Improved Variational Methodology Towards Enhancement of Marathi Printed Degraded Documents

M.S.Sonawane, C.A.Dhawale

Abstract: ***Optical Character Recognition (OCR) system aims to translate scanned text to a machine understandable text. To do so, numerous tactics exist for several scripts and so far for good quality documents. Conversely, only a delimited permutation of the same has been investigated for degraded printed Marathi documents. This work offers learning which aims to discover and fetch out a marginal and competent policy of pre-processing in treating OCR for degraded printed Marathi documents. An effective estimation of the offered substitute has been considered by exposing it to documents having bleed-through, border smear, smear inside, low illuminations, unclearness etc. Proposed methodology's results are examined in MATLAB R2015a. The work produces preprocessed images having better lucidity. Subsequent phases like segmentation, feature extraction, classification etc., offers better results with such preprocessed images having better clarity.***

Keywords : bleed-through, degraded, low illuminations, Marathi, preprocessing, smear, unclearness.

I. INTRODUCTION

In today's technological world, it is necessary to have entire existing information in a digital format acknowledged by machines. In our country, where there is a richness of information available in the form of books, documents, manuscripts, ancient texts,etc. those are conventionally presented in handwritten or printed form, such things are unsuitable when it comes to examining information between hundreds of pages. It should be digitized and renewed to a textual form with an aim to diagnose by machines doing explorations of a thousand of pages per second. It is also required to recover and mine the text which is deprecated. Only at that moment, the proper information about culture, tradition, history,etc. would be obtainable to the crowds. OCR has become one of the greatest fruitful applications in an area of artificial intelligence and pattern recognition. To distinguish handwritten or printed texts in commonly used languages like English, Japanese, Chinese, etc. judiciously competent and reasonable OCR packages are commercially presented. OCR is the utmost indispensable fragment of a document analysis system, which converts the scanned magazines, books, textinto machine comprehensible forms. Document analysis and recognition can be separated into 2 portions, which are printed, handwritten character recognition. The printed documents could be additionally allocated into 2 slices: degraded printed documents and good quality printed documents.

Degradation of the text could have bleed-through, border smear, smear inside, low illuminations, unclearness,touching characters, broken characters etc. Refer figure1.

Huge efforts have been done in Indian script recognition. In 1970, R.M.K Sinha [1],[2] at Indian Institute of Technology commenced automatic recognition of printed Devanagari script. For Devanagari script recognition, he offered a syntactic pattern analysis method [3]. Another OCR scheme for printed Devnagari script had been provided by Palit and Chaudhuri [4], Pal and Chaudhuri [5]. First viable level invention for printed Devanagari OCR industrialized by B. B. Chaudhuri, U. Pal, M. Mitra and U. Garain. Certain general anomalies in Devanagari script writing are described in Satish [6]. Efforts of degraded Gurumukhi script found in Jindal and Lehal [7]-[10].

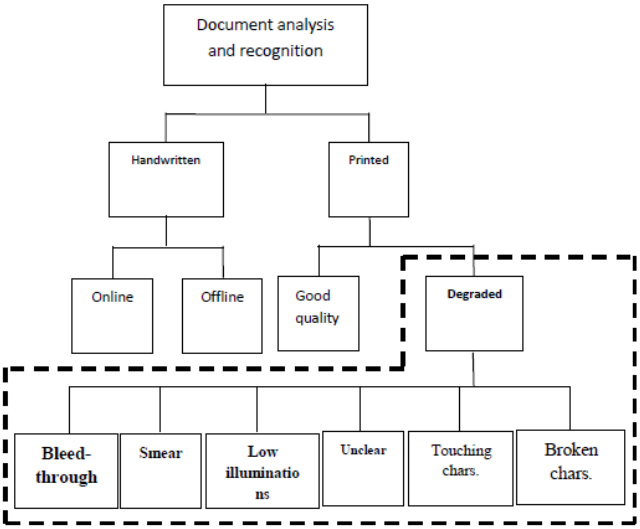

Fig. 1.Document Analysis and Recognition

Bansal [11] and several other investigators had fingered the problem of degradation of the Devanagari script, however extreme degradation was not deliberate. Therefore,it is the region where huge investigation is vital. In considering the good quality printed documents in English, Gurumukhi, Devanagari, etc. several OCR systems are presented [22]-[25]. However, no work available for recognition of degraded printed Marathi documents [12]-[15]. The outcomes of projected methodology are realized to be improved as equated to 6 present methods proposed by Otsu, Gatos, Niblack, Souvola, Bernsen and Brij Mohan Singh. This effort tries to improve the quality of degraded printed Marathi documents while doing preprocessing, so that classification, recognition can be done smoothly.

II. PROPOSED WORK

The proposed work attempts to expand the eminence of degraded printed Marathi documents which is done in the preprocessing phase of OCR. The proposed work is as follows.

Step1: To make the experiment, scanned document image will be read.

Step2: Convert the image into a grayscale image, whenever an input image is a colored.

Step3: Do wiener filtering with aim to remove the noise in an image.

Step4: In order to do the smoothing, apply Gaussian filtering.

Step5: Adaptive histogram equalization will be applied to enhance the contrast of images.

III. PERFORMANCE MEASURES

There exist numerous evaluation parameters. Peak signal to noise ratio (PSNR) and Mutual information (MI) parameters are used for assessment of proposed work.

A. Peak Signal To Noise Ratio

It is the ratio between the maximum possible power of signal and power of corrupted image. Size and class of the input image and output image should be same. PSNR is used as quality measurement between images. The higher PNSR value in the restored image provides superior class. Decibel unit is used to define PSNR. For gray scale image it is explained as,

Where,

MSE - Mean Square Error.
PSNR - Peak Signal to Noise Ratio.
M x N – Image size.
X - Imaginative Image.
R - Restored Image.

B. Mutual Information

It belongs to probability or information theory which measures the amount of information that one variable contains about another. Mutual information is the similarity measure between images. It qualifies amount of information in units like shannons or bits obtained about one variable through the other.

IV. RESULTS AND DISCUSSION

- For experimental purposes, visited various places like Deccan College Postgraduate and Research Institute, Pune, Jaykar library of Savitribai Phule Pune University,The Vagdevata Mandir Dhule, The Bhandarkar Oriental Research Institute, Pune and The Bharat Itihas Sanshodhan Mandal Pune etc. in order to find degraded printed Marathi documents. For this work more than 1000 differently degraded printed Marathi documents from several books have been used. Figure 2 to figure 11 shows differently degraded images and corresponding enhanced images. The PSNR and MI values of 10 randomly selected images are shown in table1. Table1 shows best PSNR and MI values for such images. Comparative PSNR values for six existing methods and proposed method are shown in table 2. Table 2 demonstrates that for Otsu's technique PSNR is 07.68, 39.78 is for Gatos's scheme, for a Niblack's system, it is 17.26. 29.34 for Sauvola's technique. In case of Bernsen's its 21.09. 40.42 got for Brij Mohan. Proposed method's PSNR, 57.51 is the maximum and Otsu's contains the lower most 07.68 PSNR.

Table- I: Experimental Results

Figure	PSNR Value	MI Value
Figure2	54.1868	1.1778
Figure3	56.8300	1.0000
Figure4	60.1657	1.0022
Figure5	57.2634	0.9737
Figure6	55.6829	0.8026
Figure7	58.2788	0.9458
Figure8	57.0402	1.1201
Figure9	57.5125	1.0749
Figure10	64.0813	0.9118
Figure11	58.8876	0.7849

Table- II: Comparative Results

Method	PSNR Value
Otsu [17]	07.68
Gatos [18]	39.78
Niblack [19]	17.26
Sauvola [20]	29.34
Bernsen [21]	21.09
Brij Mohan [16]	40.42
Proposed	57.51

V. FIGURES

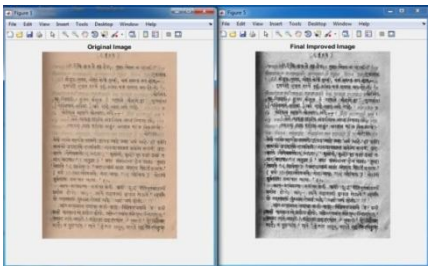

Fig. 2.Bleed-through Sample1(Left- Degraded image, Right-Enhanced image)

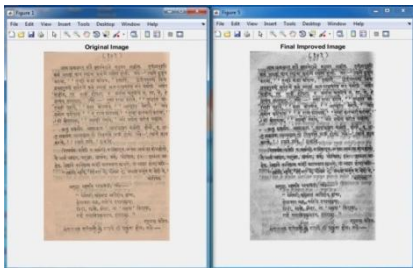

Fig. 3.Bleed-through Sample2(Left- Degraded image, Right-Enhanced image)

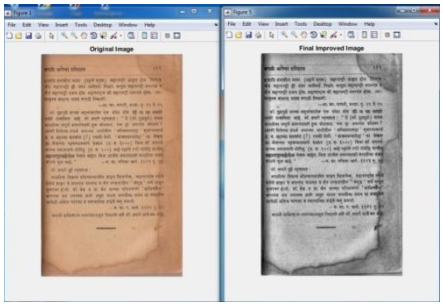

Fig. 4.Border Smear Sample1(Left- Degraded image, Right-Enhanced image)

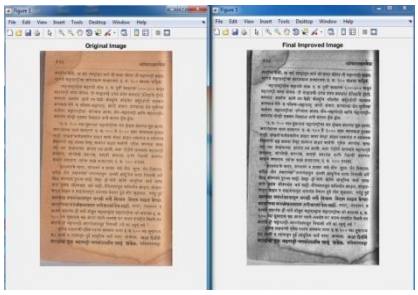

Fig. 5.Border Smear Sample2(Left- Degraded image, Right-Enhanced image)

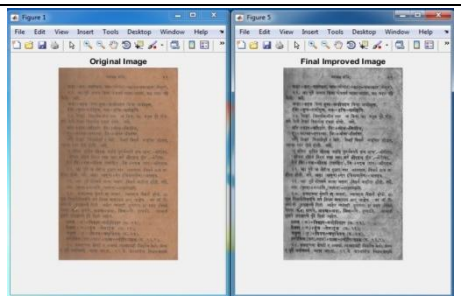

Fig. 6.Low Illumination Sample1(Left- Degraded image, Right-Enhanced image)

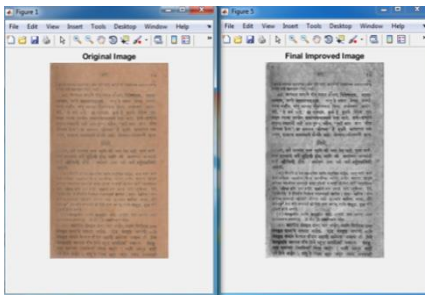

Fig. 7.Low Illumination Sample2(Left- Degraded image, Right-Enhanced image)

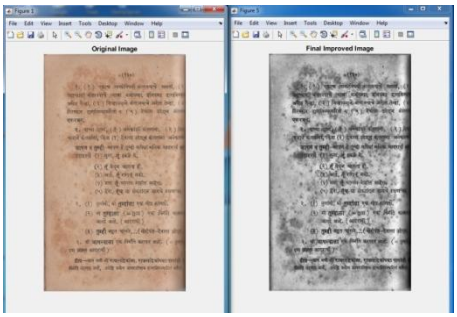

Fig. 8.Middle Smear Sample1(Left- Degraded image, Right-Enhanced image)

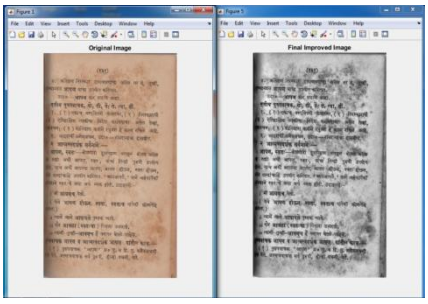

Fig. 9.Middle Smear Sample2(Left- Degraded image, Right-Enhanced image)

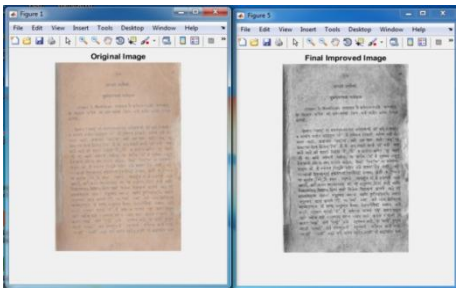

Fig. 10.Unclear Sample1(Left- Degraded image, Right- Enhanced image)

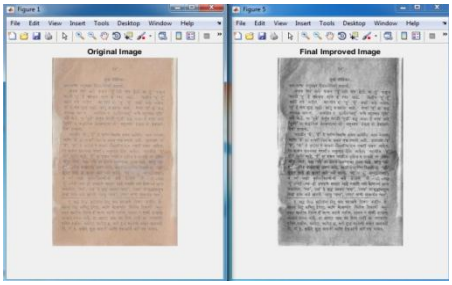

Fig. 11.Unclear Sample2(Left- Degraded image, Right- Enhanced image)

VI. CONCLUSION AND FUTURE SCOPE

Lots of work has been done for OCRs for decent quality printed documents in English, Gurumukhi, Devanagari, etc. Many researchers had handled the degradation problem of the Devanagari script, though punishing degradation was not considered. This work attempts to progress the quality of highly degraded printed Marathi documents during the preprocessing phase, so that further phases of OCR can be applied effortlessly. This work produced better quality images with higher PSNR values and good values of MI as well. Future attempt will do word segmentation and character segmentation in such a highly degraded printed Marathi document.

REFERENCES

1. V. Bansal and R.M.K. Sinha, ―Partitioning and searching the dictionary for correction of optically read Devanagari character strings, International Journal of Document Analysis and Research, Vol.4, pp.269–280, 2002.
2. K. D. Dhingra, S. Sanyal, P. K. Sharma, ―A robust OCR for degraded documents, Advances in Communication Systems and Electrical Engineering, Huang et al., (Eds.), Lecture Notes in Electrical Engineering, Springer, pp. 497-509, 2008.
3. R.M.K. Sinha, ―A Syntactic pattern analysis system and its application to Devnagari script recognition, Ph.D. Thesis, Electrical Engineering Department, Indian Institute of Technology, Kanpur, India, 1973.
4. R M K Sinha, H Mahabala, ―Machine recognition of Devnagari script, IEEE Transactions on Systems, Man and Cybernetics, Vol.9, pp.435–441, 1979.
5. U. Pal, B.B. Chaudhuri, ―Printed Devnagari script OCR system, Vivek, Vol.10, pp.12–24, 1997.
6. Satish Kumar, An Analysis of Irregularities in Devanagari Script Writing – A Machine Recognition Perspective, International Journal on Computer Science and Engineering (IJCSE) Vol. 2, No. 2, 2010.
7. M. K. Jindal, G. S. Lehal and R. K. Sharma, ―On Segmentation of touching characters and overlapping lines in degraded printed Devanagari script, International Journal of Image and Graphics (IJIG), World Scientific Publishing Company, Vol. 9, No. 3, pp. 321-353, 2009.
8. M. K. Jindal, R. K. Sharma and G. S. Lehal, ―Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts, International Journal of Computational Intelligence Research (IJCIR), Research India Publications, Vol. 3, No. 4, pp. 277-286, 2007.
9. M. K. Jindal, R. K. Sharma and G. S. Lehal, ―Segmentation of Touching Characters in Upper Zone in printed Devanagari Script, in Proceedings of 2nd Bangalore Annual Compute Conference on 2nd Bangalore Annual Compute Conference, (Bangalore, India, January 09 - 10, 2009). COMPUTE '09. ACM, New York, NY, 1-6.
10. M. K. Jindal, R. K. Sharma and G. S. Lehal, ―Structural Features for Recognizing Degraded Printed Devanagari Script, in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, April 2008.
11. V. Bansal, R. M. K. Sinha, ―Integrating knowledge sources in Devanagari text recognition , IEEE Transactions on Systems Man and Cybernetics Part A: Systems & Humans , Vol.30, No.4, pp.500–505, 2000.
12. Bolan Su, Shijian Lu and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE Transactions on Image Processing, Vol. 22, No. 4, April 2013.
13. Yung-Hsiang Chiu , Kuo-Liang Chung , Wei-Ning Yang, Yong-Huai Huang and Chi-Huang Liao, "Parameter-free based two-stage method for binarizing degraded document images", Y.-H. Chiu et al. / Pattern Recognition 45 (2012) 4250–4262, Elsevier 2012.
14. Konstantinos Ntirogiannis, Basilis Gatos and Ioannis Pratikakis, "A Performance Evaluation Methodology for Historical Document Image Binarization", IEEE 2011.
15. Vavilis Sokratis, Ergina Kavallieratou, Roberto Paredes and Kostas Sotiropoulos, "A Hybrid Binarization Technique for DocumentImages", Springer 2011.
16. Brij Mohan Singh, Mridula, "Efficient binarization technique for severely degraded document images", CSIT (November 2014) 2(3):153–161 DOI 10.1007/s40012-014-0045-5.
17. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 9(1):62–66.
18. Gatos B, Pratikakis I, Perantonis SJ (2006) Adaptive degraded document image binarization. Pattern Recognit 39:317–327.
19. Niblack W (1986) An introduction to digital image processing. Prentice-Hall, Englewood Cliffs, pp 115–116.
20. Sauvola J, Pietikainen M (2000) Adaptive document image thresholding. Pattern Recognit 33:225–236.
21. Bernsen J (1986) Dynamic thresholding of grey-level images. In:Proceedings of the eighth ICPR, pp 1251–1255.
22. C.A.Dhawale and M.S. Sonawane, ―Performance analysis of image enhancement techniques for OCR, International Journal of Pharmacy and Technology, Vol. 9, Issue No.2, 29766-29774.
23. C.A.Dhawale and M.S. Sonawane, ―Evaluation of character recognizers: Artificial neural network and nearest neighbor approach, 978-1-4799-6023-1/15 $31.00©2015IEEEDOI 10.1109/CICT.2015.30.
24. C.A.Dhawale and M.S. Sonawane, ―Performance Evaluation of Classification Techniques for Devanagari Script, 978-1-4673-9354-6/15/$31.00 ©2015 IEEE.
25. C.A.Dhawale and M.S. Sonawane, ―Evaluation and Analysis of Few Parametric and Nonparametric Classification Methods, 978-1-5090-0210-8/16 $31.00 © 2016 IEEE DOI 10.1109/CICT.2016.13.

AUTHORS PROFILE

Manojkumar Sahebrao Sonawane, have completed M.Sc.(Computer Science) in 2005, MCA in 2012. Qualified SET in year 2015, NET in 2019. Currently persuing Ph.D. in computer science in SGBAU, Maharashtra under supervision of Dr.C.A.Dhawale. Have publications in IEEE, Springer, IJPT, IJCA, IJSR etc. conference, journal, proceedings. Area of interest is image processing. Research work is on optical character recognition. Have 13 years of teaching experience in computer field. Participated in several conferences, workshops, seminars, faculty development programs, poster presentation competitions etc. Take part in syllabus framing wokshops. Organized various workshops, seminars, and guest lectures. Worked as coordinator of CSI student branch. Awarded with second price for university and state level research activity Avishkar-2009.

Dr. Chitra A. Dhawale. Awarded Ph.D (Computer Science) in 2009. subject specialization is image processing. Total 21 years teaching experience. Papers published in national journals-02, international journals-19. Papers presented in National conferences-19, International conferences-26. Ph.D. guide of computer science in SGBAU, Maharashtra. 3 book chapters are in IGI Global. Filed 1 patent. Professional memberships are life member ISTE, New Delhi, Senior member of the IACSIT, Member of International Association of Engineers (IAENG), Hong Kong. 1 proposal submitted to DST. Interaction with professional institutions like International Journal of Computer Application, International Journal of Electronics and Electrical Engineers, International Journal of Innovative Research in Information Security, International Journal of Advancements in Computing Technology, International Journal on Advances in Information Sciences and Service Sciences, International Journal of Computer Science & Systems, International Journal of Engineering and Advanced Technology, Covenant Journal of Computer Science, Vietnam, Bio-Info Publications etc.