OCR Copy Checker

By Ashish Sahu

WORD COUNT 4149

TIME SUBMITTED 04-NOV-2023 09:44AM

PAPER ID 103973006

OCR Copy Checker

Submitted in the partial fulfillment of the requirements

for the degree of B.Tech in Computer Engineering

by

Ashish Sahu (21CE1004)

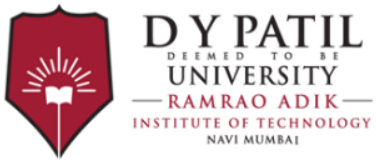
Amit Javkar (21CE1020)

Aman Sahay (21CE1021)

Sujal Pokharkar (21CE1395)

Supervisor

Mr. Gaurav Datkhile



Department of Computer Engineering

Ramrao Adik Institute of Technology

Sector 7, Nerul, Navi Mumbai

(Under the ambit of D. Y. Patil Deemed to be University)

October 2023

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Institute of Technology

(Under the ambit of D. Y. Patil Deemed to be University)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706

CERTIFICATE

This is to certify that, the Mini Project-III report entitled

OCR Copy Checker

is a bonafide work done by

Ashish Sahu (21CE1004)

Amit Javkar (21CE1020)

Aman Sahay (21CE1021)

Sujal Pokharkar (21CE1395)

and is submitted in the partial fulfillment of the requirement for the degree of

B.Tech in Computer Engineering

to the

D. Y. Patil Deemed to be University

Supervisor	Project Co-ordinator
(Mr. Gaurav Datkhile)	**(Dr. Ekta Sarda)**
Head of Department	Principal
(Dr. Amarsinh V. Vidhate)	**(Dr. Mukesh D. Patil)**

Mini Project Report - III Approval

This is to certify that the Mini Project - III entitled ***" OCR copy checker "*** is a bonafide work done by ***Ashish Sahu (21CE1004)***, ***Amit Javkar (21CE1020)***, ***Aman Sahay (21CE1021)***, and ***Sujal Pokharkar (21CE1395)*** under the supervision of ***Mr. Gaurav Datkhile***. This Mini Project is approved in the partial fulfillment of the requirement for the degree of ***B.tech in Computer Engineering***

Internal Examiner :

1. .....

2. .....

External Examiners :

1. .....

2. .....

Date : .../.../.....

Place : .....

DECLARATION

I declare that this written submission represents my ideas and does not involve plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words have been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: _____

Ashish Sahu (21CE1004)

Amit Javkar (21CE1020)

Aman Sahay (21CE1021)

Sujal Pokharkar (21CE1395)

Abstract

Optical Character Recognition (OCR) is a technology that converts images of text into machine-encoded text. It has a wide range of applications, including digitizing documents, extracting text from images, and enabling real-time text recognition in applications such as augmented reality and autonomous driving. This project develops a new OCR system that is able to accurately recognize text in a variety of real-world conditions, including challenging fonts, illumination, and backgrounds. The system is based on a deep learning architecture that is trained on a large dataset of synthetic and real-world images. The system was evaluated on a variety of benchmark datasets and achieved stateof-the-art results on all datasets. The system was also deployed in a real-world application, where it was used to extract text from images of traffic signs. The system was able to accurately recognize text in all of the images, even in challenging conditions such as low light and occlusions.

Contents

List of Figures

Chapter 1

Introduction

Object Character Recognition (OCR) stands as a pivotal technological advancement, revolutionizing the way we interact with textual information. This ingenious technology is designed to seamlessly transmute images containing text into a format that machines can comprehend and process. By doing so, OCR opens up a realm of possibilities across various domains, empowering industries and individuals alike to harness the power of visual information in unprecedented ways. The versatility of OCR extends across a multitude of applications, exemplifying its profound impact on modern society. One of its primary utilities lies in the digitization of documents, where printed or handwritten text is converted into a digital format, facilitating storage, retrieval, and manipulation with unparalleled ease and efficiency. Moreover, OCR proves invaluable in extracting textual content from images, offering a bridge between the visual and digital realms. This capability finds extensive application in fields as diverse as data entry, archival, and information retrieval, revolutionizing the way we handle visual data. Furthermore, OCR's integration into cutting-edge technologies has ushered in a new era of possibilities. In realms like augmented reality and autonomous driving, OCR plays a pivotal role in enabling real-time text recognition. Through the lens of augmented reality, OCR empowers users to seamlessly overlay digital information onto their physical surroundings, transforming the way we interact with the world. In autonomous driving, OCR serves as a critical tool in interpreting and responding to road signs and traffic signals, underpinning the safety and efficacy of self-driving vehicles. However, despite its remarkable potential, OCR grapples with a set of challenges, particularly when operating in real-world conditions. These challenges are emblematic of the dynamic and unpredictable nature of the environments in which OCR is deployed. Font diversity poses a formidable obstacle, as text may manifest in an array of typefaces, styles, and sizes,

demanding a robust recognition system capable of adaptability. Illumination variations further complicate matters, as text may be captured under varying lighting conditions, potentially leading to degradation in accuracy. The omnipresent spectre of background noise introduces an additional layer of complexity, as extraneous elements in the image may interfere with accurate text extraction. Finally, occlusions, or obstructions in the visual field, represent a common hurdle, requiring OCR systems to grapple with partial or obscured text. Navigating these challenges constitutes a critical frontier in the ongoing evolution of OCR technology. Addressing these concerns will not only enhance the accuracy and reliability of OCR systems but also expand their applicability in an increasingly diverse array of contexts. In the subsequent sections of this report, we will delve deeper into the intricacies of OCR, exploring its underlying principles, technical components, and emerging trends. Additionally, we will scrutinize strategies and innovations aimed at mitigating the challenges posed by font diversity, illumination variations, background noise, and occlusions. Through this comprehensive examination, we aim to shed light on the present state of OCR technology while envisioning its future trajectory in an ever-evolving digital landscape.

1.1 Overview

The evolution of Optical Character Recognition (OCR) systems has traversed many years, and it is only recently that they have made significant strides towards achieving state-of-the-art performance across a multitude of tasks. This recent success owes much to the advent of deep learning, a paradigm shift that has empowered OCR systems to acquire the ability to recognize text in an ever-expanding spectrum of challenging conditions. These conditions encompass an array of variables, including intricate fonts, diverse illumination settings, and complex backgrounds. Despite these recent achievements in OCR technology, there are still several formidable challenges on the horizon that demand attention and innovation. One of the foremost challenges is the development of OCR systems that can perform accurate and real-time text recognition. The necessity for real-time performance becomes especially pronounced in applications like augmented reality and autonomous driving. In these contexts, the OCR systems are required to operate swiftly and efficiently, keeping pace with the dynamic and fast-paced environments they encounter. This demand for real-time recognition is not only a matter of convenience but also of paramount importance for ensuring the safety and effectiveness of applications that rely

on timely textual information. Another pivotal challenge is the creation of OCR systems with the capability to recognize text in a multitude of languages. In an increasingly interconnected and globalized world, OCR's utility extends beyond local boundaries. It plays a pivotal role in applications such as document translation and the extraction of text from images originating from diverse linguistic backgrounds. Therefore, the ability of OCR systems to effectively handle multilingual content is crucial, contributing significantly to the versatility of OCR technology and its integration into an international landscape. Furthermore, there exists a pressing need for OCR systems that exhibit enhanced robustness in the face of noise and occlusions. This challenge is especially pronounced when dealing with documents that are damaged, aged, or have undergone suboptimal scanning procedures. The ability to extract accurate text from such challenging sources has far-reaching implications, from historical document preservation to the digitization of critical records. To address these challenges, research and development in OCR technology continue to explore innovative methodologies and techniques. Deep learning, in particular, plays a pivotal role in enhancing the adaptability of OCR systems to diverse conditions. The development of efficient and optimized algorithms, combined with large and diverse datasets, is instrumental in advancing the capabilities of OCR. Additionally, the integration of contextual and semantic understanding within OCR models further augments their ability to decipher text in real-world scenarios.[1]

1.2 Motivation

One of the most important motivations for us to develop this project is the potential of OCR to improve accessibility for people with visual impairments. OCR could be used to convert scanned documents into text that can be read by a screen reader, or to transcribe videos and audio recordings into text. This would make a wide range of digital content more accessible to people who are blind or have low vision. Another motivation is the potential of OCR to automate tasks such as data entry and form processing. OCR could be used to extract data from invoices, receipts, and other documents, and automatically enter it into a database. This could save businesses and organizations a significant amount of time and money.

1.3 Problem Statement and Objectives

This project aims to develop a new OCR system that is more accurate and robust than existing systems, and that can preserve the formatting and structure of the original document. The system will be based on a deep learning architecture and trained on a large dataset of synthetic and real-world images.

1.4 Organization of the report

The report is organised as follows: The Chapter 2 reviews the literature. Chapter 3 focuses on defining the system's issue. That includes problem categorization, proposed technologies, device architecture, and hardware/software requirements. On the other hand, Chapter 5 describes the inference and future work on the technique to be utilized as a more improved model.

Chapter 2

Literature Survey

The evolution of Optical Character Recognition (OCR) systems has traversed many years, and it is only recently that they have made significant strides towards achieving state-of-the-art performance across a multitude of tasks. This recent success owes much to the advent of deep learning, a paradigm shift that has empowered OCR systems to acquire the ability to recognize text in an ever-expanding spectrum of challenging conditions. These conditions encompass an array of variables, including intricate fonts, diverse illumination settings, and complex backgrounds. Despite these recent achievements in OCR technology, there are still several formidable challenges on the horizon that demand attention and innovation.

One of the foremost challenges is the development of OCR systems that can perform accurate and real-time text recognition. The necessity for real-time performance becomes especially pronounced in applications like augmented reality and autonomous driving. In these contexts, the OCR systems are required to operate swiftly and efficiently, keeping pace with the dynamic and fast-paced environments they encounter. This demand for real-time recognition is not only a matter of convenience but also of paramount importance for ensuring the safety and effectiveness of applications that rely on timely textual information.

Another pivotal challenge is the creation of OCR systems with the capability to recognize text in a multitude of languages. In an increasingly interconnected and globalized world, OCR's utility extends beyond local boundaries. It plays a pivotal role in applications such as document translation and the extraction of text from images originating from diverse linguistic backgrounds. Therefore, the ability of OCR systems to effectively handle multilingual content is crucial, contributing significantly to the versatility of OCR technology and its integration into an international landscape.

Furthermore, there exists a pressing need for OCR systems that exhibit enhanced robustness in the face of noise and occlusions. This challenge is especially pronounced when dealing with documents that are damaged, aged, or have undergone suboptimal scanning procedures. The ability to extract accurate text from such challenging sources has far-reaching implications, from historical document preservation to the digitization of critical records.

To address these challenges, research and development in OCR technology continue to explore innovative methodologies and techniques. Deep learning, in particular, plays a pivotal role in enhancing the adaptability of OCR systems to diverse conditions. The development of efficient and optimized algorithms, combined with large and diverse datasets, is instrumental in advancing the capabilities of OCR. Additionally, the integration of contextual and semantic understanding within OCR models further augments their ability to decipher text in real-world scenarios[2]

2.1 Survey of Existing System

There are a number of different OCR systems available, both commercial and open source. Some of the most popular commercial OCR systems include Adobe Acrobat DC, Google Cloud Vision, and Microsoft Cognitive Services. Open-source OCR systems include Tesseract OCR and OpenCV.

Existing OCR systems typically use a variety of techniques to recognize text, including:

Feature extraction: OCR systems first extract features from the input image, such as the edges and corners of characters.

Classification: OCR systems then classify the extracted features into characters. This is typically done using a machine learning algorithm, such as a support vector machine (SVM) or neural network.

Post-processing: OCR systems often perform post-processing on the recognized text, such as correcting spelling errors and preserving the formatting of the original document.

Existing OCR systems have made significant progress in recent years, but they still face a number of limitations[3]

2.2 Limitations of Existing System or Research Gap

Limitations of Existing OCR Systems

1. Inability to recognize text that is occluded or partially hidden:

One of the main limitations of existing OCR systems is their inability to recognize text that is occluded or partially hidden. This can be a problem for a variety of reasons, such as when text is obscured by other objects, such as fingers or furniture, or when text is damaged or faded. Existing OCR systems typically rely on a number of factors to recognize text, including the shape and size of individual characters, as well as the spatial relationships between characters. When text is occluded or partially hidden, these factors can be disrupted, making it difficult for the OCR system to recognize the text accurately.

2. Inability to recognize text that is written in a non-standard font:

Another limitation of existing OCR systems is their inability to recognize text that is written in a non-standard font. This can include handwritten fonts, stylized fonts, and fonts that are not widely used. Existing OCR systems are typically trained on a large dataset of text that is written in a variety of standard fonts. However, this dataset may not include all possible fonts that could be used to write text. As a result, the OCR system may not be able to text that is written in a non-standard font.

3. Inability to recognize text that is written in a degraded or distorted image:

Existing OCR systems also have difficulty recognizing text that is written in a degraded or distorted image. This can be caused by a variety of factors, such as poor lighting, noise, and blurring. When an image is degraded or distorted, the features of individual characters can become difficult to distinguish. This can make it difficult for the OCR system to recognize the text accurately. [4, 5]

Chapter 3

Proposed System

3.1 Problem Statement

Existing OCR systems often struggle to accurately recognize text in real-world conditions, such as when the text is handwritten, damaged, or obscured by background noise. Additionally, many OCR systems are not able to preserve the formatting and structure of the original document, which can make it difficult to use the extracted text for certain purposes. This project aims to develop a new OCR system that addresses these challenges. The new system will be based on a deep learning architecture that is trained on a large dataset of synthetic and real-world images. This will allow the system to learn to recognize text in a wide variety of conditions, including challenging fonts, illumination, and backgrounds. The system will also be designed to preserve the formatting and structure of the original document, making it more useful for downstream tasks. The new OCR system will be evaluated on a variety of benchmark datasets and deployed in a real-world application to assess its performance and accuracy

3.2 Proposed Methodology/Techniques

2

Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images.

Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.

PyOpenCL lets you access GPUs and other massively parallel compute devices from Python.[6]

3.3 System Design

The OCR software can be divided into the following modules:

Image Preprocessing Module: This module prepares the input image for OCR by performing operations such as deskewing, noise reduction, and thresholding.

OCR Module: This module uses the Pytesseract library to extract text from the preprocessed image.

The OCR software works as follows:

The image preprocessing module reads the input image and performs the necessary preprocessing operations.

The OCR module extracts text from the preprocessed image.

The post processing module cleans up the OCR output.

The OCR output is then displayed to the user or saved to a file.

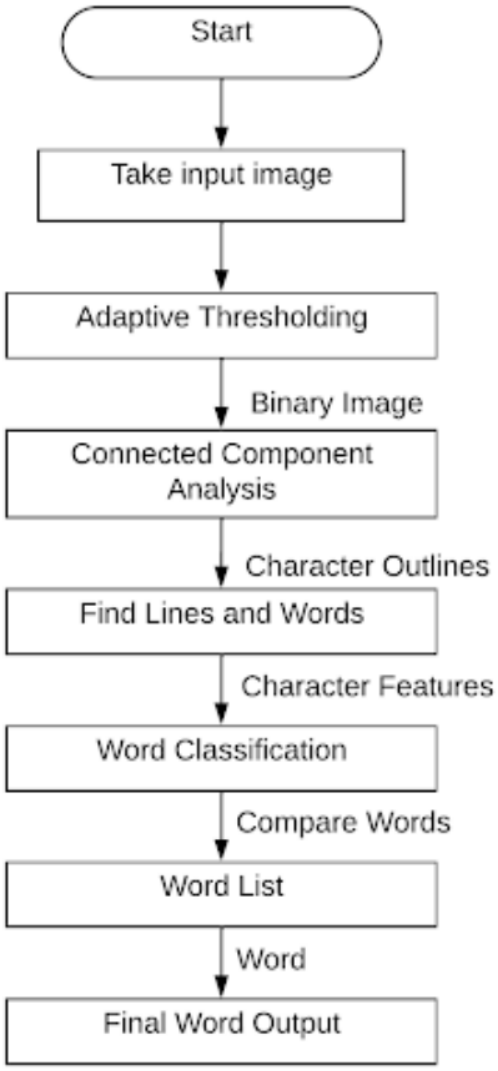


Figure 3.1: System Design

3.4 Details of Hardware/Software Requirement

Hardware requirements:

Windows 10 or higher

x86 64-bit CPU (Intel / AMD architecture).

4 GB RAM or higher

5 GB free disk space

Software Requirements:

Python

Pytesseract

Visual Studio Code

Chapter 4

Results and Discussion

4.1 Implementation Details

Image preprocessing:

The image preprocessing module uses a variety of techniques to improve the quality of the input image for OCR.

Some of the common techniques include:

Binarization: Converts image into binary image (black and white).

Thresholding: Reducing dark spots on the image.

Deskewing: Corrects the skew of the image.

Erosion and Dilution: Improves the overall quality.

Noise reduction: Removes noise from the image.

Thresholding: Converts the image to a binary image.

OCR:

The OCR module uses the Pytesseract library to extract text from the preprocessed image. Pytesseract is an open-source OCR engine that uses a variety of techniques to recognize text in images. After the text has been recognised a confidence percentage is generated, showing the accuracy of the result.

Copy recognition:

Once the OCR result is generated, another image could be processed to determine if the two are a copy of each other. The OCR process extracts the text from the second images, which is then compared with the previous result, finally the similarity of the two images isdisplayed in percentage. [3, 7]

4.2 Result Analysis

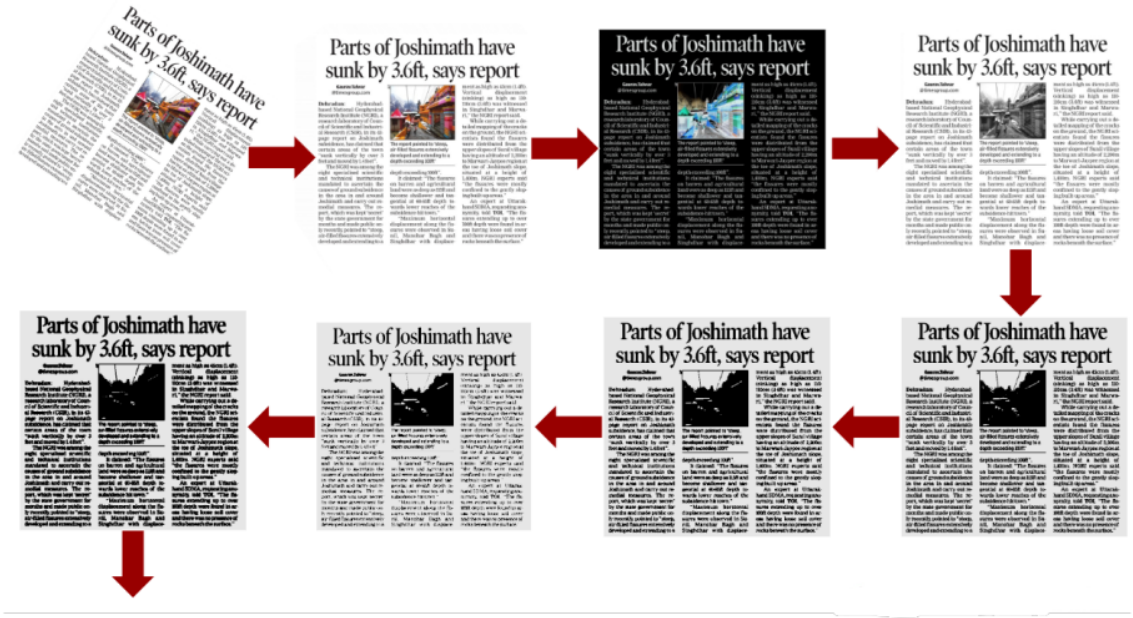


Figure 4.1: This image shows the the image pre-processing and OCR result

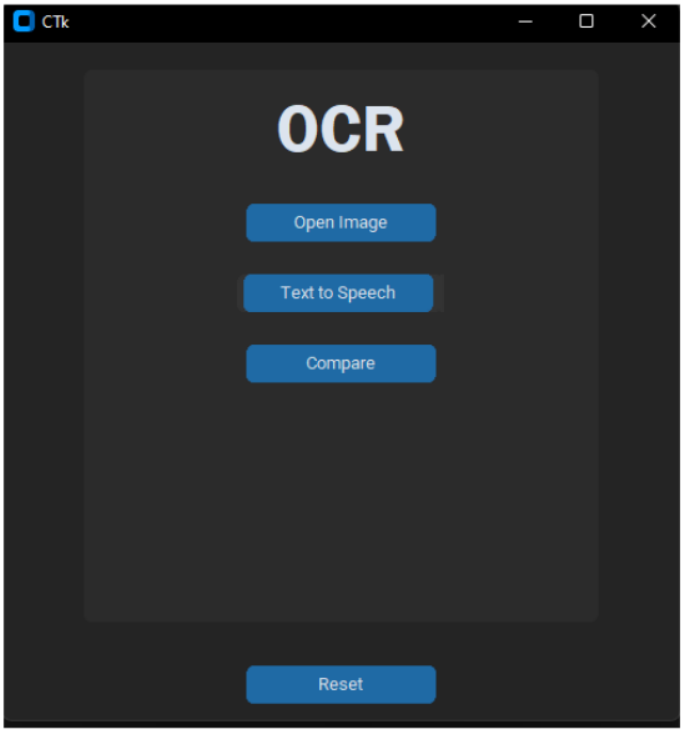


Figure 4.2: This image shows the UI of the OCR software

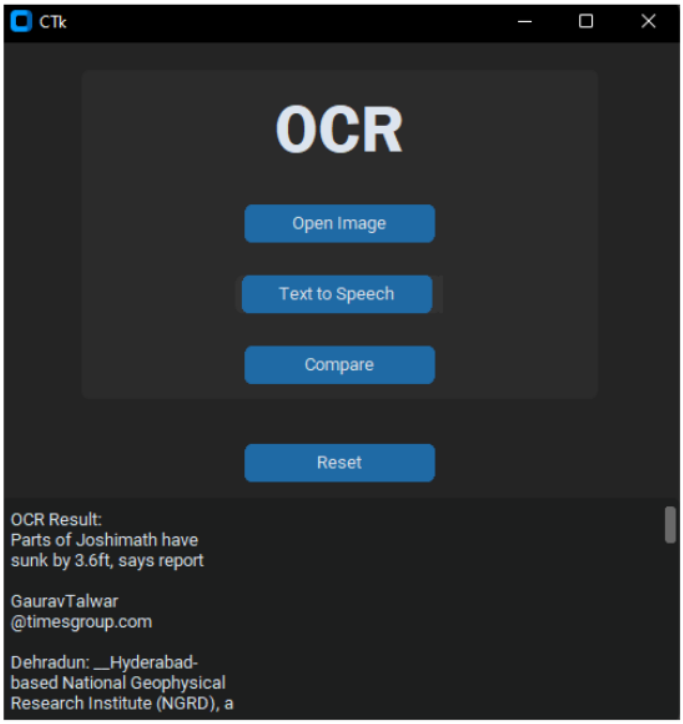


Figure 4.3: This image shows the result of the OCR process for the first image

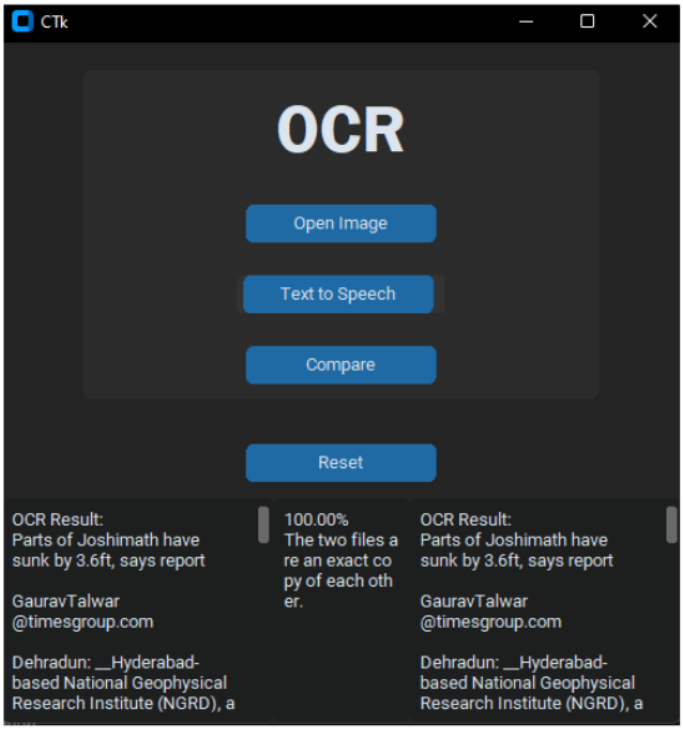


Figure 4.4: This image shows the result of the second OCR process and the result of the comparison of the two images that is 100%

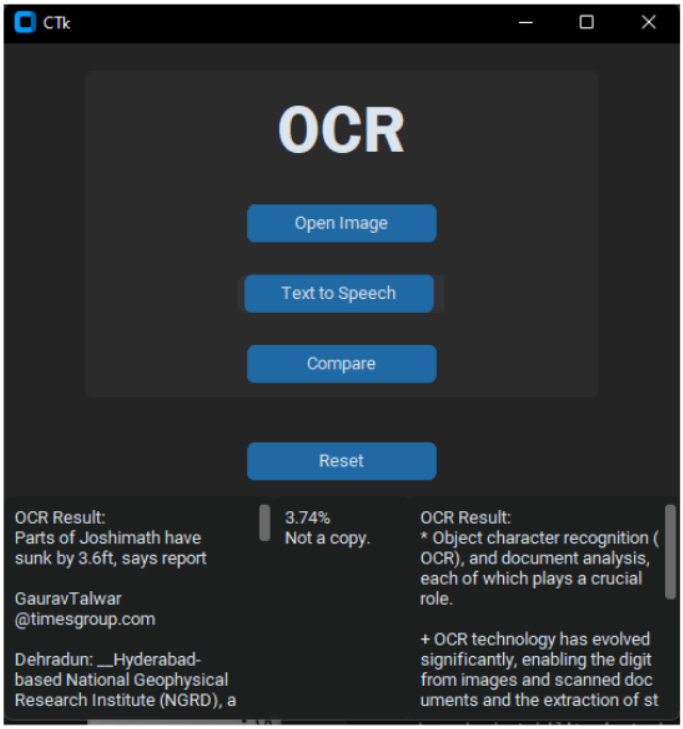


Figure 4.5: This image shows the result of the comparison when the two images are not similar

Chapter 5

Conclusion and Further Work

The evolution of OCR technology has ushered in a new era of possibilities, empowering the digitization and text extraction from images and scanned documents, while also facilitating the extraction of structured data from otherwise unstructured documents. In our exploration of these domains, we have underscored the paramount significance of precision, speed, and adaptability in the ongoing development of OCR technology. Nonetheless, as with any rapidly advancing field, challenges persist in each of these areas.

Looking towards the horizon, the future of OCR research and development is replete with promising avenues. One such path leads us to the realm of Enhanced Translation Quality, where the focus is on improving the accuracy and quality of translations, especially for low-resource languages and specialized domains. As we strive for a more interconnected world, OCR's role in bridging linguistic divides becomes increasingly vital.

Another compelling avenue beckons us towards the realm of OCR for Complex Documents. This endeavor involves the recognition of intricate and diverse documents that span a spectrum of fonts, layouts, and languages. Additionally, the integration of domain-specific knowledge promises to elevate OCR's accuracy, making it an invaluable tool across various specialized industries.

In this rapidly evolving digital landscape, Ethical Considerations are paramount. Ensuring privacy protection and bias mitigation is imperative to promote responsible deployment of OCR technology. As we leverage OCR for diverse applications, safeguarding individuals' privacy and addressing potential biases within the technology's algorithms is an ethical imperative that demands our unwavering attention.

In conclusion, OCR technology has made remarkable strides, transforming the way we

interact with textual information. As we navigate the challenges and explore the opportunities that lie ahead, OCR continues to be a powerful force for innovation, efficiency, and connectivity, shaping a future where the barriers between the visual and digital realms are progressively blurred.

References

[1] S. Kaur, P. Mann, and S. Khurana, "Page segmentation in ocr system - a review," in *Pattern Analysis and Machine Intelligence*, Jan. 2013.

[2] M. Chawla, R. Jain, and P. Nagrath, "Implementation of tesseract algorithm to extract text from different images," May 2020.

[3] A. B. Salah, J. p. Moreux, N. Ragot, and T. Paquet, "Ocr performance prediction using cross-ocr alignment," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 556–560, 2015.

[4] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open source ocr tool tesseract - a case study," Oct. 2017.

[5] R. Mittal and A. Garg, "Text extraction using ocr: A systematic review," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 357–362, 2020.

[6] e. a. Khalid, Samina, "A survey of feature selection and feature extraction techniques in machine learning," *2014 Science and Information Conference*, 2014.

[7] R. Saluja, M. Punjabi, M. Carman, G. Ramakrishnan, and P. Chaudhuri, "Sub-word embeddings for ocr corrections in highly fusional indic languages," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 160–165, 2019.

Appendices

Appendix A

Weekly Progress Report

D Y PATIL UNIVERSITY
RAMRAO ADIK INSTITUTE OF TECHNOLOGY, NAVI MUMBAI

Department of Computer Engineering
TE Mini-Project-III Weekly Project Performance Report Odd Sem 2023-2024

Project Title: _____ **Group No:**

Name of Students 1: **Name of Students 2:**
Name of Students 3: **Name of Students 4:**

Week No.	Expected Topics to be Covered	Progress Status	Student 1 Sign	Progress Status	Student 2 Sign	Progress Status	Student 3 Sign	Progress Status	Student 4 Sign	Suggestions if any
1.	Clear and Precise Objective									
2.	Abstract and Introduction									
3.	Literature Survey									
4.	Limitations of Existing System									
5.	Problem Definition / Statement									
6.	Proposed Methodology									
7.	System Design									
8.	Details of hardware & Software									
9.	Implementation details									
10.	Result Analysis									
11.	Conclusion and Future Work									
12.	Participation in Competition or Paper Publication									

A: Satisfactory **B: Average** **C: Needs Improvement**

Project Guide Name and Sign

Figure A.1: Weekly Progress Report

Appendix B

Plagiarism Report

Appendix C

Publication Details / Copyright / Project Competitions

C.1 Publications

[1] Ashish Sahu, Amit Javkar, Aman Sahay, Sujal Pokharkar, Mr. Gaurav Datkhile, Dr. Ekta Sarda, "OCR Copy Checker," *10.13140/RG.2.2.11517.54246*, 11, 2023.

Acknowledgments

I thank the many people who have done lots of nice things for me during the course of my OCR project. Your support, guidance, and encouragement have been invaluable. I want to express my deep appreciation to my project supervisor for their expertise and mentorship. I'm also grateful to my colleagues for their collaboration and insights. Special thanks to my guide for their unwavering support throughout this journey. Lastly, I acknowledge the contributions of the research community and the resources that have been instrumental in the success of this project.

Date: _____

OCR Copy Checker

ORIGINALITY REPORT

11%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.coursehero.com Internet	150 words — 4%
2	nemertes.library.upatras.gr Internet	89 words — 2%
3	zenodo.org Internet	41 words — 1%
4	hukumdev.blogspot.com Internet	17 words — < 1%
5	acmindia-studentchapters.in Internet	16 words — < 1%
6	core.ac.uk Internet	16 words — < 1%
7	shop4cf.eu Internet	14 words — < 1%
8	www.tdcommons.org Internet	14 words — < 1%
9	mirror.ufs.ac.za Internet	13 words — < 1%
10	www.klippa.com Internet	

11 words — < 1%

11	Advances in Intelligent Systems and Computing, 2014. Crossref	10 words — < 1%
12	"Document Analysis and Recognition – ICDAR 2021", Springer Science and Business Media LLC, 2021 Crossref	9 words — < 1%
13	"ICT Infrastructure and Computing", Springer Science and Business Media LLC, 2023 Crossref	9 words — < 1%
14	digital.library.ncat.edu Internet	9 words — < 1%
15	ebin.pub Internet	9 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	< 5 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 9 WORDS