IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205292 |

Digital Resurrection: Restoring Fragile Documents with OCR

Aniket Rawat, Akshay Pawar, Shivam Kudal, Chirag Fulfagar, Dr (Mrs) S.P. Deore

Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India

Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India

Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India

Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India

Associate Professor, Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, India

ABSTRACT: Developing a standard Optical Character Recognition (OCR) system involves several critical steps, including preprocessing, segmentation, feature extraction, and classification. Preprocessing, a particularly intriguing and complex part of Document Analysis and Recognition (DAR), involves transforming scanned or photographed images containing printed or handwritten text, such as numbers, letters, and symbols, into a format understandable by the system. Segmentation is a key process in any OCR system as it breaks down the text in images into lines, words, and characters. The effectiveness of the OCR system largely depends on the segmentation technique employed. To address severe degradations such as cuts, blobs, merges, and vandalism, Google Cloud Vision is employed to capture the contextual relationships within the document. This method effectively integrates document restoration and super-resolution, making the process efficient and yielding high-quality results from degraded documents. Extensive testing on various document sources, including magazines and books, has shown substantial improvements in image quality

KEYWORDS: Optical character recognition, Google Cloud Vision, Natural Language Processing.

I. INTRODUCTION

In the field of preservation and academic research, the restoration and safeguarding of delicate documents are of utmost importance. These documents act as windows into historical periods, providing crucial insights into past societies, cultures, and historical developments. Unfortunately, many of these documents have become fragile, faded, and difficult to read due to the passage of time. To uncover the valuable information contained within these historical artifacts, advanced restoration techniques are essential. This research paper explores the innovative field of document restoration using state-of-the-art technology to rejuvenate old records. Our study emphasizes the critical task of separating the foreground content from the background in these documents, as this separation is fundamental for subsequent restoration and analysis. Optical Character Recognition (OCR) plays a key role in this extraction process. By combining various techniques, we offer a thorough and accurate approach to document restoration, ensuring that researchers and historians can access clearer, more readable versions of these priceless historical documents. In this paper, we will discuss the historical importance of ancient documents and the difficulties involved in restoring and researching them. We will delve into the basic principles and methods of OCR and Natural Language Processing (NLP), highlighting how these technologies can address the specific challenges posed by fragile documents. Additionally, we will present experimental results and practical applications to demonstrate the effectiveness of our approach in document preservation. The integration of deep learning and advanced image processing techniques promises to connect the past with the present, allowing us to preserve, study, and honor the treasures of our shared history. In recent years, various OCR tools have emerged, each with distinct features. These include server-based OCR solutions like Google OCR, desktop options such as Tesseract and Python OCR, and web-based alternatives like Google Cloud OCR, Amazon OCR, and Microsoft OCR. The accuracy and effectiveness of text extraction vary widely among these OCR tools due to differences in their underlying pattern recognition algorithms. This paper focuses specifically on preprocessing techniques designed for Google Cloud OCR, aiming to optimize text extraction from challenging environments, including low-light conditions and dynamically changing settings.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205292 |

II. RELATED WORK

The author Jyoti Madake et al [1] discusses Optical Character Recognition (OCR), a technology used to identify text characters in digital copies of physical records like scanned paper documents. OCR converts the text into a language that can be processed digitally. It explains the process of OCR, starting with scanning the physical document, converting it into a digital format, and analysing the text to recognize characters using techniques like pattern recognition and feature detection. The paper's focus is on preprocessing and using Google Cloud OCR to extract text from challenging environments like low light. It proposes a system using a Raspberry Pi with a Night Vision camera, speaker, and software tools like Google Cloud Vision API, gTTS, Python Image Libraries, and OpenCV as suggested by D Vaidhyanathan et al [2]. The process involves capturing the document, adjusting brightness, applying image enhancement techniques, extracting text using Google Cloud Vision, and converting it to speech using gTTS. Multilanguage translation and reading capabilities are supported using Goslate, with testing conducted in languages like English, Tamil, and Hindi. The author focused Jazmyne Lavalas et al [3] on using Optical Character Recognition (OCR) technology to transcribe handwritten letters by Dr. Blythe Owen, a prominent Seventh-day Adventist musician, composer, and pedagogue, into text-based documents. This effort aims to enhance accessibility for researchers interested in Owen's correspondence. The interdisciplinary study draws on fields such as musicology, archival science, software engineering, and Artificial Intelligence (AI) programming. It provides practical examples of OCR capabilities for archival transcription, discussing the importance of Owen's letters to American musicology and Seventh-day Adventist history. Additionally, it offers a comparative review of four OCR programs (Google Cloud Vision, Pen to Print APP, SimpleOCR, and Transkribus) applied to the Owen letter dataset, building on previous discussions in the field.

III. METHODOLOGY

The entire process is divided into three phases. First, the input text undergoes preprocessing, where the input image is converted to grayscale. Then, binarization is applied to the grayscale image to enhance text visibility. Following this, the binarized image is encoded into image format bytes (e.g., jpeg, jpg, png, etc.). This system demonstrates a method for restoring damaged documents, such as those that have been blurred, overwritten, or stained.
The system consists of three modules:

Background Noise Removal module:

- Grayscale Conversion: Transform the input document image into grayscale to simplify the image and eliminate color information.
- Binarization: Perform binarization on the grayscale image to improve text visibility.
- Encode the binarized image into image format bytes (e.g., jpeg, jpg, png).

Text Extraction Module:

- Use Optical Character Recognition to extract text from the processed image.

Text Correction Module:

- Apply Natural Language Processing techniques to correct errors in the text extracted by OCR.

Figure 1 Workflow diagram

This system takes a damages document as an input and produces a restored document image and text as output.

A. Background Noise Removal Module

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205292 |

Grayscale Conversion:
Convert the input document image to grayscale to simplify the image and remove color information. Which Includes the following steps:

- Load the Input Image: Start by loading the color image that needs preprocessing.
- Grayscale Conversion: Transform the loaded color image into a grayscale image using a method such as weighted channel averaging. Calculate the intensity of each pixel using a formula like:
- Intensity=0.299×Red+0.587×Green+0.114×Blue
- Single-Channel Image: The resulting is a single-channel grayscale image where each pixel reflects the brightness of the corresponding pixel in the original color image.
- Normalization (Optional): Optionally, adjust the pixel values to a specific range (e.g., [0, 255]).

Binarization:
Apply binarization to the grayscale image to enhance text visibility. Binarization transforms the image into a binary image with only black and white pixels, which increase the contrast between text and background.

- Thresholding: Select a threshold value to divide the grayscale image into two categories: pixels with intensity values above the threshold are turned white, while those below are turned black. A simple thresholding operation might be described as:

$$\text{Binary Value} = \begin{cases} 1 & \text{if Intensity} > \text{Threshold} \\ 0 & \text{otherwise} \end{cases}$$

Adaptive Thresholding: Alternatively, use adaptive thresholding methods that take into account local pixel neighborhoods, adjusting the threshold value dynamically based on the local characteristics of image.

Morphological Operations: Optionally, employ morphological operations like erosion and dilation to improve the binary image and remove small noise or unwanted artifacts.

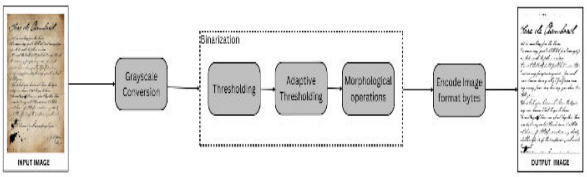


Figure 2 Preprocessed Image

- Post-Binarization Encoding: After binarization, the binary image should be encoded into a standard image format (e.g., JPEG, JPG, PNG).
- Image Conversion: Utilize a suitable library or tool to convert the binary image data into image format bytes.
- Save or Transmit: Store the encoded image bytes in a file or send them as required. This step ensures that the processed and binarized image is preserved in a standard image format.

B. Text Extraction Model

For this process, we will utilize Optical Character Recognition (OCR). Recognizing text in images is a crucial task in computer vision with a wide range of applications. OCR is a commonly used tool for extracting text from images, employing various feature extraction techniques to ensure high accuracy. These techniques include line and stroke analysis, pattern recognition, and statistical analysis.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205292 |

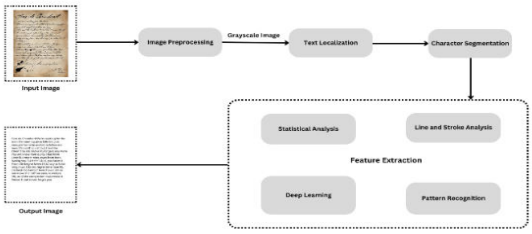


Figure 3 OCR flow

The steps for the OCR process are outlined below:

- Image Preprocessing : The input image, which may contain text, undergoes preprocessing to enhance the text quality for OCR. This can include operations such as resizing, noise reduction, contrast adjustment, and binarization.
- Segmentation: During this step, the OCR system examines the localized text regions and identifies individual characters or text lines. Accurate character segmentation is essential for correctly recognizing each character.
- Feature Extraction: OCR algorithms utilize different feature extraction techniques to capture the visual characteristics of each character. These features may encompass stroke patterns, shapes, textures, and other pertinent visual information.
- Pattern Recognition: The extracted features are subjected to pattern recognition techniques to identify and classify characters. Machine learning models, such as neural networks or support vector machines, may be used to learn patterns from training data and apply this knowledge to recognize characters in new images.
- Statistical Analysis: Statistical analysis is frequently employed to enhance OCR results. This can involve using statistical models to correct errors, manage variations in text appearance, and improve overall accuracy.
- Text Output: The final result of the OCR process is the recognized and extracted text. This text can be delivered in various formats, including plain text, machine-readable data, or structured data, depending on the needs of the application.

C. Text Correction Module

The text correction module employs Natural Language Processing (NLP) techniques to improve the accuracy and readability of text extracted from images. This module uses NLP methods to refine and enhance the clarity of the text obtained through OCR.

- This module receives the OCR-generated text as input and consolidates multiline text into a single line for easier processing.
- It utilizes the Language Tool to apply corrections to the text.
- The module then uses Spacy to identify suggested alternatives for each token in the text.
- Finally, it employs TextBlob for word correction, identifying and correcting each instance of the text and combining these corrections to enhance accuracy.

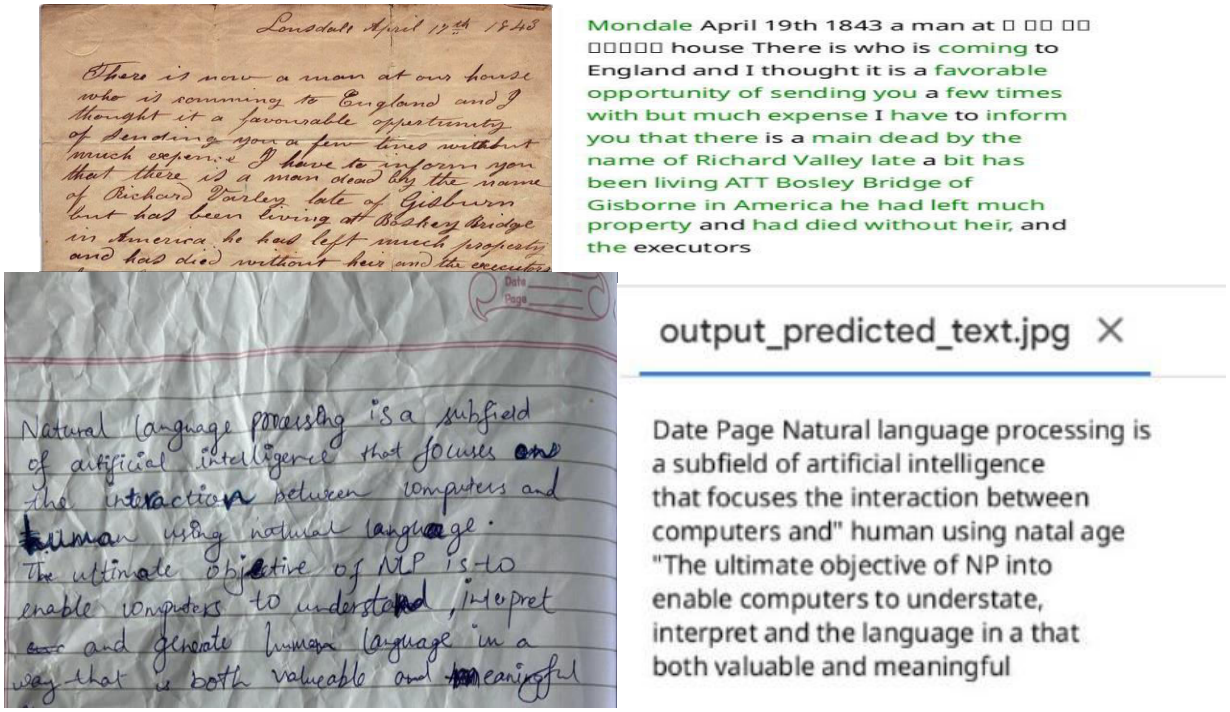
IV. EXPERIMENTAL RESULTS

In this section, we present the results of our testing and validation process, which involved a dataset of 30 images. These images included 20 handwritten images and 10 digital images, selected to evaluate the performance of our model, which comprises two phases: Optical Character Recognition (OCR) and subsequent correction using Natural Language Processing (NLP). The initial phase, OCR alone, achieved an overall accuracy of 69.6%. However, after applying the NLP correction phase, the overall accuracy improved to 76.34%.This indicates a significant enhancement in text recognition and correction capabilities provided by the NLP phase.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205292 |



V. CONCLUSION

The restoration project for fragile documents leverages optical character recognition (OCR) technology as a key component in the preservation process. OCR technology allows for the digitization of text from scanned images, enabling the creation of searchable and editable digital versions of these documents. By using OCR, the restoration team can extract text from old and deteriorated documents, even those with faded or damaged sections. This process restores clarity to illegible portions, improving readability and accessibility. Moreover, OCR aids in creating digital backups, which reduces the need to handle the original documents and minimizes the risk of additional damage. In summary, incorporating OCR technology into the restoration project significantly enhances the preservation and accessibility of fragile documents, ensuring their lasting legacy for future generations.

REFERENCES

[1] Jyoti Madake and Sameeran Pandey “Tabular Data Extraction From Documents” 2023
[2] D Vaithiyanathan, Manigandan Muniraj “Cloud based Text extraction using Google Cloud Vison for Visually Impaired applications” 2019.
[3] I Jazmyne Lavalas, Marianne Kordas, and Rodney Summerscales “Optical Character Recognition (OCR) Approaches to Cursive Handwriting Transcription: Lessons from the Blythe Owen Letters Project”. 2022
[4] Rajib Ghosh, Pooja ankri, Prabhat Kumar “RNN Based Online Handwritten Word Recognition in Devanagari Script” 2018.
[5] ShengHe,LambertSchomaker, “DeepOtsu: Document Enhancement and Binarization using Iterative Deep Learning”. 2019
[6] Mayank Wadhwani, Debapriya Kundu, Deepayan Chakraborty, and Bhabatosh Chanda“Text Extraction and Restoration of Old Handwritten Documents” 2020
[7] Y. Assael, T. Sommerschield, and J. Prag, "Restoring Ancient Text Using Deep Learning: A Case Study on Greek Epigraphy," 2019.
[8] Soumya A and G Hemantha Kumar “Enhancement and Segmentation of Historical Records. 2015
[9] Ravneet Kaur, Dharam Veer Sharma “Punjabi Text Recognition System for Portable Devices: A Comparative Performance Analysis of Cloud Vision API with Tesseract”. Vol.2, No. 2, 2021
[10] Ishwari S. Kulkarni, Anushree N. Pandit, Priya A. Kharate, Swapnali S. Tikkal, Dr. Sandeep Chaware “Proposed Design to Recognise Ancient Sanskrit Manuscripts With Translation using Machine Learning” 200
[11] S. U. Khan, I. Ullah, F. Khan, Y. Lee, and S. Ullah, ”Historical Text Image Enhancement Using Image Scaling and Generative AdversariaNetworks,” 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details