STM JOURNALS

Trends in
Opto-electro & Optical Communication

ISSN: 2231-0401 (Online)
ISSN: 2347-9957 (Print)
Volume 14, Issue 2, 2024
May–August
DOI (Journal): 10.37591/TOEOC

https://journals.stmjournals.com/toeoc

Review TOEOC

Digital Resurrection: Restoring Fragile Documents with OCR

Aniket Rawat[1,*], Shivam Kudal[2], Akshay Pawar[3], Chirag Fulfagar[4], Akshay Pawar[5], Shalaka Deore[6]

Abstract

In creating a typical optical character recognition (OCR) system, several steps are involved, such as preprocessing, segmentation, feature extraction, and classification. Preprocessing, which is a particularly interesting and challenging aspect of document analysis and recognition (DAR), deals with converting scanned or photographed images containing machine-printed or handwritten text, including numbers, letters, and symbols, into a format that the system can understand. Segmentation is a crucial task in any OCR system, as it breaks down image text documents into lines, words, and characters. The accuracy of the OCR system heavily relies on the segmentation algorithm used. To handle significant degradation like cuts, blobs, merges, and vandalism, Google Cloud Vision is utilized to capture contextual relationships within the document. Moreover, the method seamlessly combines document restoration and super-resolution, making the process efficient and producing high-quality results directly from degraded documents. Through extensive testing on various document sources like magazines and books, significant improvements in image quality have been demonstrated. The approach is robust and adaptable, particularly excelling with severely degraded documents like books, making it an ideal solution for digital libraries and similar repositories aiming to preserve and enhance document collections.

Keywords: Optical character recognition (OCR), Google Cloud Vision, document analysis and recognition (DAR), feature extraction, digital libraries

INTRODUCTION

In the realm of preservation and academic research, the restoration and preservation of fragile documents hold a unique significance. These documents serve as a time capsule to bygone eras, offering valuable insights into the past, cultural heritage, and the evolution of societies. However, the ravages of time have rendered many of these documents fragile, faded, and challenging to decipher. In the pursuit of uncovering the secrets hidden within these historical artifacts, advanced techniques are imperative. This research paper delves into the innovative realm of document restoration, where we employ cutting-edge technology to breathe new life into ancient records. Our research focuses on the vital task of extracting the foreground content and background information from the documents, as this separation forms the foundation for subsequent restoration and analysis. The extraction process is facilitated by optical character recognition (OCR). This combination of techniques offers a comprehensive and precise approach to document restoration, ensuring that researchers and historians gain access to a clearer, more legible version of these invaluable relics from the past. In this paper, we will explore

***Author for Correspondence**
Aniket Rawat
E-mail: aniket.n.rawat@gmail.com

[1-6]Student, Department of Computer Engineering, MES Wadia College of Engineering, S.P. Pune University. Pune, Maharashtra, India

Received Date: July 24, 2024,
Accepted Date: July 31, 2024
Published Date: August 11, 2024

Citation: Aniket Rawat, Shivam Kudal, Akshay Pawar, Chirag Fulfagar, Akshay Pawar, Shalaka Deore. Digital Resurrection: Restoring Fragile Documents with OCR. Trends in Opto-electro & Optical Communication. 2024; 14(2): 29–35p.

the historical significance of ancient documents and the challenges they pose for restoration and research. We will delve into the underlying principles and methodologies of OCR and natural language processing (NLP), showcasing how these technologies can be leveraged to tackle the unique challenges posed by fragile documents. Furthermore, we will present experimental results and real-world applications, demonstrating the effectiveness of our approach in the realm of document preservation. The fusion of deep learning and advanced image processing techniques promises to bridge the gap between the past and the present, allowing us to preserve, study, and celebrate the treasures of our collective history. Recent years have seen the emergence of various OCR tools, each offering unique capabilities. These include server-based OCR solutions like Google OCR, desktop options such as Tesseract and Python OCR, as well as web-based alternatives like Google Cloud OCR, Amazon OCR, and Microsoft OCR. The accuracy and effectiveness of text extraction can differ significantly between these OCR types due to variations in the underlying pattern recognition algorithms they employ. This paper specifically concentrates on preprocessing techniques tailored for Google Cloud OCR. The aim is to optimize text extraction from challenging environments, such as low-light conditions or dynamically changing settings.

LITERATURE SURVEY

Madake and Pandey [1] discuss 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 analyzing the text to recognize characters using techniques like pattern recognition and feature detection. OCR results using Cloud Vision and Tesseract are shown in Table 1.

Table 1. Optical character recognition (OCR) results using Cloud Vision and Tesseract.

Word Court	Character Count	Runtime Recognized Words Characters (Seconds) Using Vision API	Recognized Words /Characters/ Runtime (Seconds) Using Tesseract
69	386	68/365/4	62/359/25
50	252	50/252/5	44/247/20
102	513	102/513/4	84/479/45
33	139	33/138/3	18/81/8
85	348	85/348/6	74/337/22
107	528	107/528/5	83/488/18
26	137	26/137/3	22/133/10
24	104	24/104/2	24/104/6
121	610	121/610/6	86/462/32
47	217	47/217/3	47/217/10
51	262	51/262/4	40/231/21
77	434	76/433/5	55/390/40
42	232	42/232/3	40/230/20
92	524	91/523/5	72/494/25
85	376	85/376/8	68/331/34
134	707	131/704/5	90/574/90
163	762	160/758/4	136/731/85
65	356	64/355/4	41/273/22
123	589	120/586/5	103/531/50
138	662	138/662/4	138/662/25
99	430	99/430/5	81/375/79
153	810	153/810/5	141/797/90
91	508	91/508/6	67/410/60
107	497	106/406/5	88/437/69
59	305	59/305/4	54/300/32
Total: 2143	*10668*	*2129/10652/113*	*1758/9673/938*

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 (Google Text-to-Speech), Python Image Libraries, and OpenCV as suggested by Vaithiyanathan and Muniraj [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.

Lavalas et al. [3] focused on using 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.

Keshri et al. [4] focus on the preprocessing and segmentation of ancient scripts to automate the task of reading and deciphering inscriptions. It addresses the enhancement of degraded ancient document images through spatial filtering methods and binarization using the Otsu thresholding algorithm. The system achieves good results when tested on degraded epigraphic images, with better enhanced output for specific mask sizes for each filter. It also effectively samples characters from enhanced images, achieving a segmentation rate of 85% to 90% for both Drop Fall and Water Reservoir techniques. Other related work includes the use of histogram shape analysis for global binarization, which performs well with twin modal distribution histograms. Another method utilizes image contrast to recognize text stroke boundaries and produce high accuracy binarization results. A modified iterative global threshold algorithm is effective for separating object information from the foreground, especially in documents with non-uniform distribution of noises. Additionally, segmentation techniques such as connected components, nearest neighbor algorithm, multiple histogram projections, and contour tracing are explored for text line segmentation and character separation. These approaches contribute to the field of document image analysis and provide insights into enhancing and segmenting historical documents for improved automatic recognition and comprehension.

He and Schomaker's [5] paper is based on first ever restoration model that is based on restoring the missing characters form a text that has been damaged over the time using deep neural networks named Pythia, which has a character error rate of only 30.1% of character as compared to human epigraphists, which is of 57.3% of error rate and is comparatively very high.

The paper proposes a new method for image scaling using generative adversarial networks (GANs). The proposed method by Wadhwani et al. [6] can scale images up and down while preserving their quality. In this method, a GAN is trained to differentiate between low- and high-resolution images after which the GAN model is used to generate high resolution image from the low-resolution image.

METHODOLOGY

The entire process has been divided into three phase where the input text is first preprocessed where we convert the input image into grayscale after which we apply binarization to the grayscale image in order to enhance the text visibility after which it encodes the binarized image as image format bytes (e.g., jpeg, jpg, png, etc).

This system shows a method for restoring damaged documents, such as those that have been blurred, overwritten, or stained.

The system consists of three modules:

1. Background Noise Removal module:
   a. *Grayscale conversion:* Convert the input document image to grayscale to simplify the image and remove color information.
   b. *Binarization:* Apply binarization to the grayscale image to enhance text visibility.
   c. Encoding as Image Format Bytes.

2. Text Extraction module:
   a. *Text extraction:* Using OCR to extract text from the preprocessed image.

3. Text Correction Module:
   a. *Text correction:* Using NLP techniques to correct errors in OCR extracted text.

This system takes a damaged document image as input and produces a restored document image and text as output (Figure 1).

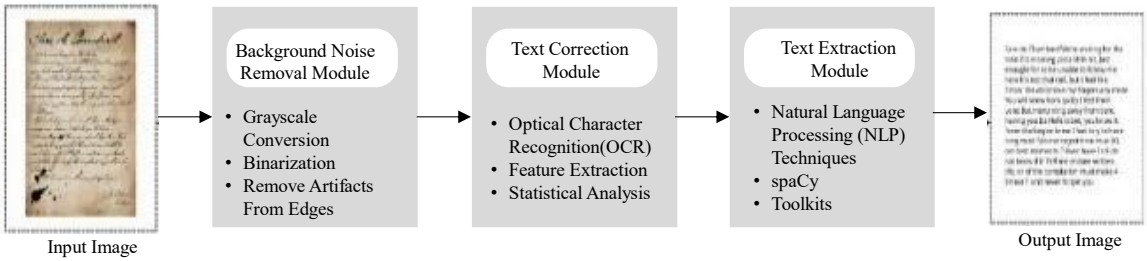

Figure 1. Workflow of the proposed system.

Background Noise Removal Module

Grayscale Conversion

Convert the input document image to grayscale to simplify the image and remove color information (Figure 2). This includes the following steps:

1. *Load the input image*: Begin by loading the color image that needs preprocessing.
2. *Grayscale conversion:* Convert 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:

   $$\text{Intensity} = 0.299 \times \text{Red} + 0.587 \times \text{Green} + 0.114 \times \text{Blue}$$

3. *Single-channel image:* The result is a single-channel grayscale image where each pixel represents the brightness of the corresponding pixel in the original color image.
4. Normalization (Optional): Optionally, normalize the pixel values to a specific range (e.g., [0, 255]).

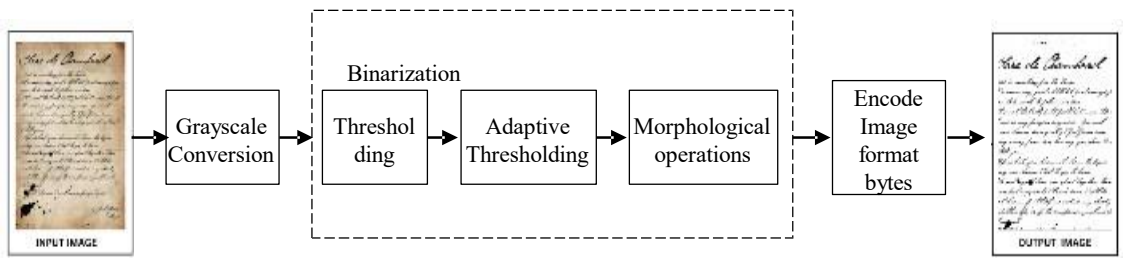

Figure 2. Preprocessing.

Binarization

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

Thresholding

- Choose a threshold value to separate the grayscale image into two classes: pixels with intensity values above the threshold are set to white, and those below are set to black.
- A basic thresholding operation might look like:

where:

I(x,y) is the intensity value of the pixel at position(x,y)
T is the threshold value.

$$\text{Binary Value} \begin{cases} 1, & \text{if Intensity } (x, y) > \text{Threshold} \\ 0, & \text{Otherwise} \end{cases}$$

Adaptive Thresholding (Optional)

Optionally, use adaptive thresholding techniques that consider local pixel neighborhoods, adjusting the threshold dynamically based on the local image characteristics.

Morphological Operations (Optional)

Optionally, apply morphological operations like erosion and dilation to refine the binary image and remove small noise or unwanted artifacts (Figure 2).

Encoding as Image Format Bytes

Encode as Image Format Bytes

- Once the binarization is done, the binary image needs to be encoded into a standard image format (e.g., jpeg, jpg, png).
- Use an appropriate library or tool to convert the binary image data into image format bytes.

Save or Transmit: Save the encoded image bytes to a file or transmit them as needed. This step ensures that the processed and binarized image is stored in a standard image format.

Text Extraction Model

For this process, we will be using OCR. The recognition of text in images is a fundamental task in computer vision with numerous applications. OCR is a widely used tool for extracting text from images, employing a combination of feature extraction methods to achieve high accuracy. The methods included in OCR are line and stroke analysis, pattern recognition and statistical analysis.

The steps of OCR process are as follows:

1. *Image preprocessing:* The input image, which may contain text, undergoes preprocessing steps to enhance the quality of the text for OCR. Preprocessing may include operations like resizing, noise reduction, contrast adjustment, and binarization.
2. *Text localization:* OCR algorithms often include a text localization step to identify regions in the image that likely contain text. This step helps narrow down the focus to areas where OCR needs to be applied, improving efficiency.
3. *Character segmentation:* In this step, the OCR system analyzes the localized text regions and identifies individual characters or text lines. Character segmentation is crucial for recognizing each character accurately.
4. *Feature extraction:* OCR algorithms use various feature extraction methods to represent the visual characteristics of each character. These features may include stroke patterns, shape, texture, and other relevant visual information.
5. *Pattern recognition:* Pattern recognition techniques are applied to the extracted features to recognize and classify characters. Machine learning models, such as neural networks or support vector machines, may be employed to learn patterns from training data and generalize to recognize characters in new images.

6. *Statistical analysis:* Statistical analysis is often utilized to refine OCR results. This may involve using statistical models to correct errors, handle variations in text appearance, and improve overall accuracy.
7. *Text output:* The final output of the OCR process is the recognized and extracted text. This can be provided in various formats, such as plain text, machine-readable data, or structured data, depending on the application requirements.

Text Correction Module

The text correction module utilizes NLP techniques to enhance the accuracy and readability of text extracted from images.

i. This module accepts the OCR-generated text and an input after which it combines multi line text into a single line for processing.
ii. This module uses the Language Tool for applying correction to the texts.
iii. After which it uses Spacy to identify suggested alternatives for each token in the text
iv. Then it uses the textBlob for word correction to identify and correct each version of the text in the text and combine correction to improve the accuracy.

APPLICATION

- *Preserving legal and administrative documents:* Not only can it be used to preserve documents but it can also be used to protect and preserve government records.
- *Education:* Improved and enhanced document provides a wider range of knowledge to the students and help academic research by restoring and preserving valuable documents for scholarly investigation.
- *Legal and forensic investigations:* Assist in legal investigations by restoring and interpreting historical legal documents relevant to legal cases and disputes [7, 8].
- *Expanding digital libraries:* By contributing in digital libraries and archives by digitizing and restoring rare and fragile documents, expanding the availability of resources online [9, 10].

FUTURE SCOPE

Increasing resolution: The primary objective is to generate higher-resolution images from lower-resolution inputs.

Enhancing visual quality: The goal is to make the resulting images visually appealing and natural-looking.

Cost efficiency: Super-resolution can offer a cost-effective solution to obtain high-quality images without the need for high-end camera equipment.

CONCLUSION

The restoration project for fragile documents employs OCR technology as a crucial tool in the preservation process. OCR technology enables the digitization of text from scanned images, allowing for the creation of searchable and editable digital versions of the documents. By utilizing OCR, the restoration team can extract text from aged and deteriorated documents, even those with faded or damaged text. This enables the restoration of clarity to illegible portions of the text, enhancing readability and accessibility. Additionally, OCR facilitates the creation of digital backups, reducing the need for handling of the original documents and minimizing the risk of further damage. In conclusion, the integration of OCR technology into the restoration project enhances the preservation and accessibility of fragile documents, ensuring their enduring legacy for future generations.

REFERENCES

1. Madake J, Pandey S. Tabular data extraction from documents. In: Mahapatra RP, Peddoju SK, Roy S, Parwekar P, editors. Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022. Singapore: Springer; 2023. pp. 429–439.

2. Vaithiyanathan D, Muniraj M. Cloud based text extraction using Google Cloud Vison for visually impaired applications. In: 2019 11th International Conference on Advanced Computing (ICoAC), Chennai, India, December 18–20, 2019. pp. 90–96.
3. Lavalas J, Kordas M, Summerscales R. Optical character recognition (OCR) approaches to cursive handwriting transcription: lessons from the Blythe Owen Letters Project. J Adventist Arch. 2022; 2: 53–71.
4. Keshri P, Kumar P, Ghosh R. RNN based online handwritten word recognition in Devanagari script. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, August 5–8, 2018. pp. 517–522.
5. He S, Schomaker L. DeepOtsu: document enhancement and binarization using iterative deep learning. Pattern Recogn. 2019; 91: 379–390.
6. Wadhwani M, Kundu D, Chakraborty D, Chanda B. Text extraction and restoration of old handwritten documents. In: Mukhopadhyay J, Sreedevi I, Chanda B, Chaudhury S, Namboodiri VP, editor. Digital Techniques for Heritage Presentation and Preservation. New York, NY, USA: Springer International; 2021. pp. 109–132.
7. Assael Y, Sommerschield T, Prag J. Restoring ancient text using deep learning: a case study on Greek epigraphy. arXiv preprint. arXiv:1910.06262. October 14, 2019. Available at https://arxiv.org/abs/1910.06262
8. Soumya A, Kumar GH. Enhancement and segmentation of historical records. In: Wyld DC, Meghanathan N, editors. Computer Science & Information Technology (ACITY, DPPR, VLSI, WiMNET, AIAA, CNDC). Proceedings of Fifth International Conference on Advances in Computing and Information Technology. Chennai, India: Academy & Industry Research Collaboration Center; 2015. pp. 95–113.
9. Kaur R, Sharma DV. Punjabi text recognition system for portable devices: a comparative performance analysis of Cloud Vision API with Tesseract. J Computer Sci Eng. 2021; 2 (2): 104–111.
10. Kulkarni I, Tikkal S, Chaware S, Kharate P, Pandit A. Proposed design to recognize ancient Sanskrit manuscripts with translation using machine learning. In: Proceedings of the International Conference on Innovative Computing & Communication (ICICC), New Delhi, India, February 19–20, 2022.

