International Journal of Control and Automation
Vol. 12, No. 4, (2019), pp. 194-198

Hybrid Binarization Approach for Typewritten Gurmukhi Script Documents

Rajan Goyal[1], Rajesh Kumar Narula[2], Manish Kumar Jindal[3]

[1] *Research Scholar, I.K. Gujral Punjab Technical University Kapurthala-144603, Punjab, India*

[2]*Assistant Professor, Department of Mathematical Sciences, I.K. Gujral Punjab Technical University Kapurthala-144603, Punjab, India*

[3] *Professor, Panjab University Regional Centre, Muktsar-15026, Punjab, India*

Abstract: In this digital world having motto "Save Paper Save Trees", most of the documentation work is being transformed into digital format. This paper presents a hybrid approach for the binarization of Typewritten Gurmukhi script documents. These typewritten documents suffer from various degradations such as ageing, faded characters and noise. In typewritten documents, retaining of characters is very important for character segmentation. Binarization is the preprocessing phase of the Optical Character Recognition system. We have presented a hybrid approach for image restoration by using the blend of local and global threshold method. We have also compared results of our technique with other binarization techniques: Sauvola, Otsu, Bernsen and H- DIBCO'10. Our approach combines the noise removal and image restoration goals into a single framework, thus producing high quality document images from degraded typewritten documents.

Keywords: Binarization, Typewritten, Gurmukhi, OCR, Image Restoration.

1. Introduction

In the world of digitization, there is need to transform the text documents into machine processable form. During the transformation process, enormous amounts of document records are conserved through electronic scanning. These text documents are accessible from different sources, for example, old documentation, official records of different departments, security-related records, and land records etc.

The performance of an Optical Character Recognition (OCR) system often depends upon the quality of a scanned document image. Binarization is the pre-requisite step for any OCR system. Document image binarization is the transformation of scanned document image into bi-level image format where image pixels are divided into two categories, *i.e.,* foreground pixels and background pixels. There are various binarization techniques proposed in the literature but still, deteriorated documents affect the performance of these techniques. So, there is a need to propose a technique which removes the noise from the text document image in such a way that maximum characters can be retained. In OCR system removing the noise is not enough, as character restoration is also most important. Typewritten Gurmukhi script documents suffer from various problems such as natural ageing, low paper quality, noise, and broken/ heavy printed characters. Earlier techniques remove the noise from these documents but suffer from broken/misprinted characters. Specifically, old and chronicled text documents are difficult to peruse due to their debasement as far as low contrast and presence of noise. A few articles are available on the binarization of typewritten documents [1].
We propose a way to deal with restoring typewritten Gurmukhi text images using a hybrid approach. Un- like earlier methodologies that use recently learned prior models to retain a document image; we can gain proficiency with the content model from the corrupted record itself.

We have collected 503 typewritten Gurmukhi documents from various departments of Punjab such as Local Audit, Municipal Corporation, Panchayati Raj, Block Development office, Court, etc. Some of these documents are in the photocopied form and some are scanned directly. The quality of photocopies is also not very good as it depends upon the machines used in government offices. As the data collection is from various sources, so the noise level, character quality, paper quality of all the documents are different. After the collection of documents, these were scanned in RGB form at 300 dpi resolution over the Cannon iR3035 machine which is then passed to preprocessing phase of the typewritten recognition system.

2. Related Work

Most traditional text image enhancement algorithms have been designed primarily to extract text from noisy documents with uneven background and suffer from restoration of faded characters.
Some techniques work well under the normal conditions of an image which is sufficient for binarization, but these techniques are not suitable for degraded document images [2],[3]. Simple global thresholding produce good results for normal images. Some local threshold window-based techniques also produce fine results under noisy document images but suffer from restoration [4, 5]. There are various historical documents which suffer from blobs, vandalized, cuts, and merges which can be removed by modeling the document image over large patches [6]. This technique is independent of font, style and script etc. To find out the best binarization algorithm there are various quality measures such as PSNR, NRM, MPM etc. [7].

Degradations such as smear through, faded characters, bleed through, contrast variation, uneven illumination, blur issue, thin or weak text and deteriorated documents make the binarization process more challenging and an extensive review about these deficiencies and related work has presented in [1]. These are some of the notable difficulties faced by an OCR framework. The number of articles reported on binarization of typewritten documents is few. An effort to improve bi-level and different contrast typewritten images has made by Cannon et al. [8]. An attempt has made by A. Antonacopoulos [9] to extract the typewritten data based on expert user specified semantic information. Then again A. Antonacopoulos [10] has presented an approach for typewritten documents based on location of characters and applied various binarization methods at reviewing the document at precise level. These two methods were used to facilitate the OCR system.

Sometimes contrast in the image may vary due to lighting effect so there is need to calculate the contrast differences in different regions of the image and then corresponding enhancement techniques can be applied [11]. Some post-processing techniques such as de-speckle, preserve stroke connectivity may improve the binarization results [12]. Recently the use of various machine learning techniques has been done for binarization process. To select the optimal global thresholding, support vector machine is used for binarization by dividing the image in various segments[13]. Iterative Global thresholding is applied by [14] to remove the uncertain noise and extracting the character forms by applying multiple binarization techniques.

Three popular methods, namely Bernsen's dynamic thresholding method [5], Otsu's thresholding technique [2], and the Sauvola local threshold window-based technique [4], and B. Su Method [15],[16] are analyzed and compared.

3. Problem Analysis

The typewritten Gurmukhi documents collected from various departments of Punjab are considered as an important part of official record. These records suffer from deterioration and therefore risk content disappearing. In order to restore and use these type of documents, digitization is a definite advantageous. There are various methods available, as mentioned in related work review section, for conserving the ancient archives. But still these techniques can't generate productive outcome as required.
Generally, there are three types of degradations in the quality of typewritten Gurmukhi document images. First, the original paper document is suffered from natural ageing leading to degradation of the paper media, and added dirt. The second problem is of faded text which may lead to loss of characters during scanning process, so restoration of document image content is major objective. The third problem is original document paper itself. We have observed that some original typewritten documents are of thin low-quality paper, so the paper quality and the working mechanism of typewriter produce a degraded document although the document may be printed by typewriter at the same instant. Due to above degradations of typewritten documents, selection of right binarization technique is must and challenging task. In any OCR system, the segmentation of content from the document image background is often depends upon the outcome of binarization technique.

Our method is focused towards enhancing images in such a way that noise can be removed and to perform effectively to restore the faded characters. First target the problem of noise found in digital document images using wiener filter. Then using the gradients of Sobel and Robert filters and Sauvola [4] approach characters of

the image is restored. The noise removal enhances the document image, making it more legible to the eye as well as leads to segmentation of the text from the document image.

4. Proposed Hybrid Approach

By the intensive review of various techniques, we have proposed a hybrid approach which combines the advantages of various techniques into a single framework. To remove the adaptive noise and to reduce the blurring effect from the image we have applied wiener filter. While, Sobel filter is applied to detect the horizonal edges from an image, as Sobel filter cannot detect diagonal edges so we have applied Robert filter. To calculate the window based local thresholding value, we have applied Sauvola [4] approach on the image. Figure 1 shows the complete hybrid binarization approach.

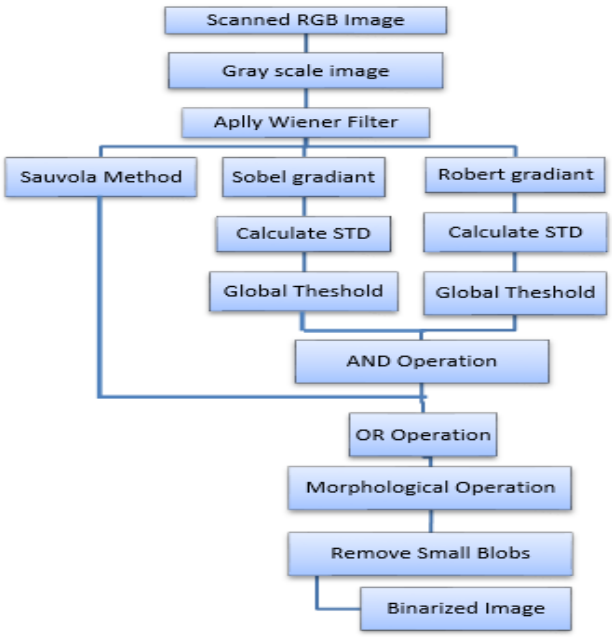

Figure 1: Proposed binarization process

We have presented a hybrid approach for image restoration by using the blend of local and global threshold method.

Step1: Firstly, scanned RGB image is converted into grayscale image.

Step2: Wiener filter is applied to remove the adaptive noise and store the result in image I_1.

Step3: Then Sauvola approach is applied on image I_1 and store the resultant in image I_2.

Step4: Now, the gradients of Sobel and Robert filters are calculated on image I_1 and calculate their edges by using standard deviation and global threshold value.

Step5: Then logical AND operation is applied on resultant of Sobel and Robert filter and store it in image I_3.

Step6: Now, the logical OR operation is applied on images I_2 and I_3.

Step7: Then resultant image is eroded with disk structure element having value '1'.

Step8: Finally, small blobs from the image are removed using connected component method.

We have compared the results of our algorithm visually with other binarization techniques such as Sauvola, Otsu, Bernsen and H-DIBCO'10 (B.Su Method)[15],[16] and found that the output document with our work contains better result and also faded characters were retained efficiently as shown in Figure 2. In typewritten documents, retaining of characters is very important for character segmentation.

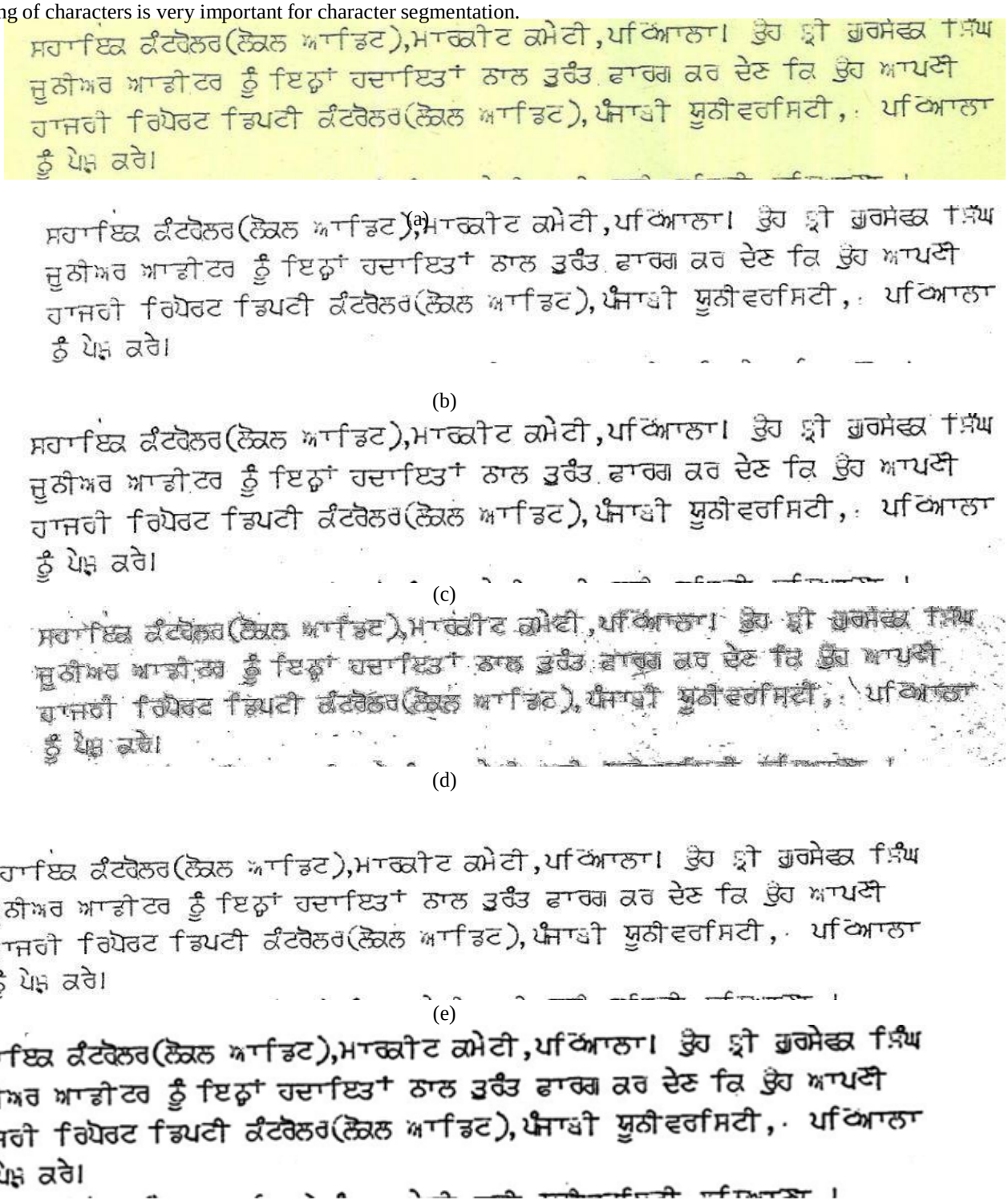

Figure 2: Results of binarization on part of document (a) Original image (b) Otsu method (c) Sauvola method (d) Bernsen method (e) H-DIBCO'10 method. (c) Proposed method.

5. Conclusion

Binarization process is the basic need of an OCR process, as high character restoration leads to high accuracy of an OCR system. The proposed technique combines the advantage of various techniques to a single framework, where adaptive noises is removed and from each direction content image's edge are detected and characters are restored at great extent by further applying local thresholding. Our dataset of typewritten Gurmukhi images contains various degradations such as ageing, faded characters, as discussed in the problem analysis section, so using the proposed technique we have removed the extra noise and restored the faded characters at maximum extent. Proposed technique provides the best results on the degraded images.

Acknowledgement
One of the authors (Rajan Goyal) would like to express his special thanks of gratitude to I.K.Gujral Punjab Technical University, Kapurthala for providing unrestrained access to resources needed for research.

References

1. Sulaiman, A., Omar, K., & Nasrudin, M. F. (2019). Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions. *Journal of Imaging*, *5*(4), 48.
2. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62-66.
3. Karthika, M., & James, A. (2015). A novel approach for document image binarization using bit-plane slicing. *Procedia Technology*, *19*, 758-765.
4. Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, *33*(2), 225-236.
5. Bernsen, J. (1986). Dynamic thresholding of gray-level images. In *Proc. Eighth Int'l conf. Pattern Recognition, Paris, 1986*.
6. Banerjee, J., Namboodiri, A. M., & Jawahar, C. V. (2009, June). Contextual restoration of severely degraded document images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 517-524). IEEE.
7. Vats, E., Hast, A., & Singh, P. (2017, November). Automatic document image binarization using bayesian optimization. In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing* (pp. 89-94). ACM.
8. Cannon, M., Hochberg, J., & Kelly, P. (1999, May). Quarc: A remarkably effective method for increasing the ocr accuracy of degraded typewritten documents. In *Proceedings of the 1999 Symposium on Document Image Understanding Technology (SDIUT'99)* (pp. 154-158).
9. Antonacopoulos, A., & Karatzas, D. (2005, August). Semantics-based content extraction in typewritten historical documents. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (pp. 48-53). IEEE.
10. Antonacopoulos, A., & Castilla, C. C. (2006, August). Flexible text recovery from degraded typewritten historical documents. In *18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 2, pp. 1062-1065). IEEE.
11. Lu, D., Huang, X., & Sui, L. (2018). Binarization of degraded document images based on contrast enhancement. *International Journal on Document Analysis and Recognition (IJDAR)*, *21*(1-2), 123-135.
12. Chen, Y., & Wang, L. (2017). Broken and degraded document images binarization. *Neurocomputing*, *237*, 272-280.
13. Xiong, W., Xu, J., Xiong, Z., Wang, J., & Liu, M. (2018). Degraded historical document image binarization using local features and support vector machine (SVM). *Optik*, *164*, 218-223.
14. Boudraa, O., Hidouci, W. K., & Michelucci, D. (2019). Degraded Historical Documents Images Binarization Using a Combination of Enhanced Techniques. *arXiv preprint arXiv:1901.09425*.
15. Pratikakis, I., Gatos, B., & Ntirogiannis, K. (2010, November). H-DIBCO 2010-handwritten document image binarization competition. In *2010 12th International Conference on Frontiers in Handwriting Recognition* (pp. 727-732). IEEE.
16. Lu, S., Su, B., & Tan, C. L. (2010). Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, *13*(4), 303-314.