Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
ISSN NO. 2320-5407
Article DOI:10.21474/IJAR01/7902
DOI URL: http://dx.doi.org/10.21474/IJAR01/7902

RESEARCH ARTICLE

RECOVERING OLD DOCUMENT IMAGE BY USING HYBRID BINRAZATION TECHNIQUE.

Rana mohamed H. Zaki and Teaba Wala Aldeen Khairi.
Computer Science Department, University of Technology/Baghdad.

Manuscript Info

Manuscript History
Received: 12 August 2018
Final Accepted: 14 September 2018
Published: October 2018

Keywords:-
RGT, LT, Binarization, restore image, ancient document recovery..

Abstract

The ancient document that sored in museum and library required recovering of its contain, since the paper loss it's shiny and bright, year after year because a lots of factors. So that, the necessity of keeping the old document as the same as the new one is needed. However, keeping the old document required huge amount of money every year and sometimes, contains of old document can't recover. Most of library tried to remake new version of document by trying to copy that image in new paper. However, some of document required treatment before copying because some letter may not recognize. So to overcome this problem a program is used in this paper to enhancement to recognize the letter in old document and make the letter easy to read in order to copy it. In this paper, a hybrid Binarization algorithm used to enhance the old document. The algorithm will make some preprocessing on old document image. The preprocessing consists of four section the first one is smooth the image of old document to be ready for entering the algorithm which give good enhanced the algorithm to detect the letters. The algorithm used local and global thresholding to detecting the letters. Also, anther operation applied on the image to remove the noise from the enhanced image of the old document image.

Introduction:-

Any physical material is exposed to damage or hurt during ages. The damage may come from lots of things such as environment like humidity and sunlight or damage also may be ink drop and uneven illumination on physical material. [1]These factors may effect on scientific and important cultural material. So to maintain these physical and ancient culture materials some of algorithms and ways are used to improve the quality and remove the clutter by using lots of filters on those materials[2].

However, these algorithms and ways cost a lot of money and sometimes may not work or can't be applied on that material because the ancient's material may decay during the age and can't be repair. So for these reason, many researcher goes to createalgorithms and methods using digital way to restoration and maintain the physical and ancients culture materials. Digital algorithms are high accurate and doesn't need any cost. Also it stays forever since ages never effects on it and it can be copied to different computers or different stations. Also the digitalized algorithm helps a lot of researcher since it provided the copies of important document with high accuracy and also inherits our next generation a document that contain very important information that may be lost during the ages. So majority of museums and libraries using the digital algorithm to shares the important document to people without afraid of lose it or damage it by people[3].

Corresponding Author:-Rana mohamed .h. Zaki.
Address:-Computer Science Department, University of Technology/Baghdad.

The digital algorithms that used in this paper for documentareconsisting from sequence of procedure. These procedures are picturing or scanning document and then pass these document to filter, which used to improve the quality and smooth the image the higher the accuracy of filtering the higher the pervious process will be. After filtering process the filtered image will be processed by Binarization (BZ)[4] which is a process that converts the grey-scale picture to binary form. The BZ will be able to distinguish between background foregrounds. This operation of BZ enables to applied the image to recognition system for OCR techniques etc. after this process the removing noise process will apply to the image the results from recognition techniques that helps in distinguish the character and letters from content of image picture. The last procedure is used to remove the noise that contain in the background of the picture and helps to improve the quality of the image the three processing that applied to image is shown in figure 1.

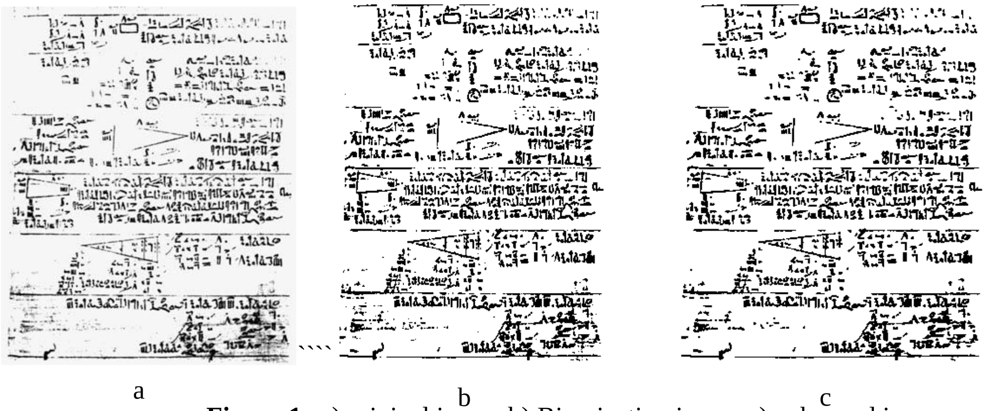


Figure 1:-a) original image b) Binarization image c) enhanced image

Theory and methods:-

The scientific and ancient documents are needed to be available to inherit to our next generation. So to save these documents from ages and maintain it forever, a system is created to preservethese documents systematically. So these documents can be previewed by future student using libraries. The system that created in this paper consists from six stages as shown in figure 2.[5] Since the ancient document that insert to the digital system may contain noise that came from different things such as strains, uneven illumination or big variation in background. So that A digital filter is applied to the image that insert to the system.[6] This filtered will try to remove as possible the noise that located in the background.[7] This helps the second process which consist from the Global thresholding which will applied on the whole picture and try to located the remaining noise that might exist on the background on the picture. The third process will be converting the picture to binary formation. While the fourth stages is detecting the noise in the picture and last process which include the local thresholding[8]is used to remove the noise that exist in the picture and the result of that process will be enhanced picture.

In this paper a hybrid binarizationimplemented or may said as combination of both Repeated Global Thresholding (RGT) and Local thresholding (LT) for improve document. The result of that image will contain small amount of noise or maynot contain depending on the noise that located in the image and final results will be more visually than that in the original image. However, the improve of picture shouldn’t effect on the content of the document and should increase the features for helpingindetects content easily.For detection of the letter neighboring sequence method is used and this area are needed more preprocessing[9].

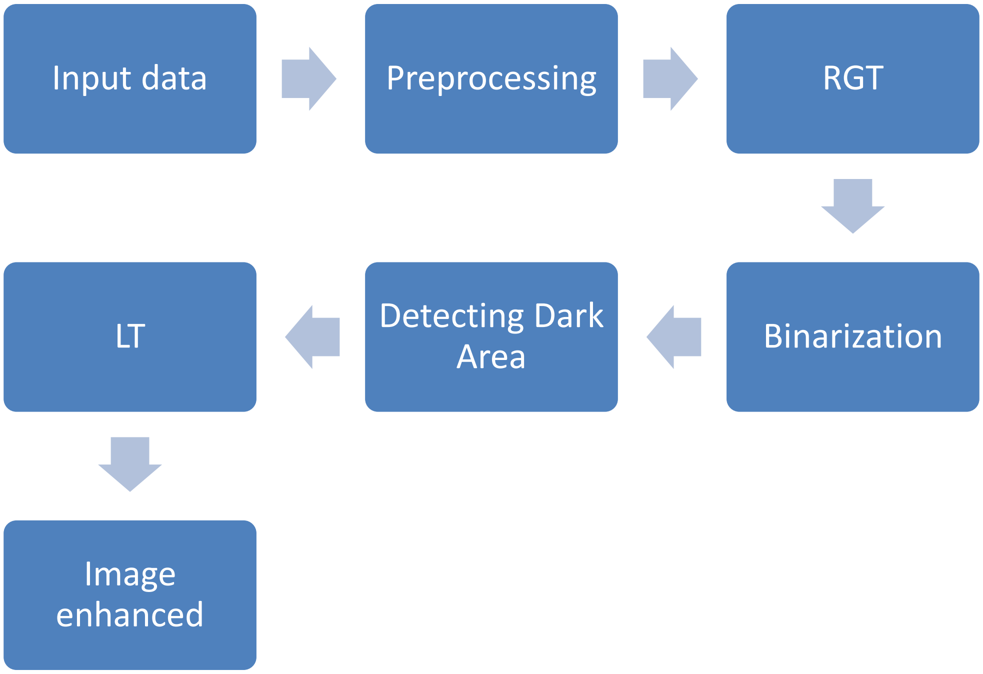


Figure 2:-system procedure of the enhanced image

Proposed methods

The proposed methods are consisting mixture of both RGT and LT together to provide higher picture enhanced. After inserting the scan or captured document image to the system, a preprocessing will apply. The preprocessing consisting from filtering to smooth and remove the noise and to detects the edges of the input image.ThenRGT will check these areas to see whether there isa noise in the background or not and try to remove it. The average of noisy area will be high than the area that contains small amount of noise. The main idea of the algorithm depends on that fact and this isconvincingsince document pictures that contain textual information only.

The RGT methods will segment the image into fixed size m*n. The summation of the black pixel that calculated from each segment and the summation that achieve this equation 1 where F(s) is black pixels frequency and M considered as mean while s mean thestandard deviation of the black pixels in the windows.

$$f(S)>m+ks \qquad (1)$$

Whereas the *f*(*S*) is the frequency of the black pixels in the segment *S* while *m* and *s* are the mean and the standard deviation of the black pixel frequency considering the segments of the entire page, respectively. RGT methods repeated several time and the iteration will stop according to the equation 1 achieved or the number of iteration exceed certain level that chosen for RGT process. RGT will make image that contains only black and white pixel which leads to BZ. Then LT is used to detect the noise in this document image. The mixture of the RGT and LT are enhanced original picture.

A row by row is used for detecting the textural that located inside the document image. The method will use the results of LT and connecting neighboring pixels in respect to their original position in the document image from the selected segments. The final results will be an enhanced image that can be visually easily to read and sight.

Simulation and results:-

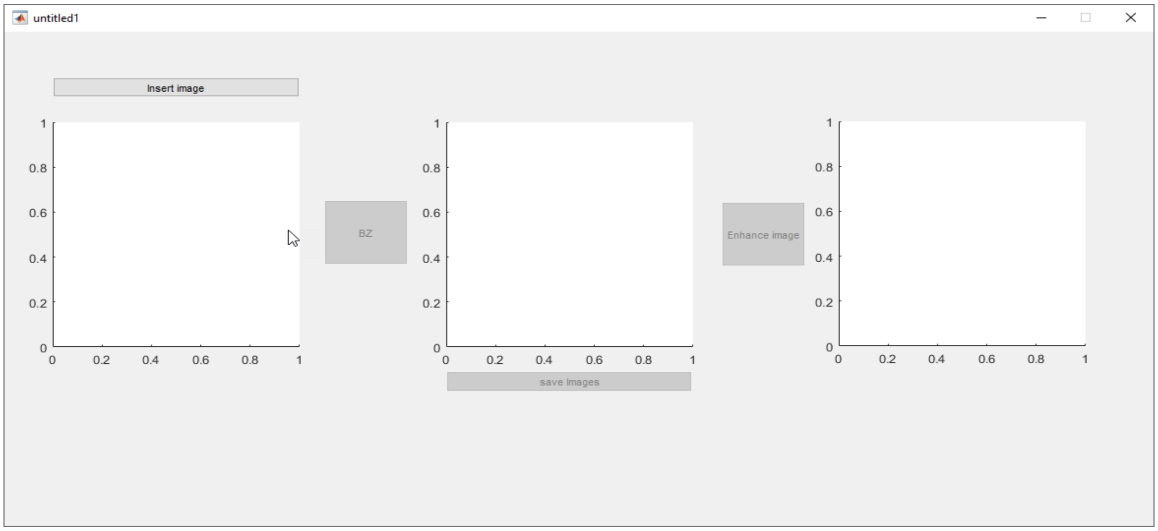


Figure 3:-Start of the program

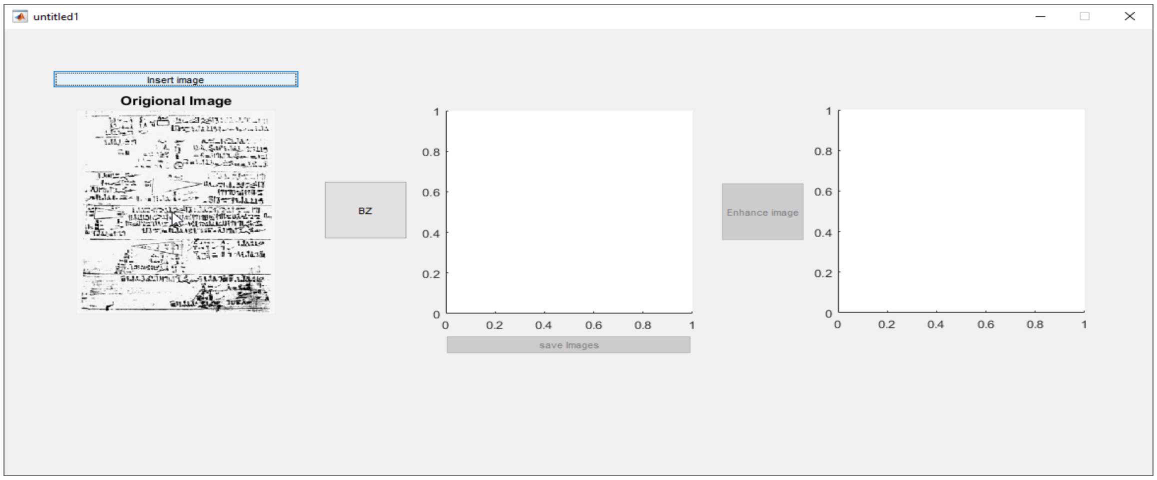


Figure 4:-original image.

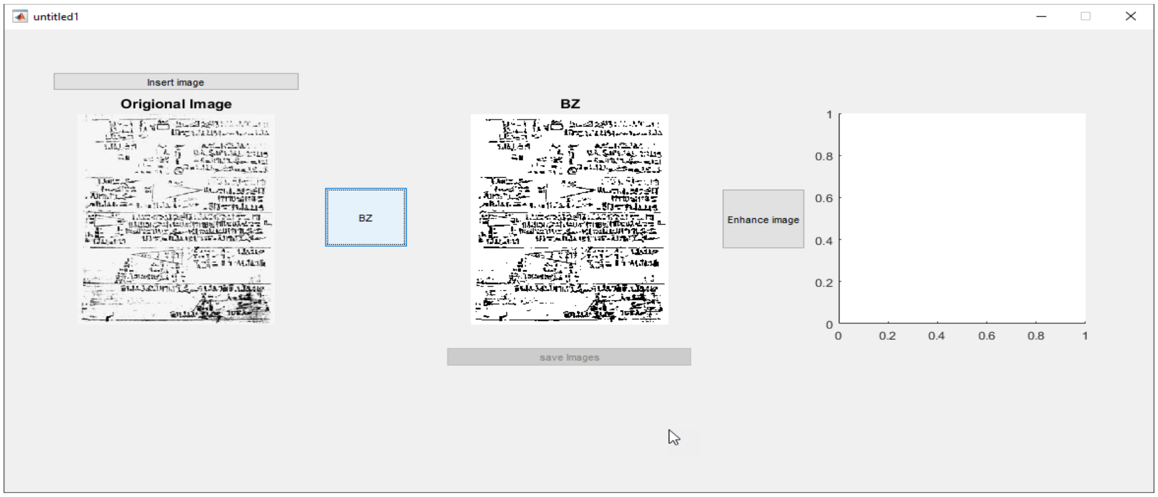


Figure 5:-Applying binarizationmethods

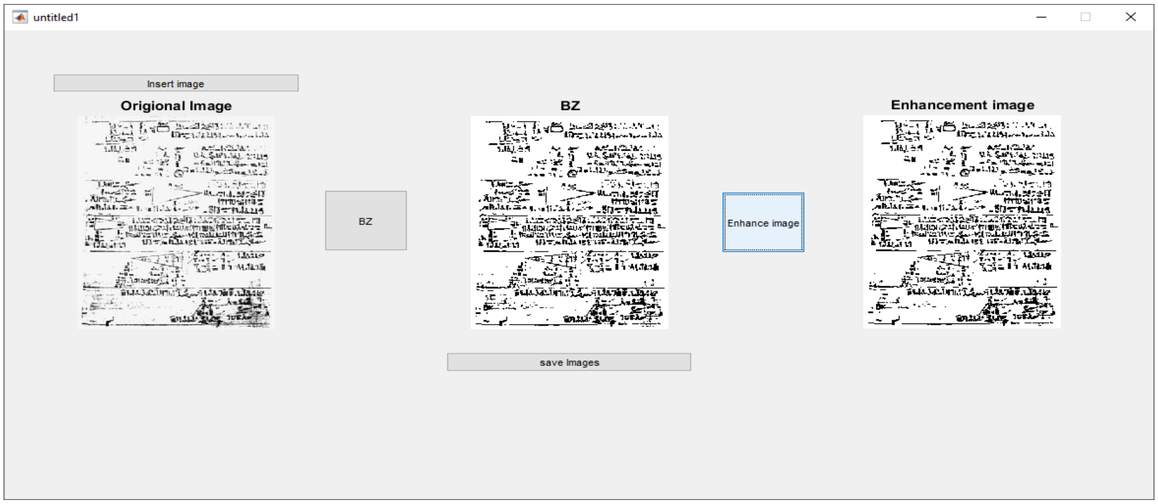
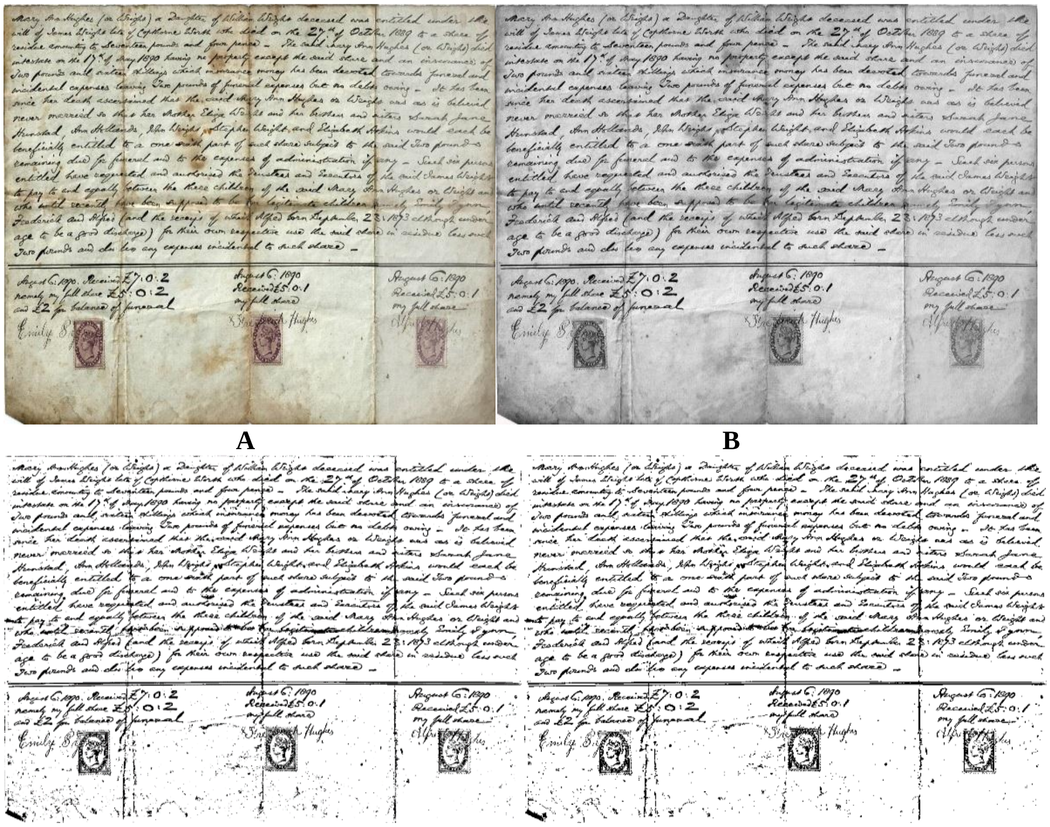


Figure 6:- remove the noise from picture


A

B

C

D

Figure 7:-a) original image b) gray scale image c) binarizationimage d) After remove noise

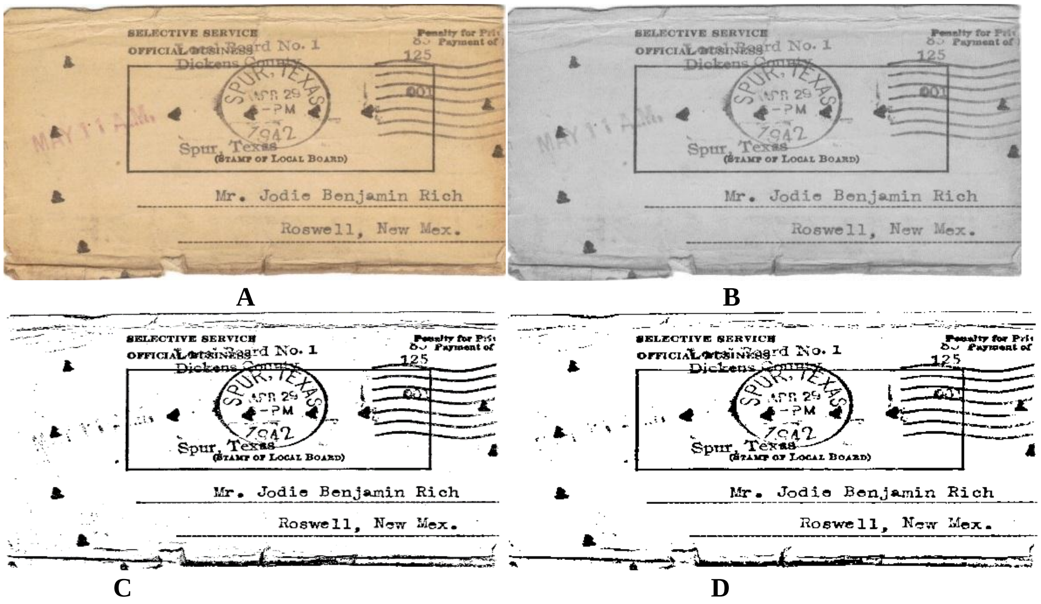


Figure 8:- a) original image b) gray scale image c) binarizationimage d) After remove noise

Conclusion:-

In this paper, a mixture of RGT and LT methods help in removing noise from the background of document image. The proposed methods enable system to deal with hard situation noise and without changing the content of data that located inside the document image. The system helps the libraries and museum and research center to provide their precious document to the world without frighteningof losing the data inside of it.

References:-

1. Rachmawati, Yahya Sitti, S. N. H. Sheikh Abdullah, and K. Omar, "Review on Image Enhancement Methods of Old Manuscript with Damaged Background," in International Conference on Electrical Engineering and Informatics, Selangor, Malaysia, 2009, pp. 62-67.
2. Muhammad Hanif , Anna Tonazzini, Pasquale Savino, and Emanuele Salerno, "Sparse Representation Based Inpainting for the Restoration of Document Images Affected by Bleed-Through†," International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), pp. 1-8, 2018.
3. Atena Farahmand, Abdolhossein Sarrafzadeh, and Jamshid Shanbehzadeh, "Noise removal and binarization of scanned document images using clustering of features," in Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 2017 , pp. 410-414.
4. Rozaida Ghazali, Mustafa Mat Deris, and Nazri Mohd Nawi, Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining, 1st ed., Jemal H. Abawajy, Ed. Johor, Malaysia: Springer International Publishing, 2018.
5. Kittipop Peuwnuan, Kuntpong Woraratpanya, and Kitsuchart Pasupa, "Local variance image-based for scene text binarization under illumination effects," in International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 2016, pp. 798-802.
6. E. Zemouri, Y. Chibani, and Y. Brik, "Enhancement of Historical Document Images by Combining Global and Local Binarization Technique," International Journal of Information and Electronics Engineering, vol. 1, no. 1, pp. 1-5, january 2014.
7. Jian Pan, XinhuaYang, HuafengCai, and BingxianMub, "Imagenoisesmoothingusingamodified Kalman filter," Neurocomputing, vol. 174, no. 3, pp. 1625-1629, january 2015.
8. Senthilkumaran N and Vaithegi S, "IMAGE SEGMENTATION BY USING THRESHOLDING," Computer Science & Engineering: An International Journal (CSEIJ), vol. 1, no. 1, pp. 1-13, Fabuary 2016.
9. Ederson Marcos Sgarbi, Wellington Aparecido Della Mura, and Nikolas Moya, "Restoration of old document images using different color spaces restoration of old document images," in International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 2014, pp. 1-7.