AUTOMATED TEXT EXTRACTION FROM IMAGES USING OCR IN UIPATH

A PROJECT REPORT

Submitted by

LOKESHWAR S (220701146)

in partial fulfillment for the course

OAI1903 - INTRODUCTION TO ROBOTIC PROCESS AUTOMATION

for the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND
ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE RAJALAKSHMI
NAGAR THANDALAM CHENNAI – 602 105

NOVEMBER 2024

BONAFIDE CERTIFICATE

Certified that this project report **"Automated Text Extraction from images using OCR in Uipath"** is the bonafide work of **"LOKESHWAR S (220701146)"** who carried out the project work for the subject OAI1903 - Introduction to Robotic Process Automation under my supervision.

SIGNATURE

Mrs. G. M. Sasikala, M.E.,
SUPERVISOR,
Assistant Professor, Department of
Computer Science and Engineering,
Rajalakshmi Engineering College,
Rajalakshmi Nagar,
Thandalam, Chennai-
602105.

Submitted to Project and Viva Voce Examination for the subject OAI1903 - Introduction to Robotic Process Automation held on ___________.

Internal Examiner **External Examiner**

ABSTRACT

This project focuses on automating the extraction of text from images using Optical Character Recognition (OCR) technology within UiPath. Many organizations deal with large volumes of image-based data, such as scanned documents, invoices, and receipts, which are difficult to process manually. This project provides a solution by automating the conversion of image content into editable text.

The system uses UiPath's OCR activities, including Tesseract OCR, Google OCR, and Microsoft OCR, to accurately extract text from various image formats like PNG, JPEG, and PDF. It also includes features for improving image quality, such as resizing and noise reduction, to enhance OCR accuracy. The extracted text can be saved in different formats, such as Excel or text files, or integrated with other systems for further use.

This solution reduces manual effort, improves accuracy, and speeds up data processing, making it ideal for industries like finance, healthcare, and logistics. It provides a fast, reliable, and cost-effective way to handle large volumes of unstructured data, enhancing operational efficiency and productivity.

ACKNOWLEDGEMENT

I express my sincere thanks to my beloved and honourable chairman MR.S.MEGANATHAN and the chairperson DR.M.THANGAM MEGANATHAN for their timely support and encouragement.

I am greatly indebted to my respected and honourable principal Dr. S.N.MURUGESAN for his able support and guidance.

No words of gratitude will suffice for the unquestioning support extended to us by my head of the department Dr. P. KUMAR for being ever supporting force during my project work.

I also extend my sincere and hearty thanks to my internal guide Mrs. G. M. SASIKALA for her valuable guidance and motivation during the completion of this project.

My sincere thanks to my family members, friends and other staff members of Computer Science and Engineering

LOKESHWAR S (220701146)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	ACCRONYM
RPA	Robotic Process Automation
AI	Artificial Intelligence
API	Application Programming Interface
CV	Computer Vision
OCR	Optical Character Recognition

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In today's digital era, businesses and organizations frequently encounter vast amounts of unstructured data stored in image formats, such as scanned documents, receipts, invoices, and forms. Extracting meaningful information from these images manually is time-consuming, error-prone, and inefficient. To address this challenge, advancements in automation and Optical Character Recognition (OCR) technology have paved the way for faster, more accurate, and reliable data extraction solutions.

This project, Automated Text Extraction from Images using OCR in UiPath, leverages UiPath's powerful automation tools and OCR capabilities to streamline the process of converting image-based text into editable and structured formats. By automating this workflow, organizations can significantly reduce manual labor, improve data accuracy, and enhance productivity.

The system is designed to handle diverse image types, including low-quality and complex documents, through advanced pre-processing techniques and multiple OCR engines. It not only extracts text but also integrates with various applications to store and utilize the extracted data efficiently. This solution is highly adaptable and can be deployed across different industries to automate repetitive tasks, minimize errors, and optimize operational processes.

Through this project, we aim to demonstrate how UiPath can simplify and automate text extraction, transforming image-based data into a valuable resource for decision-making and analysis.

1.2 OBJECTIVE

The primary objective of this project is to automate the extraction of textual information from image-based documents using Optical Character Recognition (OCR) technology in UiPath. The goal is to streamline the process of converting unstructured data from images, such as scanned documents, receipts, invoices, and handwritten notes, into structured, editable formats.

This automation aims to:

1. **Reduce Manual Effort**: Eliminate the need for manual data entry by automating the text extraction process.
2. **Improve Accuracy**: Minimize human errors associated with manual text extraction.
3. **Enhance Efficiency**: Speed up the processing and handling of large volumes of image-based data.
4. **Ensure Scalability**: Provide a solution that can handle diverse document types and volumes across different industries.
5. **Facilitate Integration**: Enable seamless integration of extracted data into various applications and systems for further processing and analysis.

1.3 EXISTING SYSTEM

In many organizations, the process of extracting text from image-based documents is still performed manually. Employees are required to review scanned documents, receipts, invoices, and other image-based data, then manually input the information into digital formats such as spreadsheets or databases. This traditional approach is highly labor-intensive, timeconsuming, and prone to human error.

Some organizations may use basic OCR software to automate parts of the process. However, these standalone OCR tools often lack integration capabilities, require significant manual intervention for preprocessing, and may struggle with complex documents containing mixed content, lowquality images, or handwritten text.

Key limitations of existing systems include:

1. **High Manual Effort**: Manual text entry or verification is still required, increasing the workload.
2. **Error-Prone Processes**: Manual handling of large volumes of data increases the likelihood of mistakes.
3. **Lack of Integration**: Standalone OCR tools are often not integrated with existing business workflows or systems, limiting their efficiency.
4. **Limited Scalability**: Existing systems may not effectively handle high volumes of documents or diverse input formats.
5. **Inconsistent Accuracy**: Poor image quality, complex layouts, and handwritten text often result in inaccurate OCR outputs.

These limitations create a need for a more robust, integrated, and automated solution that can efficiently handle large-scale text extraction tasks with minimal manual intervention.

1.4 PROPOSED SYSTEM

The proposed system aims to revolutionize the process of text extraction from image-based documents by leveraging UiPath’s automation capabilities and advanced Optical Character Recognition (OCR) technology. Unlike traditional methods, which rely heavily on manual input or standalone OCR tools, this system offers a fully automated workflow that efficiently converts image content into structured, editable formats. By integrating multiple OCR engines, such as Tesseract, Google OCR, and Microsoft OCR, the system ensures high accuracy and flexibility in handling various document types, including printed and handwritten text.

To enhance the accuracy of text extraction, the system incorporates preprocessing techniques like resizing, noise reduction, and grayscale conversion, which optimize image quality before processing. It also features robust error-handling mechanisms that detect low-confidence outputs, allowing for validation and correction when needed. The extracted text can be seamlessly integrated into other systems, such as databases and reporting tools, making it accessible for further analysis and decision-making. Additionally, the system is highly scalable, capable of processing large volumes of data across multiple formats, making it adaptable to various industries and organizational needs. This solution not only reduces manual effort and human error but also improves efficiency, accuracy, and costeffectiveness in handling image-based data.

CHAPTER 2

LITERATURE REVIEW

2.1 OCR Technology and Applications:

OCR technology, which converts printed or handwritten text in images into machine-encoded text, has evolved significantly over the last few decades. Early OCR systems were limited in scope and could only recognize text in specific fonts or on high-quality images. However, with the advent of machine learning and deep learning techniques, modern OCR engines can now process a wide range of fonts, handwriting, and document types. Several OCR tools, such as Tesseract, Google OCR, and Microsoft OCR, have gained popularity due to their accuracy and open-source or commercial availability. Tesseract, in particular, has been widely used due to its flexibility, support for multiple languages, and continuous improvement through contributions from the open-source community.

2.2 Challenges in OCR Systems:

Despite the advances in OCR technology, several challenges remain. One of the primary challenges is dealing with low-quality images, such as blurry scans or photographs with noise. The performance of OCR systems tends to degrade when handling documents with poor resolution, skewed text, or complicated layouts. Researchers have proposed various image preprocessing techniques, such as noise reduction, skew correction, and thresholding, to improve OCR accuracy in these cases. Additionally, handwritten text poses another significant challenge, as it is more variable and less structured than printed text, requiring more advanced machine learning models to achieve satisfactory accuracy levels.

2.3 Integration of OCR in Automation Tools

The integration of OCR technology within automation frameworks like UiPath has become an emerging area of interest. UiPath provides built-in OCR activities that allow users to incorporate text extraction capabilities within their automation workflows. Studies have shown that combining OCR with robotic process automation (RPA) can significantly enhance business processes by reducing manual data entry, improving operational efficiency, and ensuring faster decision-making. The combination of OCR with RPA tools allows for automated data extraction from documents such as invoices, receipts, and contracts, which are often stored as image files, and can then be processed for further analysis or storage in databases.

2.4 Post-OCR Text Processing

Once text is extracted from images, it often requires additional processing to clean, validate, and structure the data for further use. Research in this area has focused on natural language processing (NLP) techniques that help identify and extract key information, such as names, addresses, and dates, from large volumes of text. This post-processing is essential in real-world applications like invoice processing or form recognition, where the raw OCR output may contain errors or inconsistencies. Several studies highlight the need for error-handling mechanisms that can flag low-confidence OCR results and provide opportunities for manual validation, ensuring higher quality and reliability of the final data.

2.5 OCR in Industry Applications

OCR technology has found diverse applications in industries ranging from healthcare and finance to logistics and government. In healthcare, OCR is used to extract information from handwritten patient records, prescriptions, and medical forms. In finance, OCR enables automated invoice processing and accounts payable workflows, while in logistics, it helps with digitizing shipping labels and product barcodes. Researchers have also explored the use of OCR for historical document digitization, enabling the preservation of valuable records and making them searchable.

2.6 Future Trends and Advancements

The future of OCR technology is closely tied to advancements in deep learning and artificial intelligence (AI). Researchers are exploring the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to improve the accuracy of OCR systems, especially for complex documents. The integration of AI-powered models with OCR is expected to improve text recognition in noisy, distorted, or low-resolution images, making OCR even more robust and reliable across diverse use cases. Additionally, the use of cloudbased OCR services is growing, as they offer scalability and ease of integration with various business applications, allowing organizations to deploy OCR without the need for heavy infrastructure investments.

CHAPTER 3

SYSTEM DESIGN

3.1.1 SYSTEM FLOW DIAGRAM

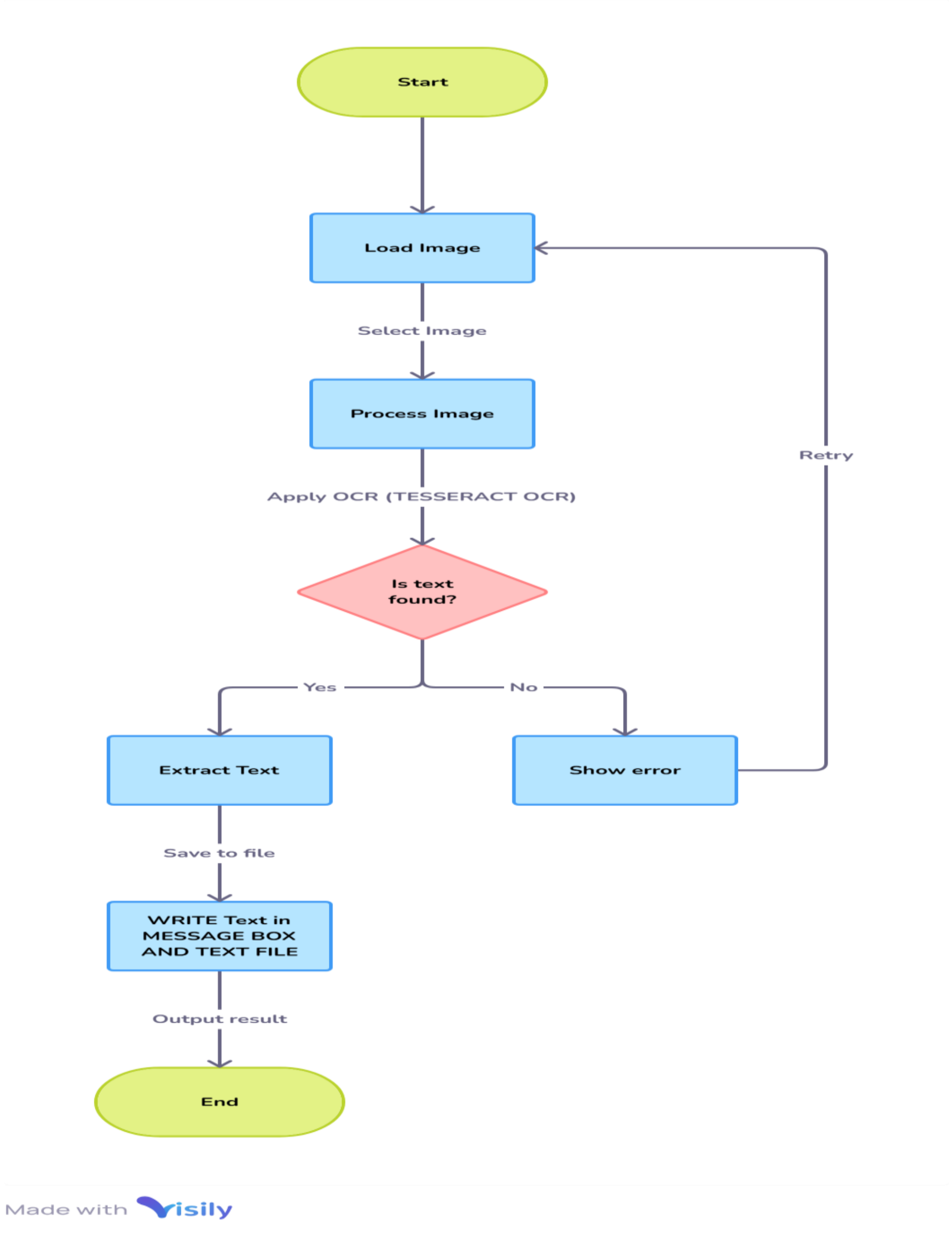


Fig 3.1.1 System Flow Diagram

3.1.2 ARCHITECTURE DIAGRAM

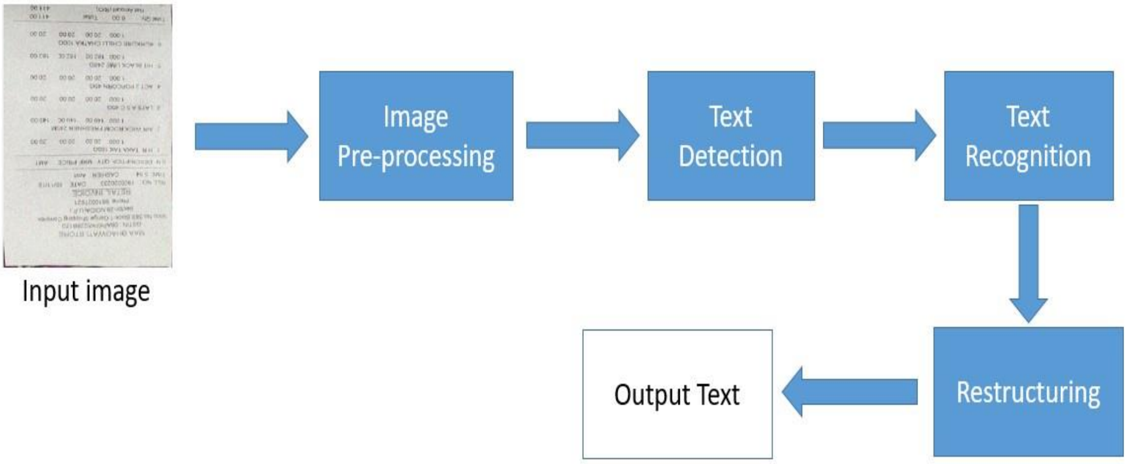


Fig 3.1.2 Architecture Diagram

3.1.3 SEQUENCE DIAGRAM

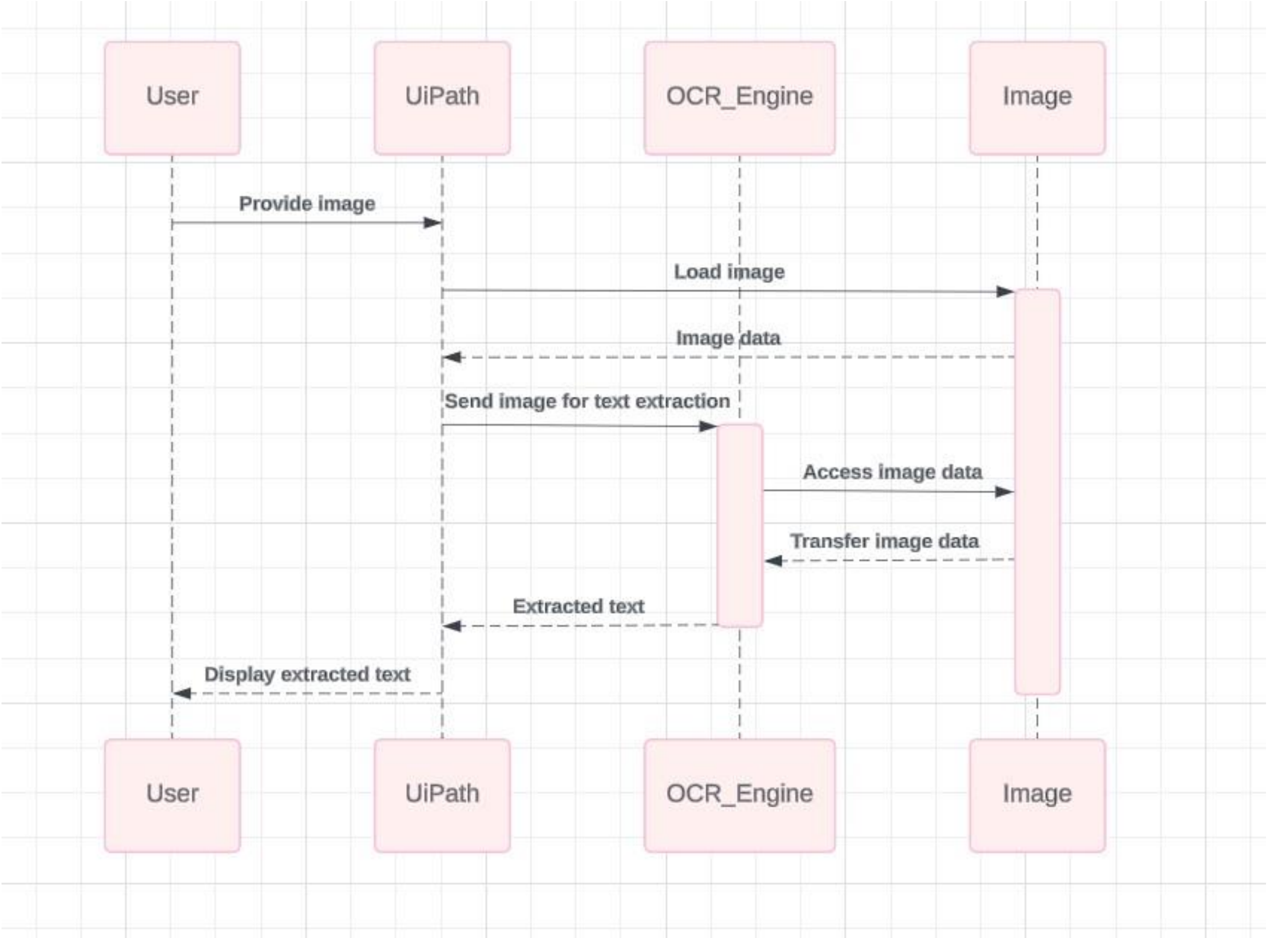


Fig 3.1.3 Sequence Diagram

CHAPTER 4

PROJECT DESCRIPTION

4.1. METHODOLOGIES:

In the proposed automated text extraction project, the following methodologies have been applied to streamline the process of extracting text from images and automating the typing of extracted text into a Notepad file.

1. Image Upload using Load Image:
The process begins by using the **Load Image** activity, which allows users to select an image file from the local system. This step is essential for obtaining the image containing text that needs to be extracted. By integrating this activity, the system can process images stored on the local machine, enabling flexibility in handling various image file types such as PNG, JPEG, or PDF.

2. Text Extraction with Tesseract OCR:
Once the image is loaded, the **OCR Activity** with **Tesseract OCR** is utilized to extract the text from the image. Tesseract OCR is an open-source and highly reliable optical character recognition engine that converts the visual content of the image into machine-readable text. The OCR activity reads the image file and identifies any printed or handwritten text present in the image. This method provides a high degree of accuracy in text extraction, especially for images with standard fonts or clear handwriting.

3. Displaying Extracted Text via Message Box:
After extracting the text from the image, the **Message Box** activity is used to display the extracted content. This step allows the user to verify the accuracy of the extracted text before proceeding further in the automation process. It serves as a confirmation point for the system's output and gives users the opportunity to make adjustments if necessary.

4. Writing the Text into Notepad using Application Browser and Type Into Activity:

To automate the input of the extracted text, the **Application Browser** activity is employed to focus the Notepad application on the screen. This activity ensures that the automation targets the correct window where the text will be typed. After establishing the target application, the **Type Into** activity is used to automatically type the extracted text into the open Notepad window. This eliminates the need for manual entry, thereby enhancing productivity and ensuring that the text is inputted exactly as it was extracted.

5. Completion Confirmation with Message Box:

The final step in the process is to display a **Message Box** indicating that the task is complete. This message serves as an end notification for the user, confirming that the entire text extraction and input process has been successfully carried out. The use of the message box adds clarity to the automation, providing users with an acknowledgment that the system has finished its task.

4.2 MODULES DESCRIPTION:

The project is divided into several distinct modules that handle different aspects of the process, ensuring smooth execution and automation of the entire task. These modules include:

1. Image Upload and Load Module

The first module is responsible for loading the image file from the user's local system. Using the Load Image activity, the system prompts the user to select an image from their file system. This module allows the flexibility to process different image formats and serves as the initial step in the automation workflow.

2. Text Extraction Module

Once the image is loaded, the OCR Activity with Tesseract OCR engine processes the image to extract the text. This module handles the core functionality of converting image-based text into a machine-readable format. The accuracy of this module is key, as it ensures the extracted text is as close to the original content as possible, minimizing errors in subsequent steps.

3. Text Display Module

After the text has been extracted, this module leverages the Message Box activity to display the extracted content to the user. This module provides an opportunity for users to review and verify the text before further automation, allowing for quick validation and error-checking.

4. Notepad Input Automation Module

This module is responsible for automating the process of typing the extracted text into Notepad. The Application Browser activity is used to bring the Notepad window into focus, ensuring that the text is typed in the correct application. The Type Into activity follows to simulate the typing process, ensuring that the extracted content appears in the Notepad window exactly as it was recognized.

5. Completion Confirmation Module

The final module displays a Message Box to inform the user that the task is complete. This module serves as a confirmation that the automation workflow has been successfully executed, and no further actions are required. It ensures a clean end to the process and provides a visual cue to the user that the automation has finished.

CHAPTER 5

CONCLUSION

The automated text extraction from images project successfully demonstrates the application of Optical Character Recognition (OCR) and Robotic Process Automation (RPA) to streamline the process of extracting and inputting text into a digital format. By leveraging UiPath's **Load Image**, **OCR Activity (Tesseract)**, and **Type Into** functionalities, the system efficiently handles tasks that would otherwise require manual intervention.

The integration of these technologies allows for improved productivity, enhanced accuracy, and the elimination of repetitive tasks. Additionally, by automating the process of extracting text and typing it into Notepad, the solution reduces human error and ensures consistency in data entry. This project highlights the potential of combining OCR with RPA to automate complex workflows, ultimately contributing to more efficient document processing and data management.

REFERENCES

1. Extract Text From Images Using Machine Learning Overview:

- **Source: https://addepto.com/blog/text-extraction-from-imagesusing-machine-learning/**

2. Image Preprocessing Techniques for OCR

Image preprocessing plays a critical role in enhancing OCR accuracy by improving the quality of the input images. This article discusses various preprocessing methods, such as noise reduction, binarization, and skew correction.

- Source: "Image Preprocessing Techniques for OCR" by ResearchGate
- URL: ResearchGate - OCR Image Preprocessing

3. Microsoft OCR Documentation

This is the official documentation for Microsoft OCR, another OCR engine used in the project. It offers details on integrating Microsoft's OCR services for text extraction.

- Source: "What is OCR and How Does it Work?" by TechTarget
- URL: Microsoft OCR Source: Microsoft OCR

4. Tesseract OCR Documentation:

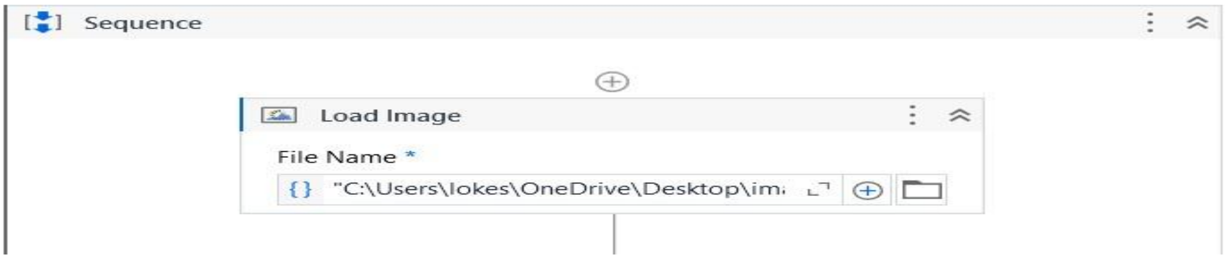
Tesseract OCR is an open-source optical character recognition engine used in this project for text extraction from images. It is widely used for recognizing printed and handwritten text in scanned documents or images.

- **Source:** Tesseract OCR

5. Overview of Text Extraction from Images using Uipath:

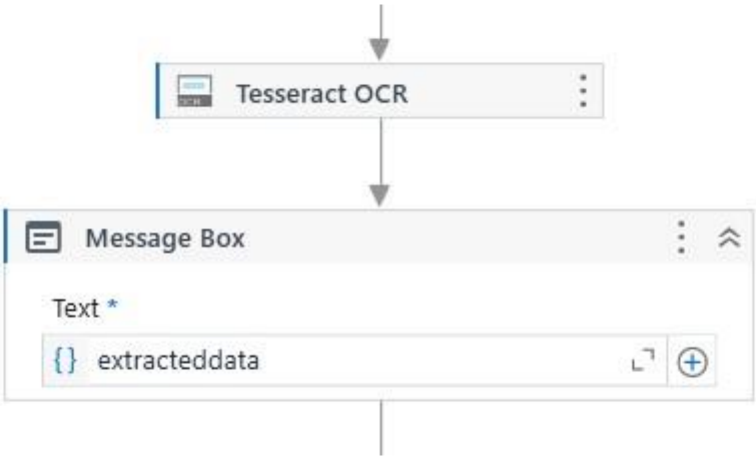
- **Source** :https://excelcult.com/how-to-extract-text-from-imageusing-ocr-in-uipath/

OUTPUT SCREENSHOTS



"C:\Users\lokes\OneDrive\Desktop\im

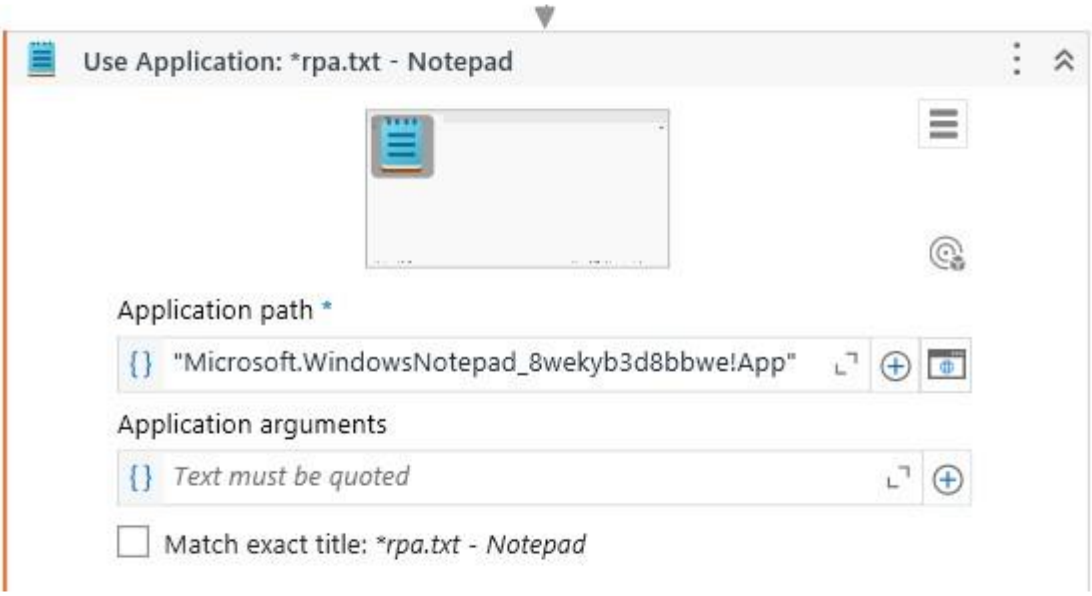
Fig 5.1 Load Image activity



Text *

{} extracteddata

Fig 5.2 Tesseract OCR Activity with Message box to display the extracted text



Application path *

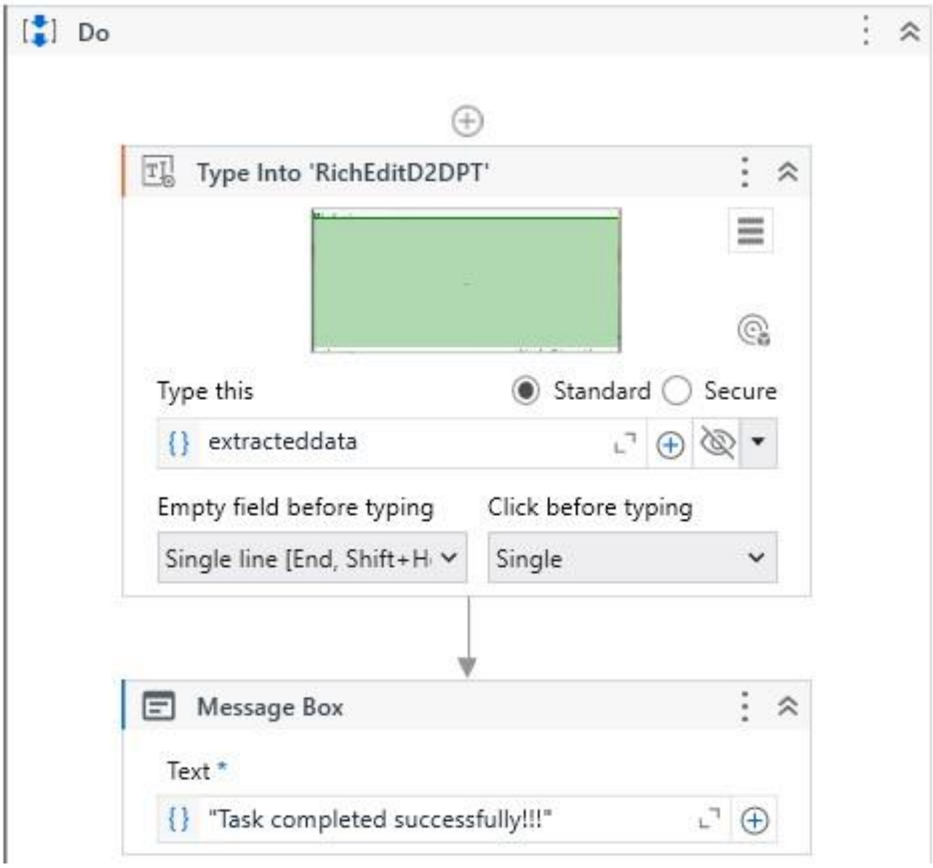
{} "Microsoft.WindowsNotepad_8wekyb3d8bbwe!App"

Application arguments

{} *Text must be quoted*

□ Match exact title: **rpa.txt - Notepad*

Fig 5.3 Open Browser activity to make a target on Notepad file



Do

Type Into 'RichEditD2DPT'

Type this: ● Standard ○ Secure

extracteddata

Empty field before typing: Single line [End, Shift+H

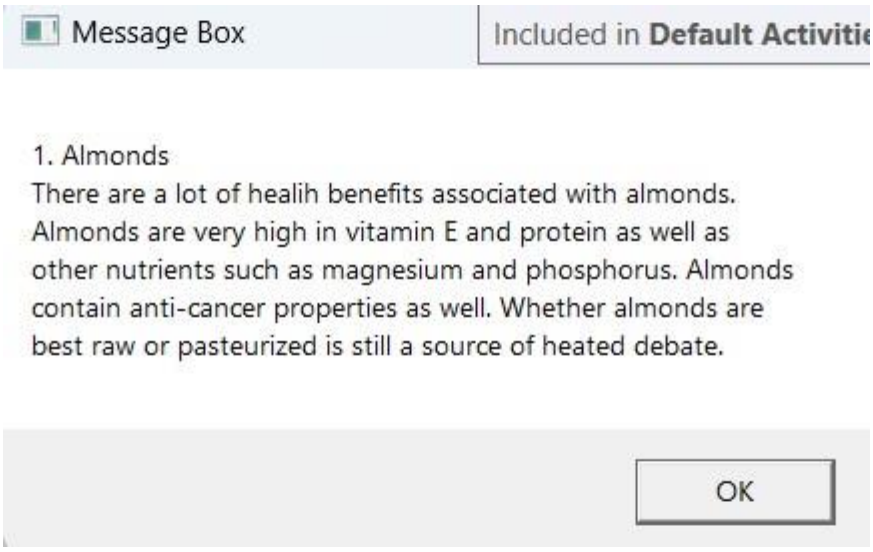
Click before typing: Single

Message Box

Text *

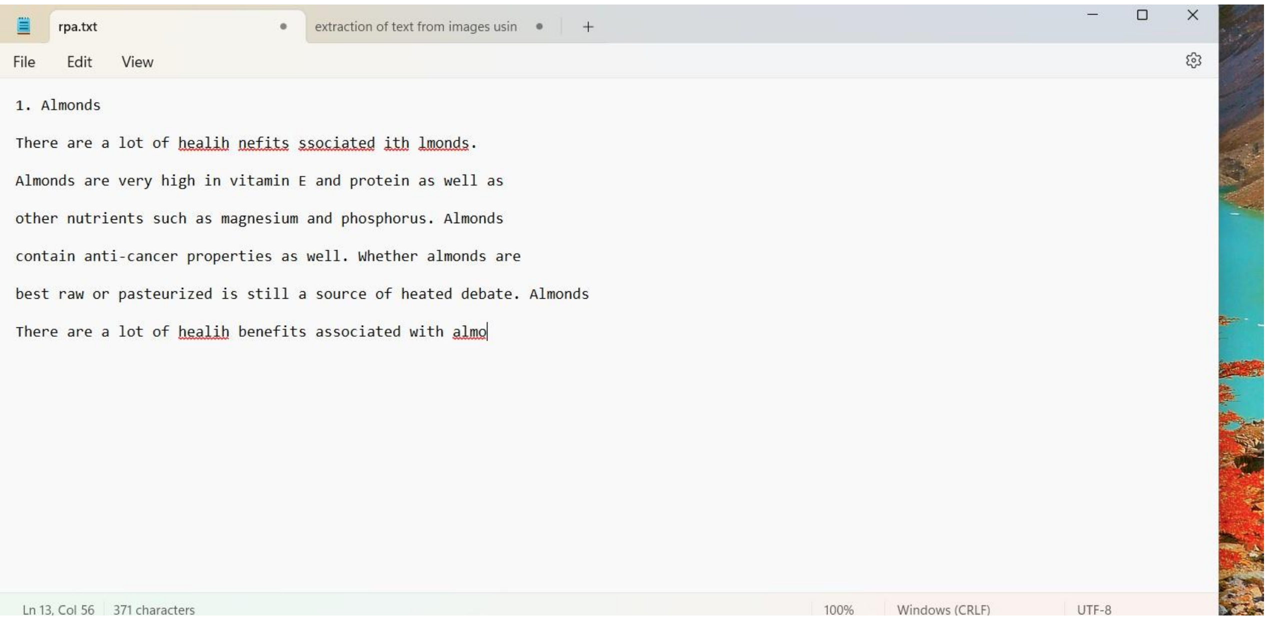
"Task completed successfully!!!"

Fig 5.4 Do section of Open Browser with Type into activity and Message box



1. Almonds
There are a lot of healih benefits associated with almonds.
Almonds are very high in vitamin E and protein as well as
other nutrients such as magnesium and phosphorus. Almonds
contain anti-cancer properties as well. Whether almonds are
best raw or pasteurized is still a source of heated debate.

Fig 5.5 Message box activity with extracted text from input image



```
1. Almonds
There are a lot of healih nefits ssociated ith lmonds.
Almonds are very high in vitamin E and protein as well as
other nutrients such as magnesium and phosphorus. Almonds
contain anti-cancer properties as well. Whether almonds are
best raw or pasteurized is still a source of heated debate. Almonds
There are a lot of healih benefits associated with almo
```

Fig 5.6 Notepad file with the content of extracted text

APPENDICES

The appendices section includes supplementary information that supports the main content of the project report. These may include raw data, code snippets, additional explanations, or detailed instructions that are referenced in the body of the report.

Appendix A: Sample Workflow

This section contains the UiPath workflow used in the project. It outlines the steps involved in loading an image, performing OCR, and typing the extracted text into Notepad.

- **Step 1**: Use the **Load Image** activity to import an image file from the local system.
- **Step 2**: Apply the **Tesseract OCR** activity to extract text from the image.
- **Step 3**: Use a **Message Box** activity to display the extracted text.
- **Step 4**: Open **Notepad** using the **Application/Browser** activity and perform the **Type Into** activity to type the extracted text into Notepad. •
- **Step 5**: Use another **Message Box** to indicate the task completion.

Appendix B: UiPath Activities Used

This section lists and describes the activities used in the workflow:

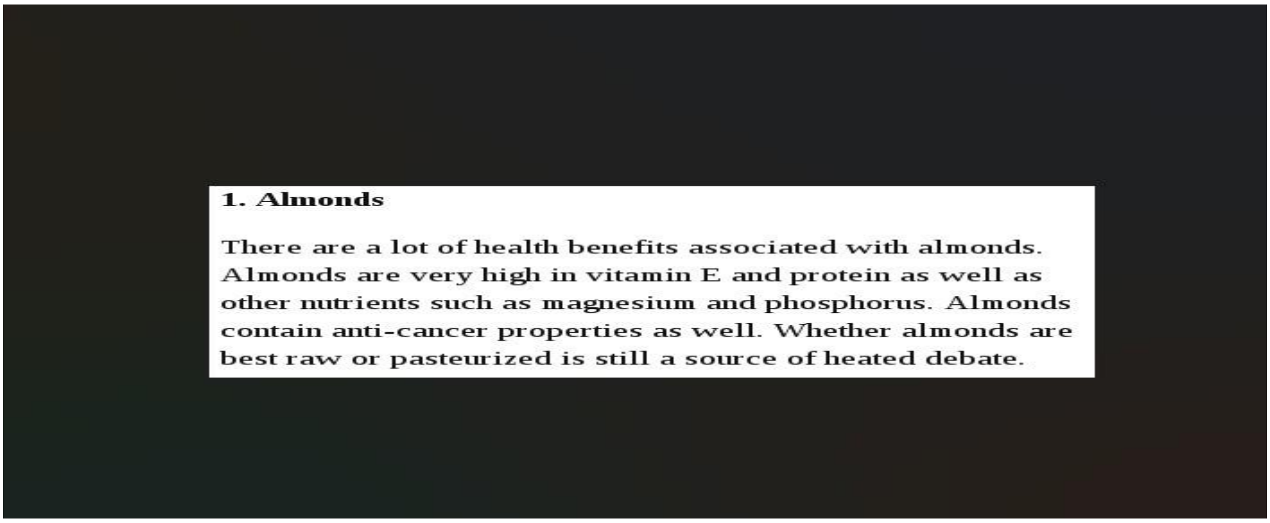
- **Load Image**: Used to load the image file from the specified path.
- **Tesseract OCR**: Extracts text from the image using the Tesseract engine.
- **Message Box**: Displays extracted text for validation.
- **Application/Browser**: Opens Notepad for typing the extracted content.
- **Type Into**: Types the extracted text into Notepad.

- **Message Box**: Displays a message indicating that the task has been completed successfully.

Appendix C: Sample OCR Output

This section shows a sample of the OCR text extracted from an image. For example, if the input image contained the text “extractedtext”, the OCR output might be displayed as follows:

- **Input Image**: A screenshot of text in a document or image.



1. Almonds

There are a lot of health benefits associated with almonds. Almonds are very high in vitamin E and protein as well as other nutrients such as magnesium and phosphorus. Almonds contain anti-cancer properties as well. Whether almonds are best raw or pasteurized is still a source of heated debate.

Appendix D: Code Snippets

In this appendix, relevant code snippets from the UiPath workflow are provided. For example: ○ **Load Image("C:\Users\Documents\image.jpg")**

○ **Perform OCR using Tesseract ○ OCR Text = Tesseract-OCR("C:\Users\Documents\image.jpg")**

○ **Display the extracted text in a message box ○ Message Box (OCR Text)**

○ **Type the extracted text into Notepad ○ Type-Into("Notepad", OCR Text)**

Appendix E: System Requirements

This section lists the hardware and software requirements necessary to run the project successfully.

- **Hardware**:
  ○ Processor: Intel Core i3 or higher ○ RAM: 4GB or more ○ Disk Space: 2GB free space
- **Software**:
  ○ **UiPath Studio**: Version 2023 or higher ○ **Tesseract OCR**: Installed and configured ○ **Windows Operating System**: Windows 10 or higher ○ **.NET Framework**: Version 4.7.2 or higher