Document Image Analysis and Enhancement for OCR Applications

Paramita Maji
Assistant Professor
Department of Computer Science & Engineering
Dream Institute of Technology, Kolkata, India

Abstract: Image enhancement is one of the challenging issues in low level image processing. Camera captured and scanned document images can suffer from some form of degradation. Degraded, noisy scanned documents can appear due to improper paper feeding into the scanner. Camera captured images also may be degraded due to lighting intensity of the environment, blur, imperfect focus, shadow, uneven illumination, use of the flash light while capturing the image and distorted by a non-planar document shape. So, the document images are often hard to read, have low contrast, became faded, washed out, crumpled and are corrupted by various artifacts, as a result it substantially deteriorate the performance of document processing systems. An improved algorithm has been proposed for document image analysis and enhancement to enhance the quality and improve the readability of degraded document images. The main idea of this paper is to enhance the quality of document images and extracts the text components. Here, pre-processing performed on the document images, is composed of several steps and the steps are represented through a Graphical User Interface (GUI). At first dataset are collected, converted into grayscale, then splitted into two halves vertically and then both halves binarized individually. An adaptive binarization has been developed here. Splitted portions merged after filtering is applied on each half, then skew-corrected and segmented into separate characters and saved into a folder. Here, Skew has been corrected based on their entropy calculation of Horizontal Projection Profile (HPP). Experimental results show that this method performs well on bad or uneven illumination document images.

Index Terms - **Adaptive Thresholding, Character Segmentation and Horizontal Projection profile, Camera Captured Document Images, Scanned Document Images, Image Enhancement, Skew Detection & Skew Correction.**

I. INTRODUCTION

Document processing is too much essential in any organization whether operations are manual or computerized. The main purpose of document processing is to recognition of documents. Several good solutions and techniques [11] [14] [15] exist for document processing and analysis. OCR systems are used to store, search, and extract text from paper-based documents. Pattern Recognition is used actively for recognition of document images. Image enhancement is one of the challenging issues in low level image processing.

Pre-processing is essential prior to binarization, because direct binarization of such images lead to unsatisfactory result. H. K. Sawant et al. [8] proposed a technique for image enhancement, image pixel interdependency which is based on curvelet transform and linear perceptron network. It provides a better result for contrast enhancement and brightness are preserved. G Bal et al. [7] described a system which is capable to process document images with high levels of degradations and can be used to ground truthing of degraded document images. Here image foreground is separated from the background, then the foreground image is enhanced, after that the original image is enhanced, and then two enhanced images are blended using a linear blending scheme. The foreground-background separation component of this system is based on a probabilistic model estimated through expectation maximization (EM). In [19] Image processing techniques such as noise filtering, histogram equalization and power law transformation have been used to result in better representation of images. Here two performance assessment parameters; entropy and standard deviation, is used.

In [20] a hybrid binarization approach, combination of global and local thresholding techniques is proposed for improving the quality for the old documents. S. A. Tabatabaei et al. [16] described a local method for document image binarization technique

that separates text from background in badly illuminated document images. This technique is based on background estimation by using morphological closing operation. In [18] an improved adaptive method for document image binarization has been proposed. In first step wiener filter is used to reduce noises; in second step an improved adaptive Otsu's method is used for binarization; and in third step dilation and erosion operators are used to preserve stroke connectivity and fill possible breaks, gaps, and holes.

Skewed document is one of the major problem in document image analysis. Segmentation and recognition of characters become more difficult due to skewed characters. A. F. Mollahet al. [3] has presented a fast skew correction technique for extracted text regions from camera captured business card images. The skew angle is estimated by analysing the bottom and/or top profiles (height/depth from a horizontal base line) of a text region. Average deviation of the skew corrected text lines is within ±3 degree while the average processing time is between 17-110 milliseconds for 0.45-3.0 mega pixel images. In [4] a skew detection method for texts in scanned documents has been described. The algorithm works well within the skew angle in the interval $[-5^0 +5^0]$ of the input image. In [5, 12] and [10] a robust skew detection algorithm has been proposed based on Hough transform and entropy base model respectively.

In [9], [17] a method for document page decomposition by the bounding-box projection techniques have been described. A. F. Mollah et al. [1] [2] presented a complete Optical Character Recognition (OCR) system and an efficient Business Card Reader (BCR) for camera captured image/graphics embedded textual documents for handheld devices respectively. Here text regions are identified, extracted and segmented using horizontal histogram profile and horizontal and vertical histogram.

In this paper an improved algorithm has been proposed for the quality enhancement to improve the readability of the degraded document images. This research work only deals with textual portion of document images. Here, pre-processing performed on the document images, is composed of several steps and the steps are presented through a Graphical User Interface (GUI). Here an adaptive thresholding is implemented for binarization. Skew detection and Correction have been done by Horizontal Projection Profile (HPP). Bounding box method is used for character segmentation.

II. PRESENT WORK

Documents can be degraded by various factors. It can happen due to improper paper feeding into the scanner or incorrect snapshot. So, it is necessary to enhance the quality of the document image for subsequent classification and recognition. Here an algorithm has been proposed for the quality enhancement that is more efficient than some other algorithms.

III. PROPOSED ALGORITHM

The proposed approach is composed of several steps and the overall algorithm improves the performance. In this research work, mainly pre-processing operation is performed on the document images.

1. Gray-Scale conversion

A colour image consists of colour pixels represented by a combination of three basic colour components, Red, Green and Blue. The colour of each pixel is determined by the combination of these 3 colour component intensities, stored in each colour plane at the pixel's location. The range of values for all these colour components is 0-255.The reason for differentiating such images from any other sort of colour image is that less information needs to be provided for each pixel. So, the corresponding gray scale value for each pixel, which also lies between 0-255, may be obtained by using Eq. 1.

$$Gray = 0.299 \times Red + 0.587 \times Green + 0.114 \times Blue \quad (1)$$

Applying this transformation for all pixels, the gray scale image is obtained and is represented as a matrix of gray level intensities. This gray level image is fed as input for further analysis.

2. Split

The input image is splitted into two halves vertically. As the intensity of the image is different at different pixel's location, so, after applying Thresholding, each thresholded portion gives more appropriate results than the whole thresholded image. This step is important for the improvement of the quality of document images.

3. Binarization

The objective of binarization is to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value. Thresholding is a process of finding an appropriate threshold value for binarization. For each pixel in the image, a threshold has to be calculated. If the pixel value is below the threshold it is set to the background value, otherwise it assumes the foreground value. Thresholding can be categorized into Global thresholding and Local thresholding.

In the global thresholding technique, a unique threshold value is selected for the entire input image and each pixel is compared with the selected threshold value at the time of transformation. Global thresholding [14] is an efficient, less time consuming and region independent one. It is more appropriate in the uniformly illuminated images than the badly illuminated images. In images with uniform contrast distribution of background and foreground like document images, global Thresholding is more appropriate. It has drawback based on the shape of the intensity distribution. Degraded document images do not always contain well-distributed foreground and background intensities due to poor contrast and noise. As a results most of the pixels cannot be easily classified as foreground or background. Otsu's method [13] is popular method for many computer vision applications. First, calculations are made of the ratio of between-class variance to within-class variance for each potential threshold value. The classes are the foreground and background pixels, and the purpose is to find the threshold that maximizes the variance of intensities between the two classes and minimizes them within each class. But some experiments, shown in Figure 2 indicate that documents are not well segmented using this method and thus results are poorer than local thresholding.

In local Thresholding technique, a threshold value is calculated for each pixel, based on some local statistics such as range, variance, or surface-fitting parameters of the neighborhood pixels within a local block of size. In this technique, a threshold has to be calculated for each pixel in the image. If the pixel value is below the threshold it is set to the background value, otherwise it assumes the foreground value. In this research work, an efficient adaptive binarization technique has been developed. It examines the intensity values of the local neighborhood of each pixel and then each split portion is binarized. The *mean* of the *local* intensity distribution, *T=mean* is included. The size of the neighborhood has to be large enough to cover sufficient foreground and background pixels, otherwise it will results into a poor threshold value. On the other hand, choosing regions which are too large can violate the assumption of approximately uniform illumination. This method is better than global thresholding.

Algorithm for adaptive thresholding

Input: Left and right part of document image file, two constants (C1 and C2) and two different local window size for left half and right half.

Output: Binarized left and right portion of image.

Step 1: Convolve the left half and right half with a *mean* operator.

Step 2: Subtract the left portion from the left convolved image and right portion from the right convolved image.

Step 3: Threshold the left difference image with *C1* and right difference image with *C2*.

Step 4: Invert the thresholded left part and right part.

4. Noise Remove

After binarization, document images are usually filtered to reduce noise. A document to be scanned or captured by camera can itself be impure with spots etc. which constitute noise. Scanning or data capturing itself can introduce some amount of noise. Noise is also incorporated due to the degeneration, ageing, photocopying. In order to make it suitable for further processing, a scanned document image is to be freed from any existing noise. This can be achieved by a method known as image enhancement-this means improvement of the image being viewed by the machine or human. Smoothing operations in document images are used for blurring and for noise reduction. Blurring is used in preprocessing steps such as removal of small details from an image. In binary (black and white) document images, smoothing operations are used to reduce the noise. Smoothing and noise removal can be done by filtering. Filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. Various methods can be applied to reduce noise. The most important reason to reduce noise is to obtain easy way of recognition of documents. Another reason is that noise reduction reduces the size of the image file, and this in turn reduces the

time required for subsequent processing and storage. The objective in the design of a filter is to reduce noise or to remove as much of the noise as possible while retaining all of the essential pixels. Salt-and-pepper noise (also called impulse and speckle noise, or just dirt) is a prevalent artifact in poorer quality document images (such as poorly thresholded faxes or poorly photocopied pages). This appears as isolated pixels or pixel regions of ON noise in OFF backgrounds or OFF noise (holes) within ON regions. These salt-and-pepper noises are removed from each part using median filtering**.**

5. Merge

After noise removal the split parts are merged again. The joined binary image is taken as input in the next step.

6. Skew Detection and Skew Correction

A text line is a group of characters, symbols, and words that are adjacent relatively close to each other, and through which a straight line can be drawn (usually with horizontal or vertical orientation). The dominant orientation of the text lines in a document page determines the skew angle of that page. A document originally has zero skew, where horizontally or vertically printed text lines are parallel to the respective edges of the paper. However when a page is manually scanned or camera captured, non-zero skew may be introduced. Since a reader expects a page displayed on a computer screen to be upright in normal reading orientation, it is important to perform skew estimation and correction before displaying document pages, otherwise will have a severe impact on the future processing of the image. Text regions of document images have two types of pixels – black and white. The black pixels constitute the texts and the white pixels are background. In this research work skew has been corrected based on their entropy calculation of Horizontal Projection Profile (HPP) [10]. HPP of a document image is a vector where each vector element contains the sum of the pixel values in the corresponding row, given in equation no. 2.

$$\boldsymbol{HPP(i) = \sum_{j=1}^{Number\ of\ columns} I(i,j)}\textbf{, Where } \boldsymbol{I(i,j)} \textbf{ is the image.} \quad (2)$$

The input is a black and white skewed document image and the skewness of the images may vary between [-15^0 $+15^0$]. At first black text pixels are extracted. Then horizontal projection profile is estimated from extracted black pixels. The randomness associated with the projection profiles increases as the skew of the images varies. As entropy is a statistical measure of randomness, so, this property is used to detect the skew. It is given by equation no 3.

$$\boldsymbol{E(i) = \sum_i -HPP(i) * \log(HPP(i))} \quad (3)$$

Then entropy values are obtained from their corresponding HPP's. The angle for which the entropy is minimum, is taken as skew angle. After obtaining the skew angle, the image is rotated by that angle in the opposite direction using nearest neighbor interpolation. Thus the document image is skew corrected.

Algorithm for Skew Detection and Skew Correction

Input: A skewed binary document image file, maximum allowed angle for skew correction, angle resolution of rotation.

Output: skew angle in degrees, skew corrected image.

Step 1: First set the default values for maximum angle, angle resolution of rotation.

Step 2: Extract black text pixels for further analysis.

Step 3: Estimate Horizontal Projection Profile of output obtained from step 1 using equation 2.

Step 4: Estimate entropy of the objects from their corresponding HPP using equation 3.

Step 5: Find the skew angle for which entropy is minimum.

Step 6: Find the skew corrected image by rotating through the skew angle in opposite direction.

7. Character Segmentation

After skew-correction, document image is segmented into characters. The objective in character segmentation is to identify each individual character. Document segmentation is one of the critical phases in machine recognition of any language. Correct segmentation of individual symbols decides the accuracy of character recognition technique. At first skew corrected binary document image is taken as input for this step. Then all the connected components of the binary image are computed [9]. Then the connected components are marked by the respective rectangular bounding boxes which just enclose them completely. Bounding Boxes for connected components are the properties of the labeled connected component regions. A bounding box of a labeled region is a rectangle that just encloses the region completely and it specifies the boundaries of the corresponding connected component. Then characters have been segmented in a top to bottom order, again move over to the next column looking for the next TRUE pixel, and so on across all columns. Then all the characters have been saved in a folder.

Algorithm for character segmentation

Input: Skew corrected binary document image.

Output: The image with bounding box, a folder containing the characters in JPEG file format.

Step 1: Connected Component Labeling: Label the connected components and measure the properties of input image regions.

Step 2: Character Bounding Box Format: Plot the Bounding Box using the output of step 1.

Step 3: Object Extraction: Extract the objects entangled in the bounding box.

Step 4: Save the objects i.e., the characters obtained in a newly created folder.

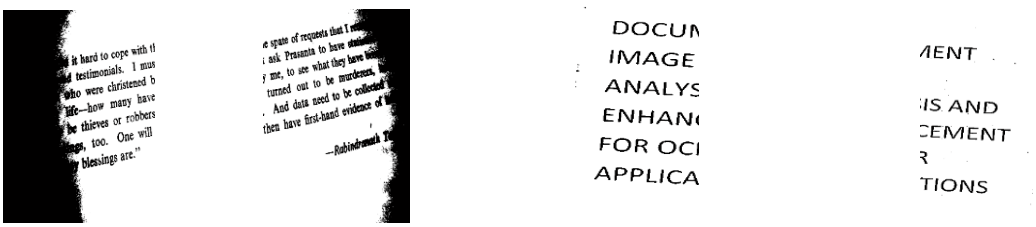
IV. Result Analysis

The proposed algorithm is tested on a database of camera captured and scanned document images with low contrasts, images with dark background and images with very sparse foreground ground content etc. Degraded color document images are taken as raw data (textual document images in English) and stored in a JPEG file format, an example is shown in figure 1.

(a) (b)

Figure 1. Input image (a) Camera captured degraded document image. (b) Scanned degraded document image.

Otsu's method [6] [13] is a popular method for global thresholding. But most of the degraded document images do not always contain well-distributed foreground and background intensities due to poor contrast and noise. As a results most of the pixels cannot be easily classified as foreground or background. Therefore, documents are not well segmented, example shown in Fig. 2.



(a) (b) (c) (d)

Figure 2. Binarized image using Otus's method (a), (b) Binarized left half and right half of Fig. 1(a).

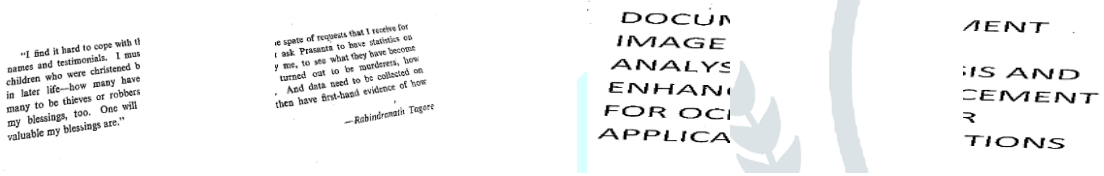
(c), (d) Binarized left half and right half of Fig. 1(b).

So, an efficient adaptive binarization technique is developed. It selects an individual threshold for each pixel based on the range of intensity values in its local neighborhood. Using this statistic, all pixels which exist in a uniform neighborhood are set to background. The main problem with this thresholding technique is the choice of window size. The window size have to be chosen properly, otherwise it will results as a poor thresholded image, shown in Fig. 3.

(a) (b) (c) (d)

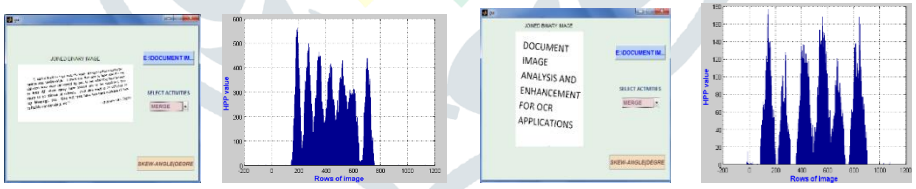
Figure 3.Binarized image using adaptive thresholding with improper window size: (a), (b) Binarized left half and right half of Fig. 1(a) with a neighborhood 7 ×7 & constant 0.1. (c), (d) Binarized left half and right half of Fig. 1(b) with a neighborhood 7 ×7 & constant 0.1

So, the chosen window size should be large enough to guarantee that the number of background pixels included is large enough to obtain a good estimate of average value but not as large as the average over non-uniform background intensities. However, the intensity of the pixel's location in the image varies, so a different window size has been taken. The result with different window sizes and different constant values are shown in Fig 4.



(a) (b) (c) (d)

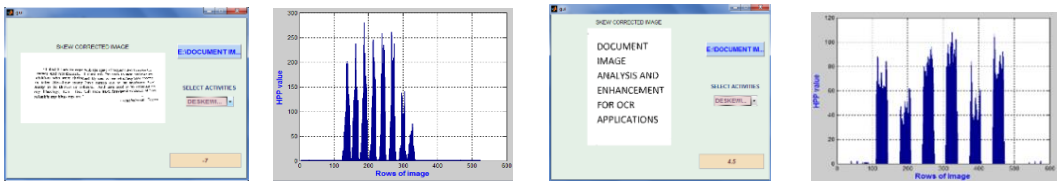
Figure 4.Binarized image using adaptive thresholding with proper window size: (a) Binarized left half of Fig. 1(a) with a neighborhood 14×14 and constant 0.04. (b) Binarized right half of Fig. 1(a) with a neighborhood 13×13 and constant 0.03. (c), (d) Binarized left half and right half of Fig. 1(b) with a neighborhood 13×13 and constant 0.03.



(a) (b) (c) (d)

Figure 5.Skewed image and their HPP. (a) Camera captured skewed image. (b) HPP of camera captured skewed image. (c) Scanned skewed image. (b) HPP of scanned image.

In this research work skew has been corrected based on their entropy calculation of Horizontal Projection Profile (HPP) [10]. At first black text pixels are extracted. Then horizontal projection profile is estimated from extracted black pixels using equation no. 2. The skewed image has a similar pattern of increase and decrease in the HPP shown in Fig. 5. The HPP values actually repeat in case of a skew corrected image shown in Fig. 6 . There exists a repetitive pattern as compared to the HPP of a skewed image. The randomness associated with the projection profiles increases as the skew of the images varies. As entropy is a statistical measure of randomness, so, this property has been used to detect the skew. Then entropy values are obtained from their corresponding HPP's using equation no. 3. The angle for which the entropy is minimum, is taken as skew angle. After obtaining the skew angle, the image is rotated by that angle in the opposite direction using nearest neighbor interpolation shown in Fig. 5. Thus the document image is skew corrected. The images can be rotated up to ±15 degrees. Fig. 6(a) and Fig. 6(c) have a skewed angle of -7 degrees and +4.50 degrees respectively.



(a) (b) (c) (d)

Figure 6.Skew Detection and Correction and HPP (a) Skew corrected camera captured image with a skew angle -7 degrees. (b) HPP of (a). (c) Skew corrected scanned image with a skew angle +4.5 degrees. (d) HPP of (b).

Finally, the deskewed binary images are segmented into characters, shown in Fig. 7(a) and Fig. 27(b) and saved into a folder.

(a) (b)

Figure 7.Character Segmentation (a) Deskewed camera captured image enclosed with the bounding boxes.(b) Deskewed Scanned image enclosed with the bounding boxes.

V. Conclusion and Future Scope

A good enhancement technique is vital to any document digitization system. The entire outcome of this process is relying heavily on the correctness and accuracy of the binarization, scanning and preprocessing methods, with an important part played by the skew correction strategy. In this research work, the set of scanned document images are segmented appropriately with average time consumption 3 seconds for 4.1 MP images. Scanned document images are well segmented into characters above 12 font size. Average time consumption is 7 second for camera capture document images. For images of non-planar paper surface (books, page curl) causes additional distortion, which poses an even greater problem due to its nonlinearity. As a result, camera captured document images are not skew corrected and segmented properly. This algorithm cannot work well with cursive text. So, cursive texts are not properly segmented using this method. So, such texts have been ignored while calculating the computational complexity. However, the technique has certain limitations too.

The present technique mainly deals with poorly illuminated English text documents. However, there are limitations that reflect for cursive texts and font size below 12. Main future improvements are segmenting characters with a minimum computational complexity and font size below 12, also, De-warping of the curled paper surface.

VI. References

[1] A. F. Mollah, N. Majumder, S. Basu and M. Nasipuri “Design of an Optical Character Recognition System for Camera based Handheld Devices” IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

[2] A. F. Mollah, S. Basu, M. Nasipuri "Segmentation of Camera Captured Business Card Images for Mobile Devices" International Journal of Computer Science and Applications, 1(1): 33-37.

[3] A. F. Mollah, S. Basu, N. Das, R. Sarkar, M. Nasipuri, M. Kundu "A Fast Skew Correction Technique for Camera Captured Business Card Images" IEEE INDICON (2009) 629-632.

[4] Boiangiu Costin-Anton, Bogdan Raducanu, Andrei-Cristian Spataru, "High-Precision Orientation and Skew Detection for Texts in Scanned Documents" IEEE 978-1-4244-5007-7/09/2009.

[5] Bin Yu, Anil K. Jain "A Robust and Fast Skew Detection Algorithm for Generic Documents" Pattern Recognition, Vol. 29, No. 10, pp. 1599-1629, 1996.

[6] C´eline THILLOU "Degraded Character Recognition" Diplˆomed'EtudesApprofondies en. Sciences Appliqu´ees 2003-2004.

[7] G. Bal, G. Agam, O. Frieder, G. Frieder "Interactive Degraded Document Enhancement and Ground Truth Generation" SPIE 6815, Document Recognition and Retrieval XV, 68150Z (28, January 2008).

[8] H. K. Sawant, Mahentra Deore "A Comprehensive Review of Image Enhancement Techniques" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 2.

[9] Jaekyu Ha, Robert M. Haralick Ihsin, T. Phillips "Document Page Decomposition by the Bounding-Box Projection Technique" Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95) 1995 IEEE.

[10] K. R. Arvind, Jayant Kumar and A. G. Ramakrishnan, "Entropy based skew correction of document images" Proc. PReMI 2007, Kolkata, Dec 18-22, 2007, pp. 495-502.

[11] L. Gorman, R. Kasturi "Document Image Analysis Systems" IEEE Computer Society Executive Briefings, July 1992, Computer 25, pp.5-8.

[12] Mike Adams, "Algorithm for Text Document De-Skewing", EECS 490, December 2004.

[13] N. Otsu, "A Threshold Selection Method from Gray-level Histograms", IEEE Trans. Systems, Man, and Cybernetics, Vol. SMC-9, No. 1, Jan. 1979, pp. 62- 66.

[14] R. Kasturi, L. W. O'Gorman and V. Govindaraju "Document Image analysis: A primer" Sadhana Vol. 27, Part 1, February 2002, pp. 3–22.

[15] S. Akram, Dr.Mehraj-Ud-Din Dar, A. Quyoum "Document Image Processing - A Review" International Journal of Computer Applications (0975 – 8887, Volume 10– No.5, November 2010.

[16] S. A. Tabatabaei, M. Bohlool "A Novel Method for Binarization of badly Illuminated Document Images" Proceedings of 2010 IEEE 17th International Conference on Image Processing September 26-29, 2010, Hong Kong.

[17] Shashidhara B and BhaskaraRao. N. "Word Segmentation for Document Images by Successively Merging Adjacent Character Bounding Boxes by Iterative Dilation" IJCSET, February 2012 Vol. 2, Issue 2,873-876.

[18] Y. Zhang , 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 (2011) 1886-1892. [19] Neetu Mittal ; Arjun Sehgal ; Sunil Kumar "Khatri Enhancement of historical documents by image processing techniques" 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) 20-22 Sept. 2017.

[20] Preeti Kale ; G.M. Phade ; S.T. Gandhe ; Pravin A. Dhulekar "Enhancement of old images and documents by digital image processing techniques" 2015 International Conference on Communication, Information and Computing (ICCICT) 15-17 Jan. 2015.