A
Mini Project Report on

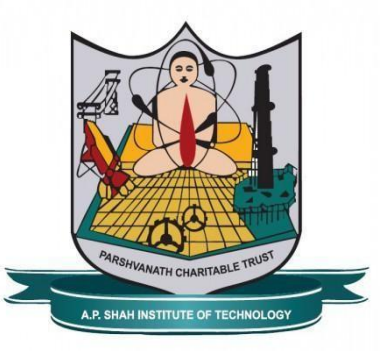
Ink to Insight using OCR

Submitted in partial fulfillment of the requirements
for the degree of
BACHELOR OF ENGINEERING
IN
Computer Science & Engineering
Artificial Intelligence & Machine Learning

by

Atharva Patil (22106039)
Shraavani Salunkhe (22106031)
Brahmjot Singh (22106004)
Aniruddha Pawar (22106009)
Under the guidance of

Prof. Nirali Arora



Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai
2024-2025

A. P. SHAH INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**Ink to Insight using OCR"** is a bonafide work of Atharva Patil (22106039), Shraavani Salunkhe (22106031), Brahmjot Singh (22106004), Aniruddha Pawar (22106009) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Nirali Arora	Dr. Jaya Gupta
Mini Project Guide	Head of Department

A. P. SHAH INSTITUTE OF TECHNOLOGY

Project Report Approval

This Mini project report entitled "**Ink to Insight using OCR"** by Atharva Patil, Shraavani Salunkhe, Brahmjot Singh, Aniruddha Pawar is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AI and ML) ***2024-25***.

External Examiner: _____

Internal Examiner: _____

Place: APSIT, Thane

Date:

Declaration

We declare that this written submission represents my ideas in my own words and whereothers' ideas or words have been included, I have adequately cited and referenced the original sources. 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 by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission hasnot been taken when needed.

Atharva Patil (22106039)	Shraavani Salunkhe (22106031)	Brahmjot Singh (22106004)	Aniruddha Pawar (22106009)

ABSTRACT

The "Ink to Insight" project tackles the challenge of transforming handwritten documents into structured knowledge using an integrated approach that combines Optical Character Recognition (OCR), image processing, machine learning, and Natural Language Processing (NLP). OCR converts handwritten text into machine-readable form, enhanced by image preprocessing with OpenCV and NumPy. Machine learning models, built using PyTorch, refine text recognition, while extractive and abstractive summarizations are generated using SpaCy and the T5 model from the Transformers library, respectively. This framework offers a comprehensive solution for digitizing and extracting actionable insights from handwritten documents.

Keywords: Optical Character Recognition (OCR), Image Processing, OpenCV, NumPy, Machine Learning, Natural Language Processing (NLP).

Index

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

In today's information-rich environment, the ability to swiftly and accurately convert handwritten documents into structured digital text is crucial for various applications, including archival, research, and automated data analysis. The traditional process of manually transcribing handwritten content is not only labor-intensive but also prone to errors. Our project aims to address these challenges by integrating advanced technologies to streamline the conversion of handwritten documents into actionable insights.

Project Motivation

Handwritten documents, ranging from historical manuscripts to modern notes, contain valuable information that often remains inaccessible due to the limitations of traditional digitization methods. Optical Character Recognition (OCR) has made significant strides in recent years, but its effectiveness can be hampered by the quality of the handwriting and the conditions under which the documents are captured. Furthermore, once the text is extracted, understanding and summarizing the content remains a complex task.

Project Overview

The "Ink to Insight" project is a comprehensive system designed to convert handwritten documents into text and subsequently summarize the content through two distinct approaches: extractive and abstractive summarization. This project involves several key stages:

<u>Image Processing and OCR:</u>

Preprocessing: Handwritten documents are first processed to enhance the quality of the images using techniques such as grayscale conversion and thresholding. This prepares the images for more accurate text extraction.

Text Extraction: Optical Character Recognition (OCR) is then applied to convert the processed images into machine-readable text. This step is crucial for converting various handwriting styles into a uniform digital format.

<u>Text Refinement with Machine Learning:</u>

Word Segmentation: The extracted text is further refined using machine learning models built with PyTorch. These models are trained on diverse datasets to handle various handwriting styles and improve text segmentation and recognition accuracy.

Summarization:

Extractive Summarization: Using SpaCy, the system identifies and extracts the most relevant sentences from the text. This approach provides a summary based on the prominence of certain terms and their frequency.

Abstractive Summarization: Leveraging the T5 model from the Transformers library, the system generates a concise and coherent summary of the text. This approach utilizes advanced language generation techniques to produce summaries that capture the essence of the content in a more natural and fluent manner.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1-HISTORY

1. Early Developments in Optical Character Recognition (OCR):

The concept of OCR dates back to the early 20th century, with the first notable attempts made in the 1920s. The initial systems were based on mechanical devices that could recognize printed characters. A significant breakthrough came in the 1950s with the development of electronic OCR systems capable of processing alphanumeric characters. The work of Rosenblatt and Kurtz (1951) in the development of pattern recognition algorithms laid the groundwork for modern OCR technologies.

2. Progress in Handwritten Text Recognition:

Handwritten text recognition posed a greater challenge due to the variability in individual handwriting styles. Early efforts, such as those by Schaeffer (1960), focused on simple pattern matching techniques. The 1980s and 1990s saw the advent of more sophisticated methods, including feature extraction and statistical pattern recognition, which improved accuracy but were still limited in handling diverse handwriting styles.

3. Machine Learning and Deep Learning Era:

The early 2000s marked the shift from traditional pattern recognition to machine learning-based approaches. The introduction of Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) significantly improved the accuracy of OCR systems. The advent of deep learning in the 2010s, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), revolutionized OCR by enabling systems to learn from large datasets and generalize better across different handwriting style

2.2-LITERATURE REVIEW

The "Ink to Insight" project leverages advanced technologies to convert handwritten documents into digital text and generate insightful summaries. This endeavor integrates Optical Character Recognition (OCR), image processing, machine learning, and text summarization to address the challenges of handling handwritten text. To understand the current state of these technologies and their evolution, it is essential to review the historical development and recent advancements in each area.

[1] **Research on English Character Recognition Technology Based on OCR**

Published in: 2023 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)

We developed an English translation system that leverages Optical Character Recognition (OCR) technology for enhanced image processing. The system employs the OTSU algorithm for image binarization, converting images to black and white to simplify character recognition. To improve image quality and reduce noise, we use median and Gaussian filters for denoising. OpenCV with Python is utilized for accurate text recognition, with seamless integration enabled by PyCharm. Systematic testing confirms that the preprocessing techniques combined with OCR processing result in high accuracy in recognizing English characters, demonstrating the system's effectiveness and efficiency.

[2] **A font style classification system for English OCR**

Published in: 2017 International Conference on Intelligent Computing and Control (I2C2)

This paper addresses the challenge of reducing computational complexity in font style/size independent Optical Character Recognition (OCR) systems. We propose a method for classifying font styles by analyzing distance profile features in left, right, and diagonal directions of character images. This approach aims to simplify generic OCR systems by incorporating font style recognition. Using a support vector machine classifier, we tested our technique on a dataset of 10 common fonts for both upper and lower case letters. Our experiments, conducted on character images from non-editable documents with 5 different font styles, achieved a satisfactory accuracy of 80%, demonstrating the effectiveness of our method in reducing OCR complexity.

[3] **Image processing based degraded camera captured document enhancement for improved OCR accuracy**

Published in: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)

Over the past decade, camera-based document analysis has gained significant attention due to the widespread use of smartphones. Capturing documents with phone cameras is convenient, but often results in low-quality images due to factors like blur, uneven illumination, perspective distortion, and low resolution, which can impact OCR accuracy. Quality enhancement techniques can help mitigate these issues, making the text more readable and improving recognition. This paper evaluates the effectiveness of various deblurring techniques for enhancing the quality of noisy and blurred camera-captured document images.

[4] **Enhancement of segmentation and zoning to improve the accuracy of handwritten character recognition**

Published in: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)

Handwritten character recognition enables machines to automatically recognize characters written in a user's language. This project focuses on recognizing handwritten English cursive words using optical character recognition (OCR) and neural networks. A scanned image of the handwritten word is converted into equivalent printed characters. To enhance accuracy and address the limitations of existing OCR algorithms, a new combination algorithm is developed for segmentation and zoning processes. This approach aims to improve the recognition of handwritten text and the overall performance of OCR systems.

[5] **NLP Based Automated Text Summarization and Translation: A Comprehensive Analysis**

Published in: 2022 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)

Extractive text summarization is key in natural language processing for condensing large texts while retaining essential information. This study introduces a Python-based GUI application using TK inter for document summarization, incorporating advanced image processing and

face recognition through OpenCV and PIL for secure user authentication. By leveraging NLP, the system achieves accuracy between 91% to 95% for text summarization and language translation. Efficient algorithms allow users to extract key sentences quickly, enhancing comprehension. The combination of text summarization and security measures offers a professional solution for modern document management and information processing needs.

[6] **Abstractive Text Summarization Models Using Machine Learning Algorithms**

Published in: 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)

The rapid growth of e-documents and large data in fields like Data Mining, Information Retrieval, and Natural Language Processing has made Automated Text Summarization essential. Text Summarization reduces the size of a document while preserving its key information and meaning. Manual summarization of vast documents is impractical, highlighting the need to modernize abstractive summarization methods. This paper surveys current models, algorithms, and techniques for Abstractive Text Summarization, focusing on the Encoder-Decoder framework. The study provides insights into recent machine learning algorithms, addressing challenges and future prospects, and aims to alleviate issues like email overload and media monitoring

CHAPTER 3

Problem Statement

3. Problem Statement

Handwritten documents pose challenges in modern processing due to variable text styles, complex image preprocessing needs, and limitations in effective summarization. Conventional OCR struggles with accuracy due to inconsistencies in handwriting, while preprocessing techniques like noise reduction and binarization are essential but difficult to optimize. Additionally, summarization of extracted text is complicated by the nuances of handwritten content, with extractive methods often missing context and abstractive models requiring sophisticated language understanding.

CHAPTER 4

Experimental Setup

4. Experimental Setup

4.1 Hardware Setup

Laptop with the following configuration is recommended:

- **Processor (CPU):** Intel Core i7 (10th Gen or later) or AMD Ryzen 7 to handle image processing, OCR, and machine learning tasks efficiently.
- **Memory (RAM):** 16 GB or more to support multiple processes such as image preprocessing, text extraction, and summarization.
- **Storage:** 512 GB SSD or higher for fast data access, especially when handling large datasets and model checkpoints.
- **Operating System:** Windows 10/11.

4.2 Software Setup

The following software tools and libraries are required for the "Ink to Insight" project:

- **Programming Language:** Python 3.8 or later
- **IDE:** Visual Studio Code or PyCharm (optional for ease of development)

Packages and Libraries:

1. **Image Processing and Data Handling:**
    - **Matplotlib:** For visualizing images, data, and results.
    - **Numpy:** For numerical computations and handling arrays during image processing.
    - **OpenCV-Python:** For image processing tasks such as grayscale conversion and thresholding.

2. **Machine Learning and NLP:**
    - **PyTorch:** For building, training, and fine-tuning models to improve text recognition and summarization.
    - **Scikit-learn:** For additional machine learning utilities such as data preprocessing and model evaluation.
    - **SpaCy:** For extractive summarization using sentence selection based on word frequency and relevance.
    - **Transformers:** For abstractive summarization, using models like T5 for natural language generation.

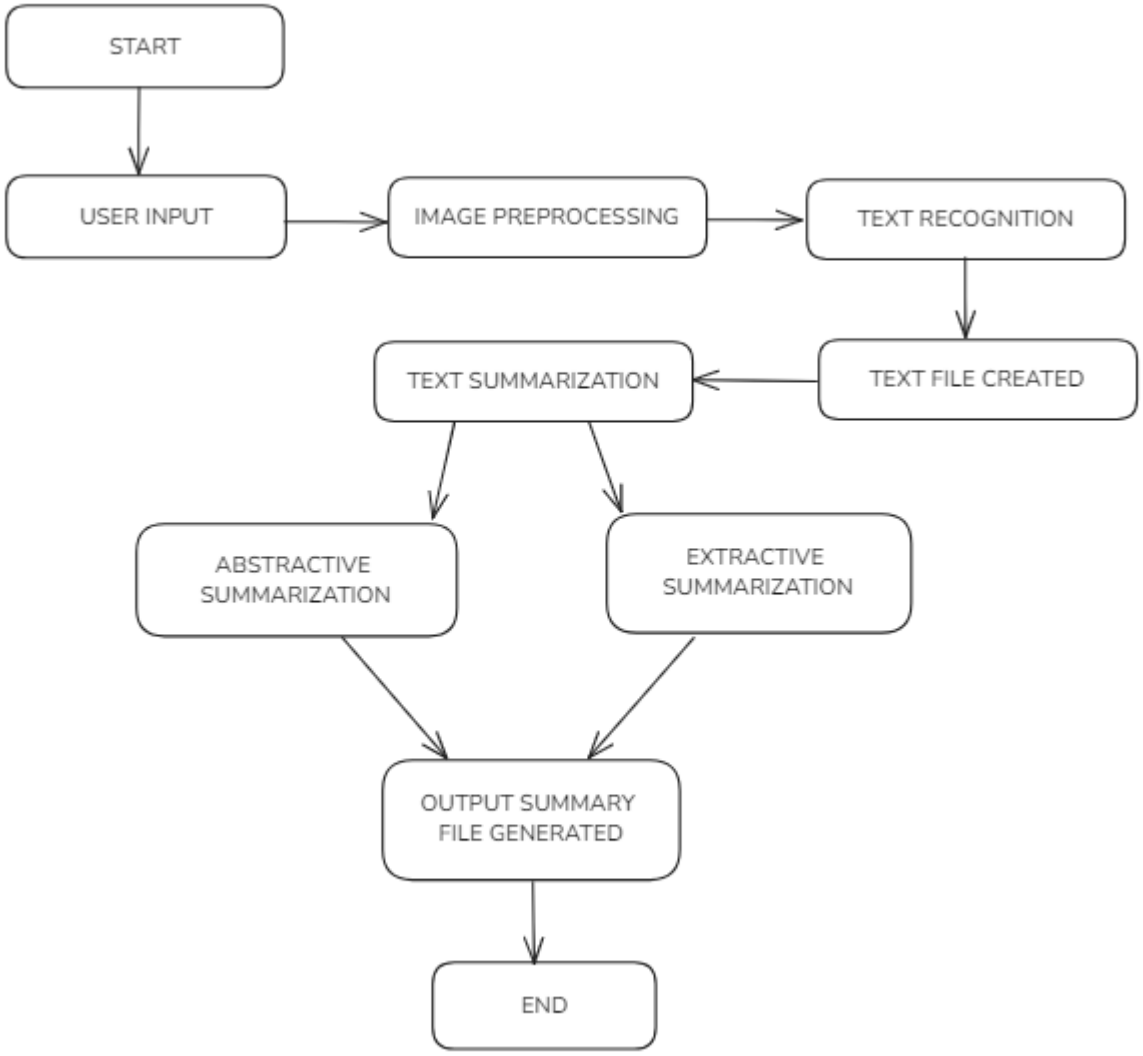
3. **Testing:**
    - **Pytest:** For testing various components of the project to ensure proper functionality.
4. **File Management:**
    - **Path:** For efficient file path handling during image and text file processing.

CHAPTER 5

Proposed System & Implementation

5. Proposed system and Implementation

5.1 Block diagram of proposed system



5.2 Description of block diagram

1. Start and User Input:

The process begins with the user providing input in the form of handwritten documents, either scanned or photographed for processing.

2. Image Preprocessing: The system applies image preprocessing techniques like noise reduction, and thresholding to improve the clarity and prepare the image for Optical Character Recognition (OCR). High resolution of document is created and word detection is done using scikit learn model.

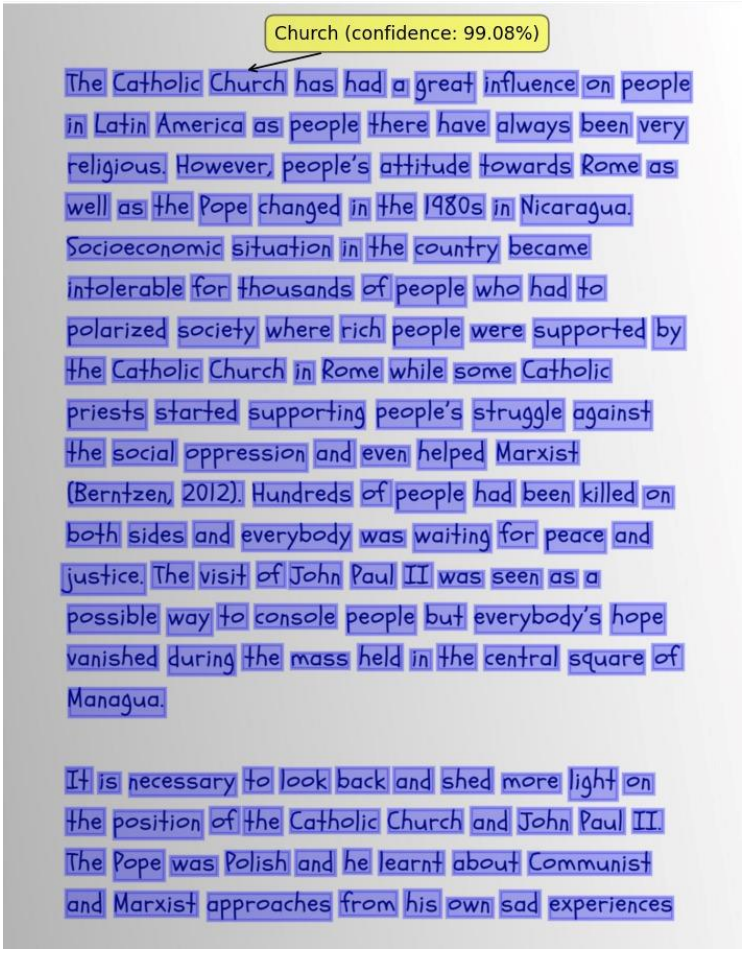
3. Text Detection and Recognition (OCR): Using OCR, the system converts the preprocessed image into machine-readable text. This stage involves recognizing various handwriting styles and extracting words via machine learning models from the words detected.

4. Text file created: Here all the text detected will be passed in a file and a text file will be created of it.

5. Summarization: the text undergoes summarization using two approaches:
Extractive Summarization: Key sentences are extracted using the SpaCy library to generate a condensed version based on term prominence.
Abstractive Summarization: The system utilizes the T5 model from Transformers to generate a coherent, concise summary that captures the main ideas of the text.
Output Summary:

6. Output summary file generated: The final summary is prepared and passed into a file for user access.

5.3 Implementation




```
The Catholic Church has had a great
influence on people in Latin America as
people there have always been very religious.
However, people's attitude towards Rome as well
as the Pope changed in the 1980s
in Nicaragua. Socioeconomic situation in the country
became intolerable for thousands of people who
had to polarized society where rich people
were supported by the Catholic Church in
Rome while some Catholic priests started Supporting
people's struggle against the social oppression and
even helped Marxist (Berntzen, 2012). Hundreds of
people had been killed on both sides
and everybody was waiting for peace and
justice. The visit of John Paul II
was seen as a possible way to
console people but everybody's hope vanished during
the mass held in the central square
of Managua. I+ is necessary to look
back and shed more light on the
```

Socioeconomic situation in the country became intolerable for thousands of people who had to polarized society where rich people were supported by the Catholic Church in Rome while some Catholic priests started Supporting people's struggle against the social oppression and even helped Marxist (Berntzen, 2012). Nonetheless, people of Nicaragua did not +hink of political paradigm, they wanted peace and consolation which could be given by a few words about the victims of the oppressive regime in the Pope's mass (Riding 1983). The Catholic Church has had a great influence on people in Latin America as people there have always been very religious. The visit of John Paul II was seen as a possible way to console people but everybody's hope vanished during the mass held in the central square of Managua. The rich obtained support of one of the most powerful institutions in the country and exploited +his in their attempts to Suppress people's resistance. Hundreds of people had been killed on both sides and everybody was waiting for peace and justice. Of course, the price of +his Success was very high and it cost the country hundreds of people's lives.

the Catholic Church has had a great influence on people in Latin America. people's attitude to Rome as well as the Pope changed in the 1980sin Nicaragua - and even helped Marxist'supportingpeople' in their struggle against the social oppression. if they were to polarized society, they would have been able to support people who had been supported by the church inRome. but they had to be reformed and sworn in to help them.

CHAPTER 6

Conclusion

6. Conclusion

The "Ink to Insight" project demonstrates the potential of integrating advanced image processing, machine learning, and natural language processing techniques to transform handwritten documents into actionable insights. It automates the conversion of handwritten text into structured digital format, followed by extractive and abstractive summarization, providing a streamlined solution for diverse handwritten content. Utilizing OCR with PyTorch for text refinement and SpaCy with the T5 model for summarization, the project effectively addresses challenges like handwriting variability and document quality. This robust, scalable approach reduces the time and effort of manual transcription while enabling deeper analysis of handwritten materials. "Ink to Insight" marks a significant advancement in document digitization and content analysis, with future enhancements aimed at improving adaptability to complex handwriting, expanding training datasets, and refining summarization through user feedback.

References

Research papers:

[1] X. Li, Z. Zhan, C. Li, C. Guo, Z. Li and J. Li, "Research on English Character Recognition Technology Based on OCR," 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2024, pp. 1210-1213, doi: 10.1109/IMCEC59810.2024.10575683.

[2] V. Bharath and N. S. Rani, "A font style classification system for English OCR," *2017 International Conference on Intelligent Computing and Control (I2C2)*, Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/I2C2.2017.8321962.

[3] P. Sharma and S. Sharma, "Image processing based degraded camera captured document enhancement for improved OCR accuracy," *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, India, 2016, pp. 441-444, doi: 10.1109/CONFLUENCE.2016.7508160.

[4] S. E. Benita Galaxy and S. S. Ebenezer, "Enhancement of segmentation and zoning to improve the accuracy of handwritten character recognition," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 2016, pp. 4732-4735, doi: 10.1109/ICEEOT.2016.7755617.

[5] N. Zade, G. Mate, K. Kishor, N. Rane and M. Jete, "NLP Based Automated Text Summarization and Translation: A Comprehensive Analysis," *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2024, pp. 528-531, doi: 10.1109/ICSCSS60660.2024.10624907.

[6] K. G. Widowati, N. Budiman, K. Foejiono and K. Purwandari, "Abstractive Text Summarization Using BERT for Feature Extraction and Seq2Seq Model for Summary Generation," *2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, Karawang, Indonesia, 2023, pp. 226-230, doi: 10.1109/ICMERALDA60125.2023.10458190.