International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Improving Document Image Readability Using Enhanced Thresholding Techniques

Sakshi Sanjay Jadhav[1], Akshada Ramdas Warungase[2], Smita Jaywantrao Ghorpade[3]

[1,2]Student, Department of Computer Science, K. R. T. Arts, B.H. Commerce, A.M. Science (KTHM) College, Nashik, Affiliated to SPPU, Maharashtra, India
[3]Assistant Professor, Department of Computer Science, K. R. T. Arts, B.H. Commerce, A.M. Science (KTHM) College, Nashik, Affiliated to SPPU, Maharashtra, India

Abstract

Document image binarization presents significant challenges, particularly when it comes to isolating text in aged or deteriorated documents. This process is a crucial initial phase in Optical Character Recognition (OCR), aiming to convert the document into a binary format—distinguishing the text (foreground) from the background. This conversion facilitates easier reading and processing by computers. Given the wide variability in document quality, researchers have created numerous techniques to address various forms of damage and degradation. This research reviews core binarization techniques and sheds light on various challenges and unresolved problems in the domain of document image binarization. Through this analysis, we assess various binarization techniques to identify their limitations and propose potential avenues for enhancing future research in this field.

Keywords: degraded document images, binarization, threshold processing

1. INTRODUCTION

Document image binarization is a crucial initial step in document analysis. Its primary objective is to distinguish the text (foreground) from the background, which is essential for enhancing readability and ensuring accurate subsequent processing, such as optical character recognition (OCR) and information retrieval. The effectiveness of a rapid and precise binarization method is crucial for the performance of later stages in the document processing workflow. Although researchers have dedicated significant effort to this issue over the years, the task of binarizing degraded documents continues to pose challenges. This difficulty arises from the considerable variability in background noise, lighting conditions, ink deterioration, and other factors that affect image quality. Consequently, determining an appropriate threshold for binarization becomes particularly complex, as the contrast between the text and background can be inconsistent across different images. In our approach, we enhance the detection of text edges by integrating a grayscale image with an innovative image segmentation technique, which facilitates the generation of a clear binarized output.

In this research, we investigate a binarization technique that initiates with the conversion of images to grayscale, followed by the implementation of two thresholding methods: global thresholding utilizing Otsu's method and local thresholding employing Niblack's method. While global approaches are

proficient for documents characterized by clean and uniform backgrounds, they frequently struggle with more intricate scenarios. Conversely, local thresholding methods are designed to accommodate pixel-level fluctuations, rendering them more effective for images that are degraded or contain noise. This paper seeks to evaluate the efficacy of these binarization techniques and processing methods, while also emphasizing the advantages and drawbacks of each in addressing various forms of degraded documents.

1.1 Problem Statement

Document image binarization serves as a crucial initial phase in the analysis of documents. The objective is to effectively extract the text from the non-textual background. An efficient and precise binarization technique enhances the quality of subsequent processing stages. Despite extensive research over the years, the challenge of binarizing degraded document images persists. This difficulty arises from the significant variations in appearance between the text and background in different images, complicating the selection of an appropriate threshold. To address this issue, we utilize edge detection of the text by employing a combination of a grayscale image and an innovative image segmentation approach, which facilitates the generation of a clear and well-defined binarized image.

1.2 Research Objectives

1. To develop an effective binarization method that can accurately separate text from the background in degraded document images.
2. To improve the quality of document images for better readability and more accurate OCR (Optical Character Recognition) results.
3. To compare different binarization techniques and identify their strengths and weaknesses.
4. To implement serial and parallel processing approaches based on the size of the document image for faster execution.
5. To provide a reliable solution that can handle different types of degradation, such as stains, smudges, and uneven lighting.

2. Literature Review

Through binarization, grayscale images are transformed into binary black-and-white images, enabling the extraction of relevant content like text from the surrounding background—an essential step in image segmentation. Thresholding is a widely used tool in image segmentation, including global, local, and various automatic thresholding methods. In this paper, we have summarized some challenges and difficulties in the field of document image binarization [1].

In our work, we propose a method that combines different thresholding approaches through a classification framework to achieve improved binarization results. Different binarization approaches demonstrate varying levels of effectiveness depending on the characteristics of the document images. A more effective performance can result from the combination of different binarization approaches, with careful evaluation [2].

Binarization is a significant advance towards the improvement of a framework for archive picture acknowledgment and it has a more extensive application in the current time wherein everything is moving towards digitization [3]. The process of binarization is a key technique in image analysis, often used in applications like scene text detection and the assessment of medical images. Binarization serves as a fundamental technique in digital image processing, particularly within document image processing, with wide-ranging applications such as text recognition and document segmentation, image morphological processing and feature extraction. This paper provides a comprehensive review and integration of widely-used document image binarization methods within an open research framework [4]. The core objective is to evaluate these techniques and highlight their limitations. Document binarization is typically execute in the preprocessing phase of several document image processing associated fields such as optical character recognition (OCR) and document vision retrieval. The binarization technique involves converting an image from a grayscale format, with up to 256 levels of brightness, to a purely black-and-white image [5].

3. Methodology

3.1 Description of Research Design

Binarization represents a significant area of ongoing research within the field of Document Image Processing. This process transforms a grayscale image into a binary image. The binarization of document images is a crucial step in the pre-processing of scanned documents, aiming to preserve as many subcomponents as possible, including text, background, and images. In the 1960s, studies on threshold segmentation predominantly concentrated on global thresholding, local thresholding, and adaptive thresholding techniques. Nevertheless, these methods faced challenges in effectively managing complex images characterized by uneven pixel distributions and noise interference. This paper aims to assess the limitations inherent in algorithms for binarizing images with degradation.

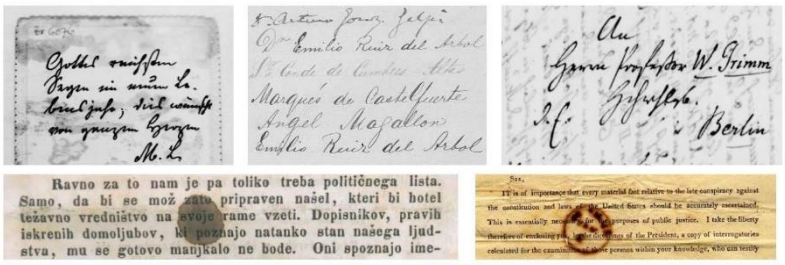


Figure 1. Example of degraded document images from DIBCO [4].

3.2 Data Collection

For this study, the datasets employed to assess and compare various document image binarization methods are derived from earlier research publications, specifically the DIBCO (Document Image Binarization Contest) and HDIBCO 2016 (Historical Document Image Binarization Contest 2016) datasets. The DIBCO dataset includes images of deteriorated documents, presenting challenges such as noise, low contrast, and inconsistent lighting, which are frequently encountered in practical scanning scenarios. The research community widely adopts this dataset for evaluating and comparing binarization algorithms. Conversely, the HDIBCO 2016 dataset concentrates on historical document images, which introduce

further complexities due to faded text, elaborate background patterns, and distortions characteristic of older documents. By leveraging these datasets from prior research, this study seeks to comprehensively evaluate the proposed binarization techniques across a spectrum of document types, ranging from contemporary to historically degraded. This methodology facilitates meaningful comparisons with existing approaches and highlights the efficacy of the proposed technique.

3.3 Data Analysis

The examination was conducted on grayscale images derived from the conversion of original color documents into various shades of gray. Subsequently, two binarization methods—global and local—were employed to evaluate their effectiveness in the binarization of document images. In the case of global binarization, a uniform threshold was applied throughout the entire image to distinguish the foreground text from the background. Conversely, local binarization utilized varying thresholds for smaller segments of the image, offering greater adaptability in managing documents with inconsistent lighting. The overall findings revealed that while global binarization was quicker and more straightforward, local binarization yielded better results in terms of accuracy and visual quality, particularly for documents exhibiting considerable degradation or uneven illumination. The balance between speed and accuracy emerged as a crucial factor in assessing the performance of both techniques.

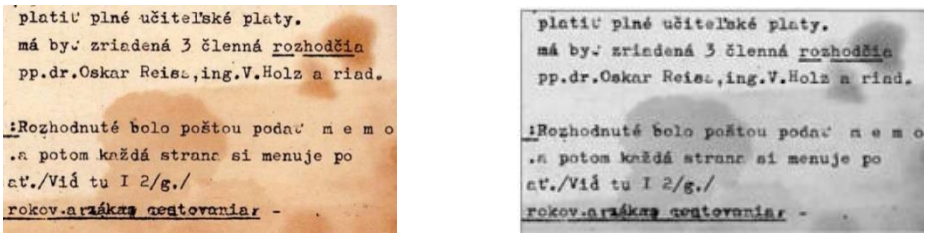
4. Implementation

4.1 Pre-processing

Pre-processing plays a crucial role in the binarization of document images by improving the quality of the input image prior to the application of the binarization algorithm. This process includes tasks such as resizing, eliminating noise, and converting to grayscale. These actions contribute to minimizing distortions like blurriness, inconsistent lighting, and smudges, thereby enhancing the clarity of text regions and increasing the precision of the binarization process.

4.2 Grayscale Conversion

Grayscale conversion is a crucial step prior to binarization, as it simplifies the image by removing color and allowing the algorithm to concentrate on variations in intensity. Binarization classifies pixels into black or white based on their intensity levels; therefore, utilizing a single-channel grayscale image enhances the accuracy of thresholding and minimizes processing time. Additionally, this approach is more effective in managing fluctuations in lighting and background noise. Most binarization algorithms are optimized for grayscale images, necessitating the conversion of RGB images to grayscale beforehand. While straightforward conversion methods, such as the luminosity formula ($g = 0.21r + 0.72g + 0.07b$), are frequently employed, they may inadvertently diminish the contrast of coloured text in the foreground. This problem is particularly pronounced when the image features multiple ink colours in the foreground.



(a) Degraded document (b) Grayscale document

4.3 Traditional Binarization techniques

4.3.1 Global threshold method

The Otsu algorithm, developed in 1979, is a prominent method of a global thresholding technique. This algorithm seeks to determine the most effective threshold value (T) by examining the grayscale histogram of an image. It segments the image into foreground and background regions, striving to reduce variation within each segment while enhancing the contrast between them. A pronounced difference in grayscale values between these segments signifies a more accurate threshold and better image segmentation. The Otsu algorithm is also commonly known as the maximized difference between classes method [4]. The optimal threshold for a given image is defined as the value that produces the maximum inter-class separation, represented by the following expression:

$$T' = \arg\max_{0 \le T \le L} \omega_0(T)\omega_1(T)(\mu_0(T)-\mu_1(T))^2 \quad (1)$$

we represent the image pixel in the grey level of the image, the image has L-order grey level, $\omega_0(T)$ and $\omega_1(T)$ are the probability distribution of the target and background when the threshold value is T, $\mu_0(T)$ and $\mu_1(T)$ represent the average grey value of the pixel of the target and background, respectively, if the pixel value of the input image is greater than T'. The pixel value is set to white, or otherwise it is black [1].

Otsu's method uses a global threshold to segment the entire image in a single pass, selecting the threshold that best separates the image into two classes.

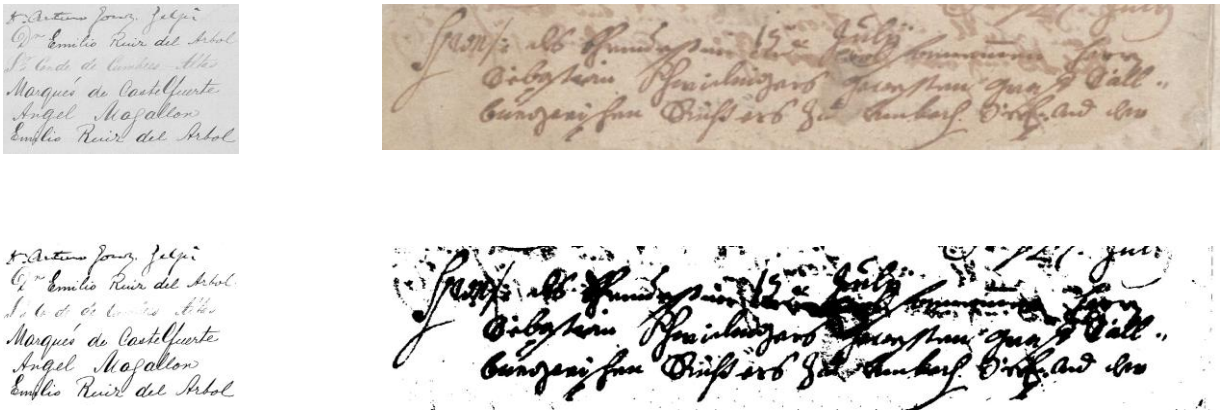


Figure 2. Original images from HDIBCO2016 dataset and binarization results of Otsu's [1].

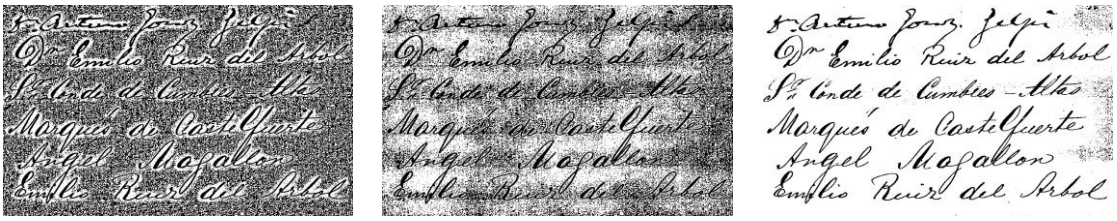
4.3.2 Local threshold method

The Niblack algorithm, on the other hand, addresses the drawbacks of global thresholding by applying a local approach. It computes a threshold for each pixel based on the local mean and standard deviation within a defined window, enabling adaptive binarization across the image. The threshold calculation formula is expressed as follows:

$T = m + k * s$ (2)

In this context, m represents the local mean grayscale value, s is the local standard deviation, and k is a configurable constant that adjusts based on the image's foreground and background properties [4].

Among local binarization methods, Niblack's algorithm is notably effective in handling grayscale images with low contrast, noise, and uneven backgrounds [1].



(a) (b) (c)

Figure 3. Niblack's binarization results [1].

Classification	Algorithm	Description	Performance
Global Threshold	Otsu [1]	The gray level corresponding to the maximum inter-class variance is selected as the global threshold.	Low complexity and fast operation. However, it cannot handle complex degraded images and is suitable for processing high-quality document images.
Local Threshold	Niblack [2]	It calculates the mean and standard deviation of the pixel within a local window of an image, laying the foundation for local binarization methods.	The processing time has increased, and obvious noise can be seen in the output binarization image, which greatly increases the foreground region.

Table 1. Traditional binarization techniques (1).

5. Conclusion

This paper examines the technique of document binarization for degraded images, a significant aspect of visual processing. The primary aim is to assess the shortcomings of existing algorithms used for binarizing degraded images. The findings indicate that each method possesses distinct advantages and disadvantages, and no single technique is universally applicable to all scenarios. Furthermore, there is no all-encompassing binarization approach that is effective for every type of document image. The paper provides an overview of two specific binarization techniques. The evaluation reveals that traditional algorithms perform adequately on images with clear or uncomplicated backgrounds but struggle with documents that feature intricate backgrounds, particularly those affected by mixed noise or severe degradation. Given the variations in characters and numerals across different languages, future research should focus on enhancing the cross-language adaptability of binarization techniques to better accommodate diverse writing styles and stroke patterns. In the near future, we plan to introduce a new algorithm that will employ a more reliable methodology to improve the effectiveness of this process.

References

1. Yang, Z.; Zuo, S.; Zhou, Y.; He, J.; Shi, J. A Review of Document Binarization: Main Techniques, New Challenges, and Trends. *Electronics*2024, *13*, 1394. https://doi.org/10.3390/electronics13071394
2. Bolan Su, Shijian Lu, and Chew Lim Tan. Integration of Document Image Binarization Techniques. National University of Singapore. September 2011.
3. Rohithkumar A., Rajath T., Bipin Nair B.J., and Dr. N. Shobha Rani. Document Image Binarization. Journal of Seybold Report. December 2020, 1533-9211.
4. Zhengxian Yang, Rui Zhang, Yanxi Zhou, Jinlong He, Jianwen Shi, and Shikai Zuo. Overview of Document Image Binarization. Xiamen University of Technology, Fujian, China. January 22, (2024)
5. Tarnjot Kaur Gill. A Review of Document Image Binarization Techniques. International Journal of Computer Applications (0975 – 8887), Volume 98, No. 12, July 2014.
6. Chris Tensmeyer and Tony Martinez. A Review of Historical Document Image Binarization. © Springer Nature Singapore Pte Ltd., May 16, 2020. https://doi.org/10.1007/s42979-020-00176-1
7. Wei Xiong, Lei Zhou, Ling Yue, Lirong Li, and Song Wang. An Enhanced Binarization Framework for Degraded Historical Document Images. EURASIP Journal on Image and Video Processing (2021) 2021:13; xw@mail.hbut.edu.cn; songwang@cec.sc.edu.