International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 06, June-2015

A Novel Approach of Data Extraction from Indian Degraded Historical Documents using Gamma Variation and Histogram Balancing Method

Neelu Maheshwari
M.Tech. Scholar,
Rajasthan Technical University,
Institute of Technology and Management,
Bhilwara, Rajasthan, India

Anurag Maloo
Assistant Professor
Institute of Technology and Management,
Bhilwara, Rajasthan, India,

Pankaj Singh Parihar
Assistant Professor
Institute of Technology and Management,
Bhilwara, Rajasthan, India

***Abstract*— Palm leaves manuscripts copper plate compositions, wooden compositions were one of the earliest manifestations of composing media and their utilization as writing material in India has been recorded from as early as the fifth century B.C. until as of late as the late nineteenth century. Palm leaf manuscripts identifying with arts, science, structural engineering, arithmetic, cosmology, astrology, and medication going back a few many years are still accessible for reference today because of numerous continuous endeavors for safeguarding of old archives by libraries and colleges not in India yet all around the globe. Such sort of original copies normally last a couple of hundreds of years yet with time the material degrades and the writing becomes illegible to be valuable in any structure. Advanced Digital Image processing can help enhance the images of these manuscripts in order to empower recovery of the written content from these degraded documents. In this paper we proposed a filter and transform based technique for recovery of data written on such historical original manuscripts. The method uses a dynamically selected pivoting background color in a linear transform to enhance the legibility of the foreground text. At that point a blend of two other image processing algorithms gamma variation method and histogram balancing are applied to the transformed image. The algorithms can be mathematically combined into one or two transformations for computational efficiency. The method is tested on a set of different historical document images in different environmental condition and the result of the proposed method is compared with Bilateral filter with Binarization method show significant improvement in readability with show noteworthy change in clarity. This enhanced image is send to a trained OCR engine for extracting Sanskrit data contents written on manuscripts. The method can likewise be utilized to improve digital images of antiquated, historical, degraded wooden, paper documents.**

Keywords*— *Digital Image Processing; Degraded Digital Documents Image Enhancement; Gamma Variation; Histogram Balancing; Bilateral filter; OCR; Sanskrit Data Trained OCR.

I. INTRODUCTION

One of the oldest medium of writing and communicating in South Asia are palm leaf manuscripts. These are also the major sources for writing and painting in South East Asian countries including Thailand, Nepal, India, Barma, Indonesia etc. Hence it is required to develop an automated system to decipher these inscriptions. The system takes the camera grabbed or scanned images of the inscriptions as an input and processes it before the character recognition is taken up. The images so captured have major problems like the broken letters, erased letters, distortion due to fossils settled and so on. The presence of unwanted marks engraved by the sculptor leads to wrong diagnosis of inscriptions. Hence this requires a lot of preprocessing before the character recognition is taken up. The need for efficient image restoration methods have grown with the massive production of digital images of all kinds, often taken in poor conditions.

There is immense measure of printed data that is installed inside pictures. Case in point, more archives is digitalized daily through cam, scanner and other gear, numerous advanced pictures contain writings, and a lot of text based data is inserted in web pictures. It would be extremely helpful to turn the characters from picture configuration to literary arrangement by utilizing Optical Character Recognition (OCR) [17]. This changed over content data is vital for document mining, document picture recovery etc. Nonetheless, by and large, the document pictures can't be straightforwardly sustained to an OCR framework because of the accompanying reasons:

- The original document papers suffer from different kinds of degradation including smear, ink-bleeding through and intensity variation, especially for historical documents when they are written on palm leaves, copper foil and paper.
- The process of obtaining digital images from the real world is not perfect. There are many factors that may cause image distortion, such as incorrect focal length,

over/under exposure, camera shaking/object movement, low resolution, etc.

Document Image Enhancement is a method that enhances the nature of a document Image to improve human observation and encourage consequent mechanized Image processing. It is generally utilized as a part of the preprocessing phase of diverse document analysis tasks. Document image enhancement issue is basically a not well postured issue, on the grounds that various enhanced images can be created from the same input image. Additionally, the nature of enhancement techniques is primarily judged by human perception, which makes the quantitative measures hard to be connected. The main aim of this study is to propose a document image enhancement technique for better accessibility to the textual information embedded in the images [18]. The specific objectives of this research are to:

- Propose some digital image processing techniques for degraded document images that achieved good performance for degraded documents and can be used in different document analysis applications.
- Propose an OCR technique by which data written on degraded historical documents in Sanskrit or Devnagri Lipi can be read and be digitalized.

A. *Scope of study*

There are many different kinds of document enhancement techniques which handle differently distorted document images, such as document image dewarping [1] and document image super-resolution[2]. In this paper, we focus on three aspects of the document enhancement techniques: document image Binarization, image enhancement and Sanskrit data retrieving. These techniques can be widely used in preserving ancient knowledge delivered by vedic maharshi, rishhi and munis. These documents are written on palm leaf and paper and we know they degrade over the time.

B. *Challenges on Degraded Document Image*

However restoration of contents of despoiled document has been considered for many years, the improvement of despoiled document images is still a hazy problem. This can be clarified by the way that the demonstration of the document foreground/background is tremendously difficult because of different sorts of document corruption for example, uneven brightening, picture contrast variety, dying through, and stretch as represented beneath[3],[4]. However, the despoiled document image binarization is not fully explored and still needs further research.

TABLE I. Pros and Cons of various image enhancement methods.

Methods	Pros	Cons
Global Thresholding	Fast, Produce good results on clean documents	Fail on degraded images
Local Thresholding	Works on degraded documents	Sensitive to window size
Background Subtraction	Produce good results when foreground varies	Performance decreased when background non-uniform
Image Contrast	Produce good results when background varies	Performance decreased when foreground non-uniform
Domain Knowledge	Preserve text info using domain knowledge	Hard to extract proper domain knowledge
Energy Based	Simple but effective	Need to tune a few parameters

II. RESEARCH PROBLEM AND PROPOSED SOLUTION

A. *Problem Description*

This paper is also focused on maintaining of data from degraded historical documents written in ancient language i.e. Sanskrit. The main objective of the planed method is to improve the quality of the captured image of despoiled documents contain important information and extract the data for digital purpose [11]. This problem can be split into two parts: First enhancing the captured image of ancient document and second is propose an OCR technique which can read this enhanced image and covert the data into digital form written in Sanskrit language.

For first, Image processing techniques can help to improve the images of manuscripts as to facilitate retrieval of the written text from these despoiled documents. But these methods do not produce satisfactory results in processing these manuscripts since the color intensity of the background varies throughout the image [12],[13].

Secondly, extracting the data from enhanced image is a problem and this problem become critical when document is written in ancient language like Sanskrit. In market there are a lot of solutions are available for optical character recognition (OCR) purpose but only a few of them are available for extracting Sanskrit data along with poor quality.

B. *Proposed Solution*

In this paper a novel method is proposed for digitally enhancing and data extraction from timely degraded historical documents.

Fig. 1. Block diagram of proposed solution.

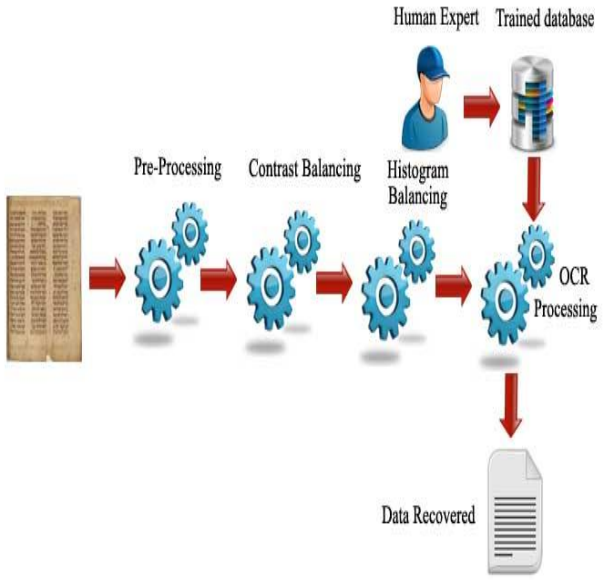


Fig. 2. Technical representation of proposed solution.

For document enhancement a lot of work have already been done like background foreground extraction, binarization etc. for still all of these solution require complex mathematical calculations which may slow down the processing method. In proposed method gamma variation method and histogram balancing are applied to the transformed image.

For Sanskrit data extraction, a very few methods are available but they lack of proper reading of Sanskrit documents as they work fine for Hindi documents. So here we also train the system for recognize the manuscripts data written in Sanskrit language. For this purpose we choose Tesseract OCR engine which is open source project [16], [17]. But this engine would not come with Sanskrit dataset, it recognize English, Hindi and only some language dataset. So we will train this OCR engine for recognize Sanskrit data.

C. *Proposed algorithm for historical document enhancement*

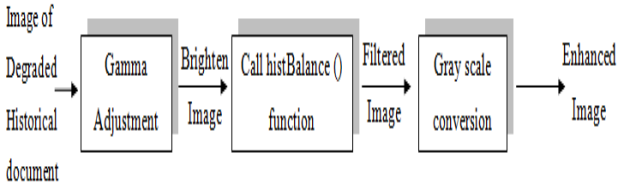


Fig. 3. Block diagram of image enhancement module (Pre- Processing).

Pre-Processing function

Require: The input Document image I

Step 1: increase brightness of I by adjusting gamma value and obtained brighten image B.

Step 2: call histBalance(B) function and get filtered image X.

Step 3: convert image B into gray scale and get the final processed image Y.

histBalance() function

Require: The brighten image of document B.

Step 1: Read value of alpha and beta value.

Step 2: For each pixel (x,y) of image B: do;

Step 3: For each color channel c of image B; do;

Step 4: X[x,y] = alpha * B[x,y] + beta;

Step 5: end for

Step 6: end for

Step 7: Return the filtered image X.

D. *Proposed Algorithm for Sanskrit data extraction from enhanced image*

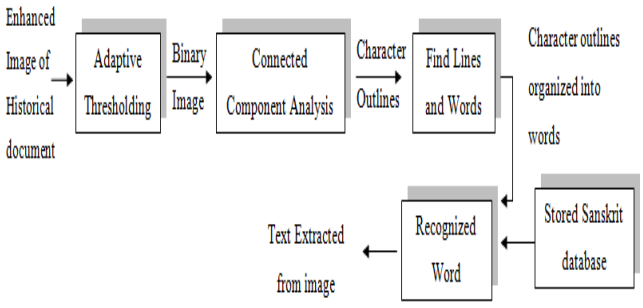


Fig. 4. Block diagram of Sanskrit data extraction form enhanced image.

III. Result Analysis

The proposed method is able to recognize most of the characters of the degraded or blurred document. The performance of proposed method is compare with bilateral filtering with binarization method and it has been found the proposed method work well is adverse condition even when background is highly contrast [14], [15]. Along with better

image enhancement we are also able to extract the most of data written on such degraded documents in Sanskrit.

TABLE II. Recognition Rate comparison on various images for bilateral filter using binarization method and the proposed method

S. No.	Input Image	Total words in document	Words recognized in Binarization method	Recognition Rate (R1) %	Words recognized in Proposed method	Recognition Rate (R2) %
1	1.jpg	415	25	6	237	57
2	2.jpg	40	22	55	35	88
3	3.jpg	161	87	54	133	83
4	4.jpg	148	11	7	25	17
5	5.jpg	60	8	13	23	38
6	6.jpg	46	2	4	8	17
7	7.jpg	105	0	0	75	71
8	8.jpg	74	5	7	44	59

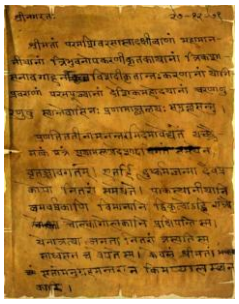
1.jpg

2.jpg

3.jpg

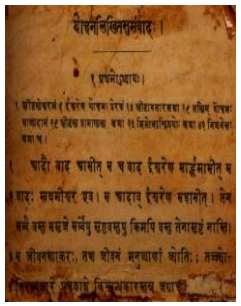
4.jpg

5.jpg



6.jpg

7.jpg



8.jpg

Fig. 5. Images of Degraded historical document used for simulation of algorithm.

IV. CONCLUSION

Quite often old documents are subject to background damage. Examples of background damages are varying contrast, smudges, dirty background, and ink through page, outdated paper and uneven background. The old Sanskrit manuscripts which are several thousand years of age, for example, are not legible even after preservation process by the library. Image processing offers a selection of approaches to counter these quality degradations and make the manuscripts readable.

The various techniques for image enhancement on old historical manuscripts have been devised. Based on previous research results, the methods have proven to improve several distinctive obstructions in the old manuscripts. This dissertation focuses on readability enhancing method of the damaged historical documents specially written in ancient language i.e. Sanskrit.

We obtained a collection of the old manuscripts images of Vedas, Granthas and other Indian ancient documents from public database available over the web and transformed them into a collection of enhanced image database. Although such manuscripts have gone through preservation process, but up to now, those manuscripts are still in poor state and few actions were taken to establish methods to make the manuscripts more readable and preservation of their contents in digital format as digital data have no time degradation affect. The processing of medium quality images of the palm leaf, paper, and wooden manuscripts is the main focus of our work.

In this paper we presented an image enhancement technique for historical degraded manuscript Sanskrit document images along with content extraction. The algorithm first adjusts the contrast and brightness of color image. Then perform histogram color balancing to reduce the degradation effect in the document by adjusting the pixel intensity value using value of α and β. This enhanced image is then converted into a grey-scale image using a linear transform to brighten the text foreground by removing most of the background colors for better contrast. Then this gray scale image is converted into binary image which fed into OCR engine for data extraction. The OCR engine performs connected component analysis to find words and line. Then engine recognize the words and lines using pre-trained Sanskrit database. The algorithm has been found to work successfully in improving readability of document images and produce high quality binarized images suitable for OCR and extracting the Sanskrit contents at comparable quality, on not only paper manuscripts but also on other aged and degraded documents such as palm leaf and historical wooden documents.

V. FUTURE WORK

There are still some limitations of our proposed methods leaving some scope for future enhancement. The image enhancement method might not work well on some document images with blurred text where color of page get light and ink of text get mixed with degraded page color. The OCR engine is trained for several text of different size for Sanskrit/ devnagri fonts. But in some cases OCR fail to recognize the text contents due to written font style difference, character shaping and special notations. Again the proposed method is able to recover and extract text contents only leaving scope for extraction of objects and graphics. In the cases when the back-side text strokes are as dark as or even darker than the front-side text strokes, the enhancement method cannot classify the two types of character strokes correctly hence fail to recover and extract the contents. In addition, the proposed method depends heavily on the high contrast image pixels. As a result, it may introduce error if the background of the degraded document images contains a certain amount of pixels that are dense and at the same time have a fairly high image contrast.

REFERENCES

[1] Z. L, Y. Zhang, and C. L. Tan, "An improved physically-based method for geometrical restoration of distorted document images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 728–734, April 2008.

[2] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," *International Conference on Pattern Recognition*, pp. 600–605, Spetember 2000.

[3] Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," *International Conference on Document Analysis and Recognition*, September 2011.

[4] Pratikakis, Gatos, and Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," *International Conference on Frontiers in Handwriting Recognition*, pp. 727–732, November 2010.

[5] R. Hedjam and M. Cheriet, "Historical document image restoration using multispectral imaging system," *Pattern Recognition*, Vol. 46, pp. 2297-2312, March 2013.

[6] S. J. Kim, F. Deng, and M. Brown, "Visual enhancement of old documents with hyperspectral imaging," *Pattern Recognition*, Vol. 44, pp. 1461-1469, July 2011.

[7] Kaminska, M. Sawczak, K. Komar, and G. Sliwinski, "Application of the laser ablation for conservation of historical paper documents," *Applied Surface Science*, Vol. 253, pp. 7860-7864, December 2007.

[8] L. Krakova, K. Chovanova, S. Selim, A. Simonovicova, A. Puskarova, A. Makova, and D. Pangallo, "A multiphasic approach for investigation of the microbial diversity and its biodegradative abilities in historical paper and parchment documents," *International Biodeterioration and Biodegradation*, Vol. 70, pp. 117-125, March 2012.

[9] R. Hedjam, R. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," *Pattern Recognition*, Vol. 44, pp. 2184-2196, June 2011.

[10] K. Khurshid, C. Faure, and N. Vincent, "Word spotting in historical printed documents using shape and sequence comparisons," *Pattern Recognition*, Vol. 45, pp. 2598-2609, April 2012.

[11] "A Review of Digital Image Enhancement Method of Degraded Indian Ancient Manuscripts " IJSRD, ISSN 2321 0613 Volume 3 Issue 3 May 2015

[12] I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, and U. Ehrlich, "Adaptive shape prior for recognition and variational segmentation of degraded historical characters," Pattern Recognition, Vol. 42, pp. 3348-3354, November 2008.

[13] R. Hedjam, R. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," Pattern Recognition, Vol. 44, pp. 2184-2196, July 2011.

[14] L. Likforman-Sulem, J. Darbon, and E. Smith, "Enhancement of historical printed document images by combining total variation regularization and non-local means filtering," Image and Vision Computing, Vol. 29, pp. 351-363, March 2011.

[15] B. Gatos et al., "Adaptive degraded document image binarization", *The journal of pattern recognition*, Elsevier publishing, doi: 10.1016, pp. 317-327, September 2005.

[16] Nitin Mishra, C. Patvardhan, et al., " Shirorekha Chopping Integrated Tesseract OCREngine for Enhanced Hindi Language Recognition", International Journal of Computer Applications, Vol. 39, No. 6, Febuary 2012.

[17] Mamata Nayak, Ajit Kumar Nayak, "Odia Characters Recognition by Training Tesseract OCR Engine", IJCA Proceedings on International Conference on Distributed Computing and Internet Technology, No. 1, December 2013.

[18] S. P. Godse, Samadhan Nimbhore, "Recovery of badly degraded Document images using Binarization Technique", *International Journal of Scientific and Research Publications*, Volume 4, Issue 5, May 2014.