Satange, Ms. Swati S. Bobde, Ms. Snehal D. Chikate.
4. Otsu, N. “A threshold selection method from gray-level histograms”. IEEE Trans. Systems Man Cybernet. pp. 6266, 9 (1), 1979.
5. Shi, Z., V. Govindaraju, "Historical Document Image Segmentation Using Background Light Intensity Normalization", SPIE Document Recognition andRetrieval XII, 16-20 January 2005, San Jose, California, USA.'
6. https://tesseract.patagames.com/
7. https://en.wikipedia.org/wiki/Optical_character_recognition
8. Sai Siddharth Kota, Raja Massand, Abhinaya Agrawal and Preety Singh "DIGITAL ENHANCEMENT OF INDIAN MANUSCRIPT, YASHODHAR CHARITRA" Computer Science and Engineering Department, The LNM Institute of Information Technology, Jaipur, India"
9. S. N. H. Sheikh Abdullah, K. Omar , M. S. Zakaria "Review on Image Enhancement Methods of Old Manuscript with Damaged Background" , 2010.