Implementation of Document Image Binarization for Removing Noise from Degraded Documents

[1]Vaishali Patil, [2]Dr. Tripti Arjariya

[1]Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India. [1]vaishupatil43@gmail.com, [2]tripti.beri@gmail.com

Abstract— In this digital world, most hardcopy documents are being convert into digital formats. In the process of conversion, large number of documents are stored and protect through electronic scanning. These documents are available from various sources such as dated documentation, old legal records, Health records, music Sheets, palm leaf, and reports on preservation-related issues. In particular, dated and historical documents are hard to read due to their degradation in terms of low disparity and existence of corrupted artefacts.In recent, corrupted reports binarization has been studied widely and several approaches were developed to deal with issues and challenges in document binarization. Reports image binarization is usually performed in the pre-processing stage of different document image processing related applications such as optical character recognition (OCR) and document image retrieval.It translate a gray-scale document image into a binary document image and accordingly opportunities the ensuing tasks such as document skew estimation and document layout analysis. To create the images, the original source image is decomposed into wavelet subband. Then, the original image is approximated by each subband separately, and nally, the multichannel image is constituted by arranging the original source image (grayscale image) as the rest methods and the approximated image by each sub band as the remaining channels. Further argument are made on the effectiveness and robustness of existing methods, and there is still a scope to develop a hybrid approach that can deal with degraded document binarization more effectively. This reports also reviews noises that might appear in scanned document images and discusses some noise removal methods. In addition to appraise binarization algorithms, we consider the available social datasets and appraise metrics, including those that require pixel-level ground truth and those that do not. We assume with suggestion for future work. Otsu's algorithm is one of the most well-known methods for instinctive image thresholding.

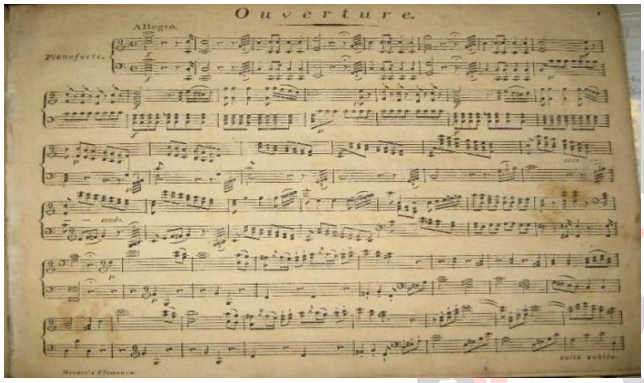
Keywords—Binarization, OCR, Otsu's Algorithm, Document Degradation, Image Quality, Accuracy

I. INTRODUCTION

The Study of image processing is a stimulating topic and has various application in several fields. One among the appliance is document image analysis. Presently most of the day-to-day activities depend upon the changes happening within the technological world, since the mode of data interchange has undergone drastic changes thanks to digital technique. just in case of data authenticity, documentation and its format also as security, storage and retrieval are the important criteria which are taken under consideration while digitizing. Document image binarization refers to the converting o gray scale image into a binary image. it's the earliest step of most document image analysis and understanding systems. Here, we've taken sample images from Document Image Binarization Contest dataset images. we've done contrast stretching, histogram equalization, noise filtering, Laplacian transformation, global and native thresholding methods to get rid of show-through noise, uneven illumination noise and shot noise from degraded document images using OpenCV open source software. Further, performance metrics were estimated to know the efficiency of the above methods in removing the aforementioned noises. Binarization in document analysis aims to differentiate text as foreground pixels from the background. This task is one among the preprocessing steps that features a significant impact on steps like feature extraction and recognition from document images that need a high-quality and accurate foreground. When a handwritten document is scanned, the standard and curvature of the handwriting pose serious problems. Nevertheless, historical documents provide a further challenge which will cause degradation. The degradation are often increased by smudges, stains, bleed-through, and disappearance of the letters in considerable segments of a document. As an important technique for computer vision related applications, image segmentation has been studied

 DOI : 10.35291/2454-9150.2021.0386

for many years. Despite the event of the many complex segmentation algorithms including deep learning-based methods, automatic thresholding remains widely adopted and keeps evolving thanks to its simplicity and effectiveness. To date, numerous automatic thresholding algorithms are proposed which may be further categorized into either local thresholding or global thresholding. Local thresholding tries to hunt multiple threshold values supported localized gray level information, while global thresholding calculates the edge value only using global information to form it simpler and more efficient. There are many adaptive thresholding methods like Otsu's method, Kapur's method, and entropy based method. Otsu's method is one among the foremost well-known and useful global thresholding algorithms proposed by Otsu in1979. it's thus far still widely utilized in many applications including document image binarization, medical image processing, bioscience, and combating infectious diseases like corona virus disease (COVID-19).



(a)Music Sheet

(b) Historical Document

Figure 1. Source sample of different degraded documents; (a) music sheet, (b) historical document

Although these tasks are nearby an equivalent, these are an equivalent challenges in historical paper binarization which makes it a troublesome task for accurate binarization. Thus, the most goal of this paper is to review the prevailing algorithms developed for degraded historical paper binarization. before understanding and review of the degraded binarization techniques, it's extremely important to know the character of defects and degradations in historical paper.

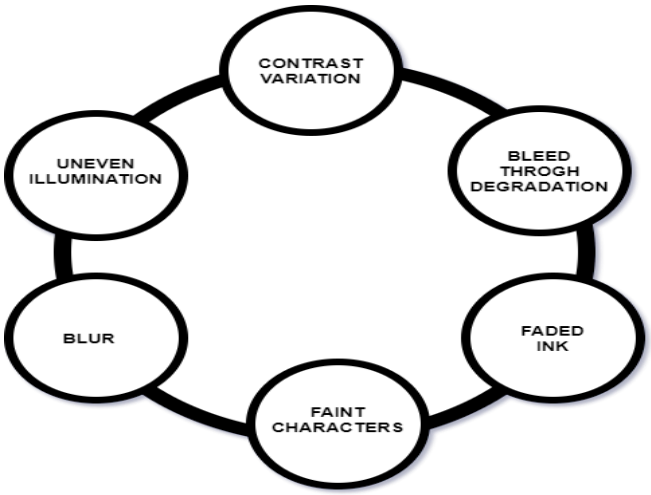


Figure 2. Frequently seen degraded defects in historical documents

II. RELATED WORK

In Otsu [1979] Optimal Threshold is of course selected supported global property but not on local property. It are often used for picture segmentation in order that maximum. Separation are often done between resultant classes of gray levels. Automatic Optimal Thresholding Selection technique are often used. [1]

Kapur & Niblack [1985\1986] By using this algorithm there are two possibilities to separate the histogram of the image. Out of which one can define object and another define background. Maximum Entropy Algorithm are often used.[2]

Solihin Y. and C.H. Leedham [1999] This procedure is additionally called Integral Ratio. it's supported two levelthresholding approach during which each pixel of handwritten image are often divided into three parts: foreground, background and fuzzy area between them. we will decide whether a pixel lies in foreground or background on the idea of Native IR and Quadratic IR. Histogram Based Global Thresholding technique are often used.[3]

Yang and Yan [2000] It use Logical thresholding method for binarization of seriously degraded complex background gray scale images. It cannot affect useful information. Logical Adaptive Thresholding technique are often used.[4]

Randolph [2001] It are often used for enhancement in Fax documents. A directional filter bank has been used which is capable for smoothing of edges and contours. Binary Domain Approach technique are often used.[5]

Wu et al [2003] In first stage global thresholding technique used. In second stage refinement of threshold value are often done. it's used for both simple and sophisticated images that have different shading like postal envelopes. Multi-stage Global Thresholding technique are often used. [6]

 DOI : 10.35291/2454-9150.2021.0386

N. Habibunnisha, K. Sivamani [2019] In this work, done image enhancement techniques to reduce the noises from degraded document images. Here, we have taken sample images from Document Image Binarization Contest (DIBCO) dataset images. We have done contrast stretching, histogram equalization, noise filtering, Laplacian transformation, global and local thresholding methods to remove show-through noise, uneven illumination noise and shot noise from degraded document images using OpenCV.open source software. Further, performance metrics were estimated to understand the efficiency of the above methods in removing the aforementioned noises.[15]

No.	TITLE	AUTHOR	METHOD	DESCRIPTION
1	Reduction Of Noises From Degraded Document Images Using Image Enhancement Techniques	N. Habibunnisha, K. Sivamani , R. Seetharaman, D. Nedumaran	Contrast Enhancement, Histogram Equalization, Median Filtering,	In this work,taken sample images from Document Image Binarization Contest (DIBCO) dataset images. We have done contrast stretching, histogram equalization, noise filtering, Laplacian transformation, global and local thresholding methods to remove showthrough noise, uneven illumination noise and shot noise from degraded document images using OpenCV open source software.
2	Binarization of Degraded Document Images Using Convolutional Neural Networks and Wavelet-Based Multichannel Images	Younes Akbari, Somaya Al-Maadeed And Kalthoum Adam	Cutting Edge Methods	This paper To create the images, the original source image is decomposed into waveletsub and finally, the multichannel image is constituted by arranging the original source image (grayscale image) as the first channel and bands. Then, the original image is approximated by each subband separately, the approximated image by each subband as the remaining channels.
3	A threshold selection method from gray-level histograms	N. Otsu	Automatic Optimal Thresholding Selection Technique	Otsu [1979] Optimal Threshold is naturally selected based on global property but not on local property. It can be used for picture segmentation so that maximum separation can be done between resultant classes of gray levels. Automatic Optimal Thresholding Selection technique can be used.
4	An adaptive logical method for binarization of degraded document images Pattern Recognition	Yibing Yang and Hong Yan	Logical Adaptive Thresholding Technique	It use Logical thresholding method for binarization of seriously degraded complex background gray scale images. It cannot affect usefulinformation. Logical Adaptive Thresholding technique can be used.

Table: Literature review

III. METHODOLOGY

Document images often suffer from different types of degradation that renders the document image binarization a challenging task. Existing system presents a document image binarization technique that segments the text from badly degraded document images accurately. Here it estimates a document background surface through an iterative polynomial smoothing procedure. This method is simple, it cannot work properly on degraded document images with a complex document background. To overcome this, we propose a new technique in which an adaptive contrast map is first constructed for a given degraded document image and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document .The algorithm suggest here is present in four steps: 1. Decision-making for locating the vector of parameters of the image to be filtered, 2. Filtering the image employing a bilateral filter, 3. Splitting the image into the RGB components, and performing their binarization employing a method inspired by Otsu's algorithm for every RGB channel, and 4. Choice of which of the RGB components best preserved the document information within the foreground, which is taken into account the ultimate output of the algorithm. Figure3 presents the diagram of the proposed algorithm. The functionality of every block is detailed

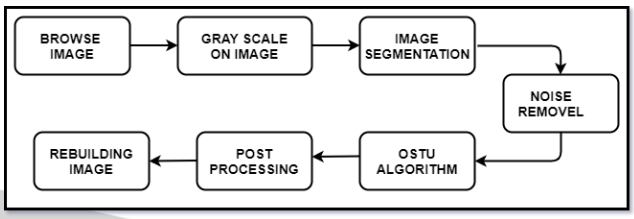


Figure 3. Block diagram of the proposed algorithm

IV. RESULTS AND DISCUSSION

A few experiments are designed to demonstrate the effectiveness and robustness of our proposed method. We first analyze the performance of the proposed technique. In this work, we have reviewed the recent advances in the field of historical document binarization. Based on our analysis, we provide some discussion on the broad trends, the current technical challenges, and directions for future work.

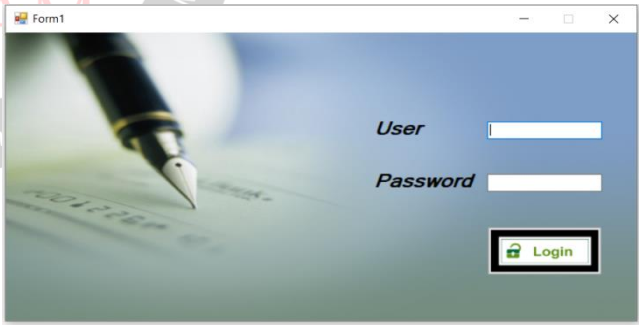


Fig: Login Page

Fig : Historical Document Image Converted Into Grayscale Image

 DOI : 10.35291/2454-9150.2021.0386

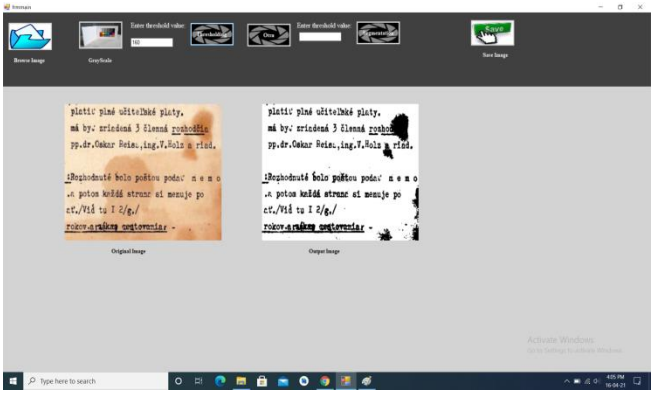


Fig: Thresholding Performed On Historical Document Image

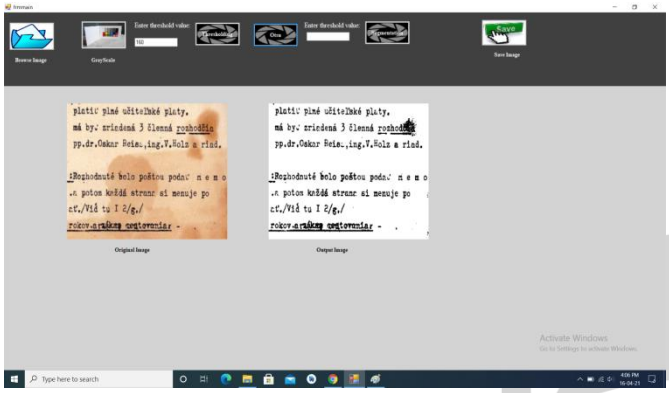


Fig : Otsu's Algorithm Performed On Historical Document Image

V. RESULTS COMPARISON

Title	Author	Method	Thresholding Technique	Threshold Value	Pixel Loss	Ideal Limit Given	Post processing
Reduction Of Noises From Degraded Document Images Using Image Enhancement Techniques	N. Habibunnisha, K. Sivamani , R. Seetharaman, D. Nedumaran	Contrast Enhancement, Histogram Equalization, Median Filtering,	Local	Static	Yes	Not given	Not performed
Binarization of Degraded Document Images Using Convolutional Neural Networks and Wavelet-Based Multichannel Images	Younes Akbari, Somaya Al-Maadeed And Kalthoum Adam	Cutting Edge Methods	Local	Static	Yes	Not given	Not performed
A threshold selection method from gray-level histograms	N. Otsu	Auto-matic Optimal Thresholding Selection Technique	Local	Static	Yes	Not given	Not performed
Implementation of document image binarization for removing noise from degraded documents	Our Proposed System	Otsu Binarization	Global	Dynamic	Less as Compared To above system	Here we mentioned ideal threshold value limit for binarization	Here we performed post processing after segmentation

VI. CONCLUSION AND FUTURE SCOPE

This paper presents a picture Binarization technique for degraded images by using Otsu's method. The proposed method is straightforward, more reliable and an efficient way. The proposed method makes use of morphological operators then Otsu thresholding. we've proposed a strong Otsu's method during this paper. during this review, different document degradation issues are discussed. These issues are related to ancient and historical documents, either in handwritten or printed form. The existence of degradation in handwritten and printed documents exhibits several challenges in developing an accurate and robust method for document binarization, which ultimately aims at improving the standard of various systems. during this review, different binarization techniques are discussed that are recently on trend. There are numerous binarization methods available which will work well with a specific sort of degradation, but a binarization method which will handle any sort of degradation remains left for future work. Binarization is a crucial step towards the event of a system for document image recognition and it's a wider application within the current era during which everything is moving towards digitization. Although several efforts are made within the past to satisfy the most objective of binarization that involves separation of text from the degraded document background, there's a compelling got to develop a quick and accurate method of binarization followed by the system

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Systems, Man, and Cybernetics, vol. 9, pp. 62-66, 1979.
[2] J., P.K. Sahoo, and A.K.C. Wong Kapur, "A new method for gray-level picture. Thresholding using the Entropy of the Histogram," Computer Vision Graphics and Image Processing, vol. 29, pp. 273-285, 1985.
[3] Y. and C.G. Leedham Solihin, "Integral Ratio: A New Class of Global Thresholding Techniques for Handwriting images," IEEE Trans. on PAMI, vol. 21, pp
[4] Yibing Yang and Hong Yan, "An adaptive logical method for binarization of degraded document images Pattern Recognition," Pattern Recognition Society, Elsevier Science, vol. 33, no. 5, pp. 787-807, 2000.
[5] T. Smith, M Randolph, "Enhancement of fax documents using a binary angular representation," in Proceedings of, Int Symp on Intelligent Multimedia, Video and Speech Processing, , Hong Kong, China, 2001.
[6] Adnan Amin, Wu Sue, "Automatic Thresholding of Gray-level Using Multi-Stage Approach," in Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE, 2003
[7] Pratikakis, S. Perantonis Gatos B, "An Adaptive Binarization Technique for Low Quality Historical Documents," Document Analysis Systems VI, vol. 3163, 2004.
[8] Chen Y. and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," IEEE Proc.-Vis. Image Signal Process., vol. 152, December 2005.
[9] Chen Y. and G. Leedham, "Document binarization using Kohonen," IET Image Process., pp. 67-85, 2007.
[10] Basilios, Ioannis Pratikakis, and Stavros J. Perantonis. Gatos, "Improved document image binarization by using a combination of multiple binarization techniques and adapted

 DOI : 10.35291/2454-9150.2021.0386

edge information," Pattern Recognition, vol. ICPR 2008. 19th International Conference on. IEEE, 2008, 2008.

[11] Basilios, Konstantinos Ntirogiannis, and Ioannis Pratikakis Gatos, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," ICDAR, vol. 9, 2009. [16] Yi, Michael S. Brown, and Dong Xu. Huang, "User assisted ink-bleed reduction," Image Processing, IEEE Transactions, pp. 2646-2658, Oct. (2010).

[12] Bolan, Shijian Lu, and Chew Lim Tan Su, "Robust document image binarization technique for degraded document images." Image Processing, IEEE Transactions on 22.4, pp. 1408-1417, 2013.

[13] Cindy M., et al. Goral, "Modeling the interaction of light between diffuse surfaces," ACM SIGGRAPH Computer Graphics, vol. 18, 1984.

[14] Michael, Dick de Ridder, and Heinz Handels Egmont Petersen, "Image processing with neural networks—a review," Pattern recognition 35.10, pp. 2279-2301, 2002.

[15] N. Habibunnisha, K. Sivamani , R. Seetharaman, D. Nedumaran, “Reduction Of Noises From Degraded Document Images Using Image Enhancement Techniques ” , International Conference On Inventive Systems And Control (Icisc 2019) Ieee Xplore Part Number: Cfp19j06-Art; Isbn: 978-1-5386-3950-4.

 DOI : 10.35291/2454-9150.2021.0386