Optimizing OCR Performance: An Investigation into Image Preprocessing Techniques

Amaan Tamboli[1], Prof. Archana Kadam[2], Ayush Talegoankar[3], Abhinav Tekam[4], Bhushan Wakchaure[5]

[1]Amaan Tamboli Computer Engineering, PCCOE
[2]Archana Kadam Computer Engineering, PCCOE
[3]Ayush Talegaonkar Computer Engineering, PCCOE
[4]Abhinav Tekam Computer Engineering, PCCOE
[5]Bhushan Wakchaure Computer Engineering, PCCOE

Abstract - This study investigates the impact of various image processing techniques on the accuracy of text extraction by PaddleOCR. The research explores how inversion, noise reduction, erosion, dilation, and grayscaling can pre-process images to improve the performance of PaddleOCR. Initially, the OCR's effectiveness is assessed using unprocessed images. Subsequently, each image processing method is applied individually to evaluate its contribution to the accuracy of text detection. The findings aim to identify the most beneficial preprocessing technique for enhancing the accuracy of PaddleOCR in text extraction tasks. This research has implications for the optimization of OCR technology in digitizing textual content from diverse image backgrounds.

***Key Words*:** OCR optimization, image preprocessing, paddleOCR, image to text

1.INTRODUCTION

Text extraction from photos is one of the most crucial activities in the digital age, as it allows non-editable material to be converted into an editable and searchable format. Optical character recognition (OCR) technology, which is the basis of this conversion procedure, has gained popularity due to PaddleOCR's robust performance. However, the quality of the input image greatly affects the accuracy of OCR. This study report presents extensive investigation into the use of several image preprocessing techniques to maximize PaddleOCR's text identification capabilities. Using raw images from the investigation, PaddleOCR is initially utilized to establish a baseline accuracy level. Next, it methodically applies a number of image processing techniques, including dilation, inversion, noise reduction, erosion, and grayscaling. The goal is to evaluate how well various preprocessing methods enhance OCR accuracy. Through data analysis, the study seeks to provide a more thorough understanding of how picture quality affects OCR performance as well as the best preprocessing technique. This research enhances the field of image processing by offering valuable insights into the preprocessing steps that can significantly boost the efficacy of text extraction systems such as PaddleOCR.

2.MOTIVATION

It is not only academic curiosity that drives the pursuit of advances in optical character recognition (OCR) technology. It is a commitment to removing the barriers standing in the way of turning visual data into accessible, editable, and searchable text. This study is motivated by the profound social impacts that higher OCR accuracy can have. The implications run deep and wide; they include anything from safeguarding historical documents to providing assistance to the visually impaired.

The drive to become a proficient paddlerThe primary motivation behind OCR's advanced photo preprocessing capabilities is its ability to digitize data in an effortless manner. This work aims not only at enhancing an already-existing technology but also at fully realizing the promise of OCR as a bridge between the analog past and the digital future. By working toward this objective, we are expanding the field of document analysis and creating opportunities for innovations that will benefit future generations. The purpose of motivation is to create, improve, and inspire a society where knowledge is publicly accessible to all.

3.LITERATURE REVIEW

The evolution of Optical Character Recognition (OCR) technology has been pivotal in the digitization of text from images and documents. Chirag Patel, Atul Patel, and Dharmendra Patel [1] conducted a case study on Tesseract, an open-source OCR tool, providing valuable insights into its performance and accuracy. This research is crucial for understanding the capabilities and limitations of Tesseract in character recognition.

In another significant contribution, Mr. Pratik and Mr. Shashank H. Yadav [2] presented their work at an IEEE conference, exploring the methodologies and algorithms behind text recognition from images. Their comprehensive overview of the field offers a deeper understanding of the text recognition process.

Ranjan Jana, Amrita Roy Chowdhury, and Mazharul Islam [3] examined the challenges and solutions in OCR from text images. Their study adds to the knowledge of OCR's applicability and potential for development across various contexts.

Kushan Mehta, Jay Patel, and Nilesh Dubey [4] focused on text extraction from book covers using Maximally Stable Extremal Regions (MSER). Their findings underscore the effectiveness of MSER in extracting text against complex backgrounds, which is essential for the digitization of printed materials.

V. Kumar and colleagues [5] discussed the practical applications of OCR in automating data entry and processing tasks through the extraction of information from bill receipts. Their research highlights the role of OCR in enhancing business operation efficiency.

Malathi et al. [6] provided a comparative analysis of Robotics Process Automation (RPA) with open-source OCR engines, including Microsoft OCR and Google Tesseract OCR. Their work is instrumental in understanding the performance of different OCR engines within the context of RPA.

Lastly, A. Chandio et al. [7] and Asghar Ali Chandio [8] addressed the complex challenge of recognizing cursive text in natural scene images using Deep Convolutional Recurrent Neural Network (CRNN). Their contributions are significant in advancing OCR technology to tackle cursive writing recognition, a notably difficult area within the field.

Collectively, these studies demonstrate the advancements and diverse applications of OCR technology. They also highlight the necessity for ongoing research to improve text recognition accuracy and efficiency, especially in challenging scenarios such as cursive text and complex image backgrounds. The literature indicates a trend towards more sophisticated OCR systems capable of handling a variety of text presentations, which is essential for the continued integration of OCR technology into modern digital workflows.

4.PROPOSED MODEL

A. Image Pre-Processing

1. **Inverted Image:** Image inversion is a fundamental image processing technique where a picture's colors are reversed. It is sometimes referred to as the negative effect. This method effectively converts bright areas of the original image into dark ones and vice versa, creating a photographic negative. For instance, flipping a photo that has a light background and dark writing would result in a light backdrop and dark text.

2. **Noise Removal:** Noise reduction, also known as denoising, is a crucial stage in image processing that eliminates unwanted random variations, or "noise," from an image. The image's clarity and visual quality are enhanced by this process, increasing its suitability for viewing or additional research. When photos are being taken, transmitted, or processed, noise may be produced. It could show up as graininess or speckles, which degrades the image quality.

3. **Dilation:** "Image dilation" is a morphological procedure that enlarges an image's object limits. It is achieved by convolving a structural element with the image, which may have varying sizes and shapes. Dilation is mostly used to make a binary image's white areas (foreground) larger. This helps to highlight specific traits, connect dissimilar components, and fill in little gaps.

4. **Erosion:** Image erosion is a fundamental technique in morphological image processing. Its primary use is to warp or deteriorate the borders of foreground objects in images1. This process works well for filling in broken or damaged object pieces and eliminating little white patches.

5. **Grayscaling:** Grayscaling is the process of converting an image from its original color space—such as RGB, CMYK, HSV, etc.—into shades of gray1. In an 8-bit digital image2, each pixel in a grayscale image represents the brightness of that specific area; this is commonly expressed as a value between 0 (black) and 255 (white). This transformation makes the picture

data less complex, which makes it easier for algorithms to handle and analyze the data.

Python's OpenCV package can be utilized to create picture pre-processing techniques. To do inversion, for example, use the cv2.bitwise_not() function; to minimize noise, use the cv2.medianBlur() function; to dilate, use the cv2.erode() function; and to perform grayscaling, use the cv2.cvtColor() method with the cv2.COLOR_BGR2GRAY argument. The OpenCV library has several methods for carrying out various image pre-processing operations.

B. Text Extraction

After applying the image pre-processing techniques, the resulting image is typically in a format that is more suitable for text recognition. The next step is to use OCR software to extract the text from the image. PaddleOCR is a popular OCR library that can be used to recognize text from images.

The pre-processed image is sent into the ocr.ocr() method, which handles OCR and generates a set of lists with the recognized text, in order to use PaddleOCR. Each sub-list contains the identified text as well as its locations within the image. PaddleOCR also provides other programmable options to allow for even more customization of the OCR process. These include adjusting the detection mode, using unique techniques for recognition, or specifying the text's language.

It's critical to keep in mind that the quality of the pre-processed image greatly influences the quality of the text that is recovered. If the original image is of low quality or the picture pre-processing techniques are not applied correctly, the text extraction accuracy may suffer. It's important to experiment to find the optimal pre-processing settings and methods for a given image.

5.SYSTEM ARCHITECTURE

The system architecture leverages optical character recognition (OCR) and multi-stage image processing to facilitate the extraction of text from user-inputted photos. The four primary parts of the architecture are each responsible for a certain function inside the system as a whole.

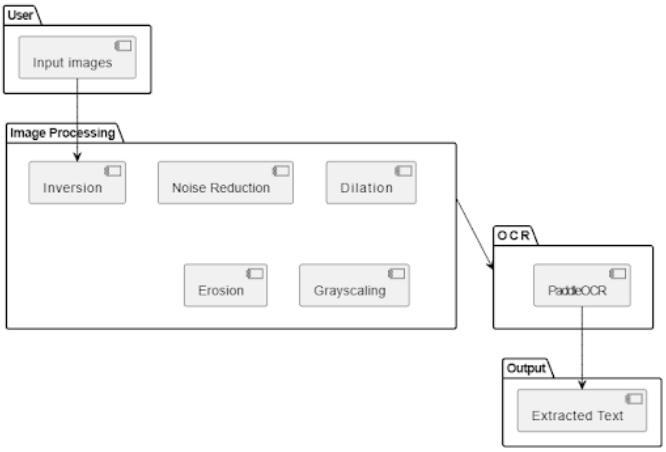

Fig 5.1 System Architecture

1. **User Input Block**: This is the area where users initially enter images into the system. By acting as an interface between the user and the subsequent processing steps, it ensures that the photographs are successfully received and queued for processing.

2. **Image Processing Block**: To enhance the quality of the images and prepare them for text extraction, this block applies a range of image processing techniques. These algorithms include inversion, noise reduction, dilation, erosion and grayscaling, among others, to improve the accuracy of the OCR process.

3. **OCR Block**: Following processing, the images are forwarded to the OCR block, which recognises and extracts text from the pictures using PaddleOCR, the most modern OCR technology. PaddleOCR is well known for its remarkable accuracy and efficiency in converting a wide range of image formats into machine-encoded text.

4. **Output Block**: The extracted text is shown or stored in the output block as the last step. The system's endpoint is this block, which allows users to access the textual information that has been taken from their photos.

6.RESULT

This study used PaddleOCR to investigate how various image processing methods affected OCR (Optical Character Recognition) accuracy. Initially, text was extracted from an uncooked photo using PaddleOCR. Subsequently, every image processing technique was applied separately to the same image.

The confidence score for each sentence in the image was calculated both before and after each algorithm run. This

approach simplified the process of objectively comparing how much each algorithm improved OCR accuracy. The results, which are presented in the table below, give a clear comparison of the confidence scores before and after the application of each image processing technique.

Sentence	Unprocessed	Inversion	Noise Reduction	Dilation	Erosion	Gray scaling
1	0.857	0.891	Not detected	Not detected	Not detected	0.908
2	0.977	0.959	0.967	0.928	0.869	0.986
3	0.915	0.951	0.923	0.939	0.955	0.973
4	0.911	0.988	0.856	0.929	Not detected	0.992
5	0.98	0.977	0.515	0.585	Not detected	0.985
6	0.974	0.961	Not detected	0.695	Not detected	0.974
7	0.957	0.965	Not detected	Not detected	Not detected	0.963
8	0.894	0.913	Not detected	Not detected	Not detected	Not detected
9	0.906	0.94	Not detected	Not detected	Not detected	0.785
10	0.902	0.916	0.551	Not detected	Not detected	0.909
11	0.615	0.837	0.733	0.707	0.758	0.697
12	0.974	0.96	0.644	Not detected	Not detected	0.97
13	Not detected	0.876	Not detected	Not detected	Not detected	Not detected

Table 1: Confidence score matrix for each line of sentences of images.

As shown in Table 1, the Inversion technique consistently improved the confidence scores across most sentences when compared to the unprocessed state. This suggests that inversion can enhance the clarity or detectability of text in images, thereby improving the confidence of subsequent analyses or detections

Fig 6.1: Unprocessed image results using paddleOCR

In Fig1, it appears that PaddleOCR performed well on most lines, with high confidence scores above 0.8. However, it struggled with one line (confidence of 0.615) and failed to detect text on another line. This further emphasizes the challenges in OCR, particularly with distorted or unclear text.

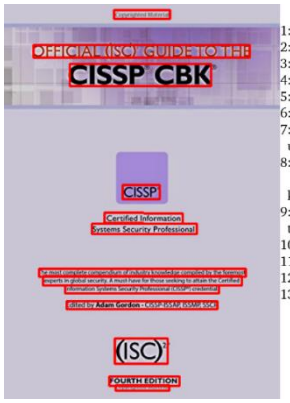

Fig 6.2 : Inverted image results using paddleOCR

In Fig2, it is clear that inverting the image not only increased the confidence scores of the detected text but also enabled the detection of text that was not recognized when the unprocessed image was used.

Moreover, results obtained from grayscaled images were somewhat less precise than those obtained from untreated images.

The erosion, dilation, and noise reduction techniques did not perform as anticipated in this experiment. The results demonstrate that these strategies did not increase the confidence ratings or the legibility of the text in images. Instead, they caused aberrations that significantly degraded the quality of the images, which significantly decreased the efficacy of the subsequent experiments.

Taking everything into account, our results demonstrate the necessity of image processing techniques for accurate text identification. More precisely, our findings suggest that integrating several image processing techniques may greatly improve the accuracy of text identification using paddleOCR engine.

7.FUTURE SCOPE

These three main areas can be the focus of this study's future revisions. The first field is language-specific OCR. Considering the variety of scripts and languages spoken around the world, a significant extension of this work may be to examine how well preprocessing techniques perform for different languages. This could be especially interesting for languages with complex scripts or those written from right to left. By tailoring OCR optimization approaches to specific languages, the accuracy of text extraction could be significantly improved.

The second domain is called the Multi-technique Approach. In this study, each preprocessing method was applied separately. However, subsequent research may look at the outcomes of

combining or ordering many strategies. This could lead to the determination of the optimal combinations.

The final section is Advanced Preprocessing Techniques. While this work focused on basic picture preprocessing techniques, future research may investigate more advanced techniques. Examples of this include adaptive thresholding, skew correction, and region of interest extraction. By looking into these complex techniques, future studies might uncover novel ways to enhance PaddleOCR's capabilities.

By focusing on these areas, future research can develop OCR technology and lead to more accurate and efficient text extraction from pictures. This will not only help academics but also have practical uses in the various industries that employ OCR technology.

8.ACKNOWLEDGEMENT

We like to extend our sincere gratitude to all the people, groups, and software libraries that have helped shape the OCR methods and resources employed in this work. We would like to express our profound gratitude to the OpenCV developers, whose perseverance and hard work produced the crucial image processing libraries needed for our pre-processing methods.

Furthermore, we would like to thank the PaddleOCR community for creating a powerful OCR program that was the foundation of our study. Our project's success has been greatly attributed to PaddleOCR's ability to convert non-editable data into an editable and searchable format.

Finally, we express our gratitude to the scientific community for their continued contributions to the advancement of OCR. Their dedication to provide unrestricted access to their studies has been a priceless asset for this endeavor. Their combined knowledge and insights have improved our comprehension and motivated us to continue pushing the frontier of OCR technology.

9.REFERENCES

1. *Chirag Patel, Atul Patel, Dharmendra Patel "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study" International Journal of Computer Applications (0975 – 8887) Volume 55– No.10, October 2020*

2. *Mr. Pratik, Mr. Shashank H. Yadav "Text Recognition from Images" IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)2019*

3. *Ranjan Jana, Amrita Roy Chowdhury, Mazharul Islam "Optical Character Recognition from Text Image" International Journal of Computer Applications Technology and Research 2021*

4. *Kushan Mehta, Jay Patel, Nilesh Dubey "Text Extraction from Book Cover Using MSER" Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM) 14 June 2019*

5. *Kumar, V., Kaware, P., Singh, P., Sonkusare, R., & Kumar, S. (2020). "Extraction of Information from Bill Receipts Using Optical Character Recognition" 2020 International Conference on Smart Electronics and Communication (ICOSEC)*

6. *Malathi, T., Selvamuthukumaran, D., Diwaan Chandar, C. S., Niranjan, V., & Swashthika, A. K. (2021). "An Experimental Performance Analysis on Robotics Process Automation (RPA) with Open Source OCR Engines: Microsoft OCR and Google Tesseract OCR" IOP Conference Series: Materials Science and Engineering*

7. *A. Chandio et al.: "Cursive Text Recognition in Natural Scene Images Using Deep CRNN" IEEE 2022*

8. *Asghar Ali Chandio "Cursive Character Recognition in Natural Scene Images" IEEE 2020*