RESEARCH REVIEW JOURNALS
NCIDCAS-2018
DEC- 2018
Special Issue

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

# A study on binarization for degraded historical document

[*1]Nagane Aniket; [2]Mali Shankar; [3]Gursale Sneha & [4]Kshetre Diksha

[*1]*Assistant Professor, Department of Computer Science, MIT ACSC,Alandi(D),Pune, Maharashtra (India)*
[2]*Associate Professor, Department of Computer Science, Dr. Vishwanath Karad MIT World Peace University, Pune (India)*
[3,4]*Student, Department of Computer Science, MIT ACSC,Alandi(D),Pune, Maharashtra (India)*

**ARTICLE DETAILS**

**Article History**
*Published Online:* 22 Dec 2018

**Keywords**
binarization, degraded document, thresholding, image enhancement

[*]**Corresponding Author**
*Email: asnagane[at]mitacsc.ac.in*

**ABSTRACT**

Binarization is one of the very initial and important step in image processing. In OCR systems also binarization is necessary as a part of image enhancement. This paper presents study on binarization of degraded historical documents. As our rich heritage and culture can be unveiled and understood from the old and ancient documents, OCR for oldest scripts is a today's need. Old ancient available documents are mostly in degraded form. This paper presents study on binarization of degraded historical documents and notes the observations as there are many different and approaches followed to achieve better binarization of historical documents. Methods studied and presented in this paper yields good results but for specific type of documents.

## 1. Introduction

Binarization is defined as converting colour/grey image into binary image, i.e. image with more than two intensities is converted into a image with only two intensity levels 0 and 1. Document binarization is considered as process of image enhancement. Image enhancement refers to a process of converting input image into a format, which is suitable for a specific application. In Optical Character Recognition systems separation of text (are of interest) from input image as foreground and non-interest area as background is necessary.[1] Unless until the text area is extracted from image, the process of recognising the text cannot be performed. So binarization is non-trivial, significant steps in OCR systems. If binarization is not performed correctly, it may affect the systems performance.

The rich heritage and culture of our history is found recorded on documents which are very old and has been degraded over the period of time.[2] To preserve the history digitization of such old ancient documents is needed. Now a days many organizations are have started maintaining all such old and ancient documents in the digitized form. But being old documents, degradation in the various forms is observed such as of bleed through effect, smear and smudge, uneven illumination etc. Because of natural degradation of the documents, these are difficult to read. Enhancement can be performed on such documents so as make them readable. Enhancement can be achieved with binarization process. Many researchers have proposed promising methods for degraded historical documents, but still there is enormous scope for the improvement in the existing and proposed methods. In this paper we have studied methods that can be applied to degraded historical documents.

## 2. Study of Literature

Karthika and James [3] displayed a technique to perform record picture binarization utilizing bit-plane cutting. The proposed strategy is sent dependent on the standard of partition and-overcome. According to this procedure, the 8 bit planes of the dim picture are extricated and prepared independently, toward the finish of which, the outcomes are consolidated to give the last binarized yield. This method is expected to yield good results for all kind of document images with different types of noises

Farrahi and Cheriet [4] have proposed an adaptive and parameter-less generalization of Otsu's technique for binarization. The proposed method is based on Otsu's binarization method. The method does not require parameters as it uses the required information/parameters from the archived images, i.e. from archived images the parameters are extracted and use in processing the images and we need not pass the corresponding parameters. The method obtains background map which is used to distinguish between content and non-content areas of the images. The method is tested on standard datasets and delivers promising results.

Moghaddam and Cheriet [5] exhibited a adaptive and parameterless generlization of Otsu's strategy. The adaptiveness is acquired by fusing grid-based modeling and the estimated background map. The parameterless actions is accomplished via automatically estimating the document parameters, for example, the normal stroke width and the normal line tallness. The proposed strategy is expanded utilizing a multiscale structure, and has been implemented on different datasets, with promising outcomes.

Liu and Shrihari [6] proposed a binarization algorithm for degraded or poor quality documents. The proposed method is suitable for the documents having poor contrast or variations in contrast. Proposed method is based on texture features of documents. Proposed method consists of three steps:

i. Calculate candidate threshold using Otsu's algorithm iteratively
ii. Extraction of texture features associated with each candidate threshold with the help of run-length histogram of the accordingly binarized image;
iii. Selection of optimal threshold to preserve document texture features.

Experiments stated that the new thresholding method, is appreciably better than those obtained by typical existing thresholding techniques.

Su et al. [2] proposed document image binarization technique for the documents which contains considerable variations in the intensities of foreground text and document background within the same document. The proposed method computes adaptive contrast map which is combination of local image contrast and local image gradient. Then the binarization is carried out on the constructed image constrast, which is further combined with Canny's edge map. This helps to locate the pixels that belongs to text stroke. Further local thresholding is applied on these located text stroke pixels to extract the text out of the document background. The proposed method is simple, robust, and involves minimum parameter tuning.

Gatos et al. [7] presented an adaptive approach for the binarization and enhancement of degraded documents. Proposed method performs pre-processing as initial step, followed by separation of foreground and background regions. Then the thresholding is executed combining estimated background with original image. After extensive experiments, method demonstrated superior performance against four (4) well-known techniques on numerous degraded document images.

Khurshid et al. [8] presents a new method which is based on Niblack's thresholding technique. Thresholding in Niblack's method and proposed method standard deviation and local mean of all pixels in a given window. The proposed method is more suitable for the ancient document images with more brightness or the images with white background. The proposed method is tested on low quality ancient document images and performs considerably well.

Su et al. [9] presents a binarization technique to distinguish or extract the text from poorly degraded historical document images. The proposed method constructs the contrast image of original input image and identifies the pixels with high contrast. By applying local threshold which is derived from high contrast pixels, the text is separated from the document. To compute the image contrast, image gradient is compared with local maximum and minimum of the image. The results of the proposed method are significant with old degraded documents.

Yang and Yan [10] presented a thresholding method for binarization of very poor quality gray-scale document images. The proposed method is suitable for document images having uneven intensities in background and foreground of the image. Initially using run-length histogram connectivity of text stroke and grouping of characters is checked. Then, authors propose a modified logical thresholding method to extract the binary image adaptively from the degraded gray-scale document image with complex and inhomogeneous background. Proposed method gives really good results for the low quality documents.

Ntogas and Ventzas [11] proposed that binarization procedure in which first de-noising is carried out with low pass filters, then local thresholding algorithms are applied on the pre-filtered image and in the last step post-processing is done as refinement with the help of erosion and dilation operations. The main contribution of this paper is to propose a simple and robust binarization procedure for pre-filtered historical manuscripts images, and simulation results are also presented.

## 3. Conclusion

From the studied literature in this paper, it has been observed that a lot of work has been done on binarization of degraded historical documents. Existing work is delivering good results but the methods are restricted to certain types of documents to some extent. Also it has been observed that binarization of degraded documents is a complicated process and only single method or approach may not give the better binarization output, whereas combining multiple binarization methods may give good results. After analysing the studied literature, it has been observed that, there is still good scope for the improvement in binarization techniques needed for degraded historical documents.

***Acknowledgement***

The authors would like to thank Dr. C. H. Patil and Mrs. Rashmi Lad for their support in the research work.

## References

1. J. Sauvola, M. PietikaKinen (1999) "Adaptive document image binarization". Pattern Recognition 33 (2000) 225–236
2. B. Su, S. Lu and C. L. Tan (2013) "Robust Document Image Binarization Technique for Degraded Document Images"IEEE Transactions on Image Processing 22(4):1408–1417
3. M. Karthika and A. James (2014) "A proposed method for document image binarization based on bit plane slicing" Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on IEEE
4. M. R. Farrahi and M. Cheriet (2012) "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization"Pattern Recognition 45(6):2419–2431
5. R. F. Moghaddam and M. Cheriet (2010) "A multi-scale framework for adaptive binarization of degraded document images" Pattern Recognition43:2186–2198
6. Y. Liu and S. N. Shrihari (1997) "Document Image Binarization Based on Texture Features" IEEE Transactions on Pattern Analysis and Machine Intelligence19(5):540–544
7. B. Gatos, I. Pratikakis and S.J. Perantonis (2006) "Adaptive degraded document image binarization" Pattern Recognition 39:317–327
8. K. Khurshid, I. Siddiqi, C. Faure, N. Vincent (2009) "Comparison of Niblack inspired Binarization methods for ancient documents"Proceedings of SPIE-IS&TElectronic Imaging 7247:72470U
9. B. Su, S. Lu and C. L. Tan (2010) "Binarization of historical handwritten document images using local maximum and

minimum filter" International Workshop on Document Analysis Systems 159–165

10. Y. Yang and H. Yan (2000) "An adaptive logical method for binarization of degraded document images" Pattern Recognition 33:787–807
11. N. Ntogas and D. Ventzas (2008) "A Binarization Algorithm for Historical Manuscripts"12th WSEAS International Conference on Communications, Heraklion, Greece, July 23-25.