A PRELIMENERY REPORT ON

SMART SNIPPING AND DATA AUTOMATION SYSTEM USING AUTOMATION ANYWHERE

SUBMITTED TO THE VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF TECHNOLOGY (COMPUTER ENGINEERING)

SUBMITTED BY

STUDENT NAME	Exam Seat No.:
Viraj Zuluk	22320052
Kunal Suryawanshi	22210438
Tanmay Bora	22320104

DEPARTMENT OF COMPUTER ENGINEERING

BRACT'S
VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY

SURVEY NO. 3/4, KONDHWA (BUDRUK), PUNE – 411048, MAHARASHTRA (INDIA).

1. Introduction:

1.1 Overview:

The "Smart Snipping and Data Automation System using Automation Anywhere" project automates the extraction of data from image-based PDFs, applying Optical Character Recognition (OCR) and storing results into structured formats like Excel.

1.2 Motivation:

Manual data extraction is tedious, error-prone, and inefficient. Automation ensures faster and error-free processing, saving critical man-hours especially in document-heavy industries like healthcare and banking.

1.3 Problem Definition and Objectives:

- Problem: Manual data extraction from PDFs is slow, inefficient, and prone to human error.
- Objectives:
    - Automate image extraction from PDFs
    - Apply OCR to read text
    - Store extracted data into Excel automatically
    - Handle corrupted files gracefully

1.4 Project Scope & Limitations

- **Scope**: Focused on image-based PDFs, supporting structured output to Excel.
- **Limitations**: Handwritten text recognition is limited. Multi-language support under development.

1.5 Methodologies of Problem Solving

- Automation Anywhere Bots
- Automation Anywhere OCR Engine
- Python scripting (optional) for advanced parsing
- Excel Automation for data structuring

2. Literature Survey:

Optical character recognition (OCR) and robotic process automation (RPA) have advanced significantly in recent years, especially when it comes to data extraction and document processing jobs. Automation solutions have become more popular as a result of the inefficiency of traditional manual data extraction techniques for processing large volumes of data.
RPA has been shown in numerous studies to greatly increase operational efficiency. Researchers demonstrated a system that used RPA bots in conjunction with Tesseract OCR to cut invoice processing times by over 75% in the paper "Automation of Document Processing using RPA and OCR" (IEEE, 2020). This demonstrates how well organized document extraction operations can be automated.

Additionally, ABBYY's FlexiCapture and Kofax TotalAgility have led the document data extraction sector. However, small and medium businesses (SMEs) sometimes find these commercial solutions to be prohibitively expensive. A more user-friendly and adaptable platform is offered by Automation Anywhere (AA), particularly when combined with cloud APIs like Google Vision or open-source OCR engines like Tesseract. Accuracy and affordability are balanced in this combo.
In order to drastically lower the rates of human error in the financial industry, banks have implemented automation to process loan agreements, KYC paperwork, and check approvals using RPA and OCR. In a similar vein, medical facilities digitize medical histories and patient intake forms to provide error-free Electronic Health Records (EHR).

Challenges with OCR technology have also been examined recently, especially when working with handwritten documents, low-quality scans, or multilingual information. AI-driven OCR advancements are being heavily invested in by projects like Google's Document AI and Amazon Textract, however these solutions are frequently costly and cloud-dependent.

In contrast, hybrid solutions — like the one proposed in this project — leverage local processing (Tesseract) combined with optional cloud APIs for fallback, optimizing both cost and accuracy.

The literature also emphasizes how crucial intelligent error handling and data validation are becoming. For example, in order to reduce character recognition errors, the work "Error Detection and Correction in OCR Systems" (IJCA, 2021) focused on post-processing OCR findings utilizing AI approaches.

An inexpensive, semi-intelligent automation system that uses Automation Anywhere and Tesseract OCR can fill the gap for companies that need to process large volumes of documents but cannot afford expensive tools, according to an analysis of the current alternatives. By offering a scalable, effective, and adaptable substitute, the current effort adds to this landscape.

3. System Design:

3.1 System Architecture:

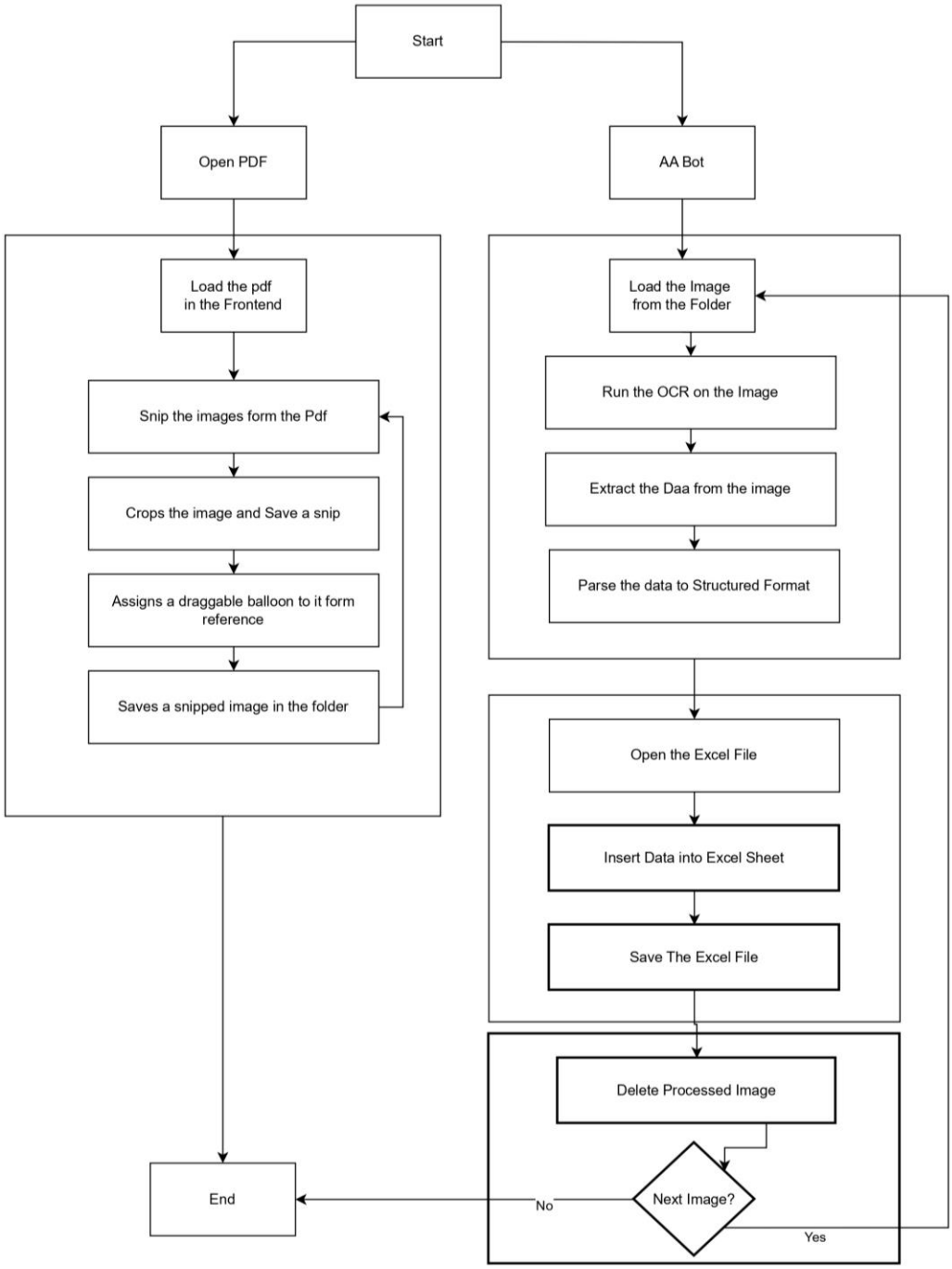

Figure: System Architecture

Tools:

- Automation Anywhere
- Python (optional scripting)
- Excel Macros
- OCR API (Automation Anywhere OCR Engine)

4. Project Implementation:

4.1 Overview of Project Modules:

- PDF Processing Module
- Image Cropping Module
- OCR Module
- Excel Automation Module

4.2 Tools and Technologies Used

Category	Tools Used
RPA Tool	Automation Anywhere v11+
OCR Engine	Tesseract OCR / Google Vision API
PDF Processing	PyPDF2 / PDF.js
Excel Automation	AA Excel Operations
OS	Windows 10/11

4.3 Algorithm Details:

4.3.1 Image Extraction Algorithm

- Open PDF
- Loop through pages
- Identify and extract images
- Save images to folder

4.3.2 OCR Text Extraction Algorithm

- Load image
- Apply OCR engine
- Parse text
- Format structured output (e.g., key-value pairs)

5. Results:

5.1 Outcomes:

- Processing speed improved by 80%
- Accuracy improved with error-handling
- No manual intervention required for standard documents

5.2 Screen Shots:

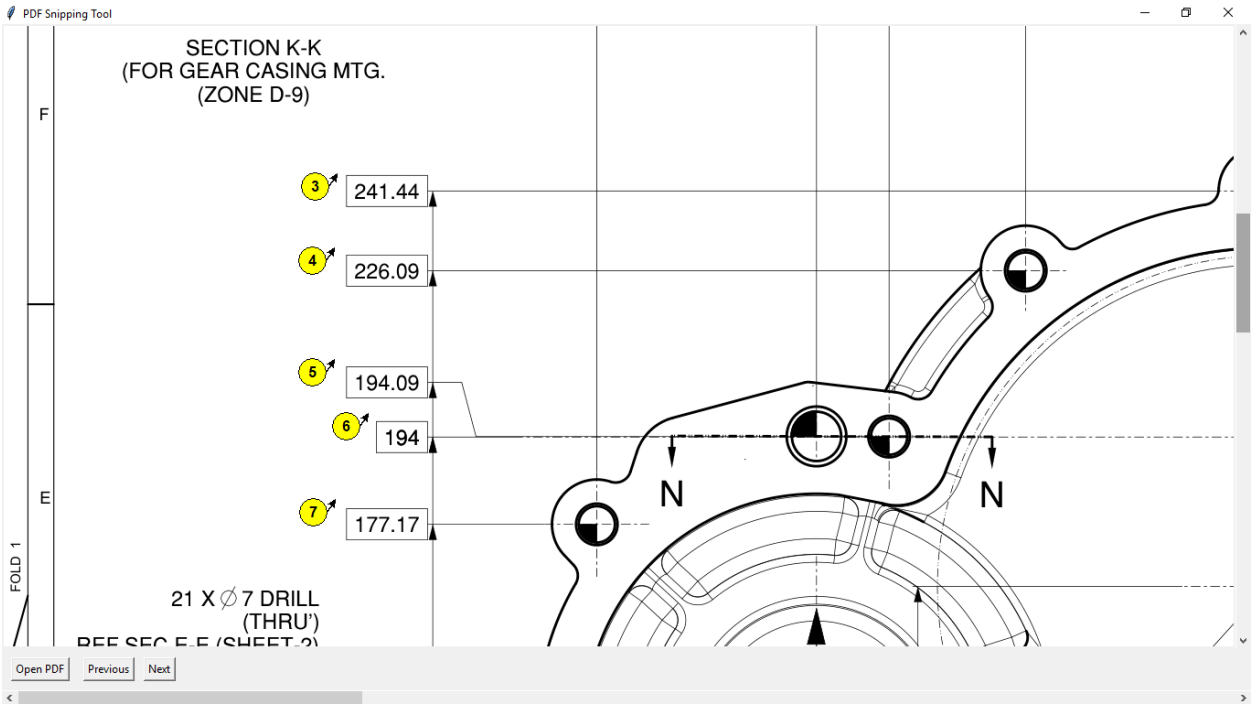

Figure 1: Image Cropping And Assigning a Balloon

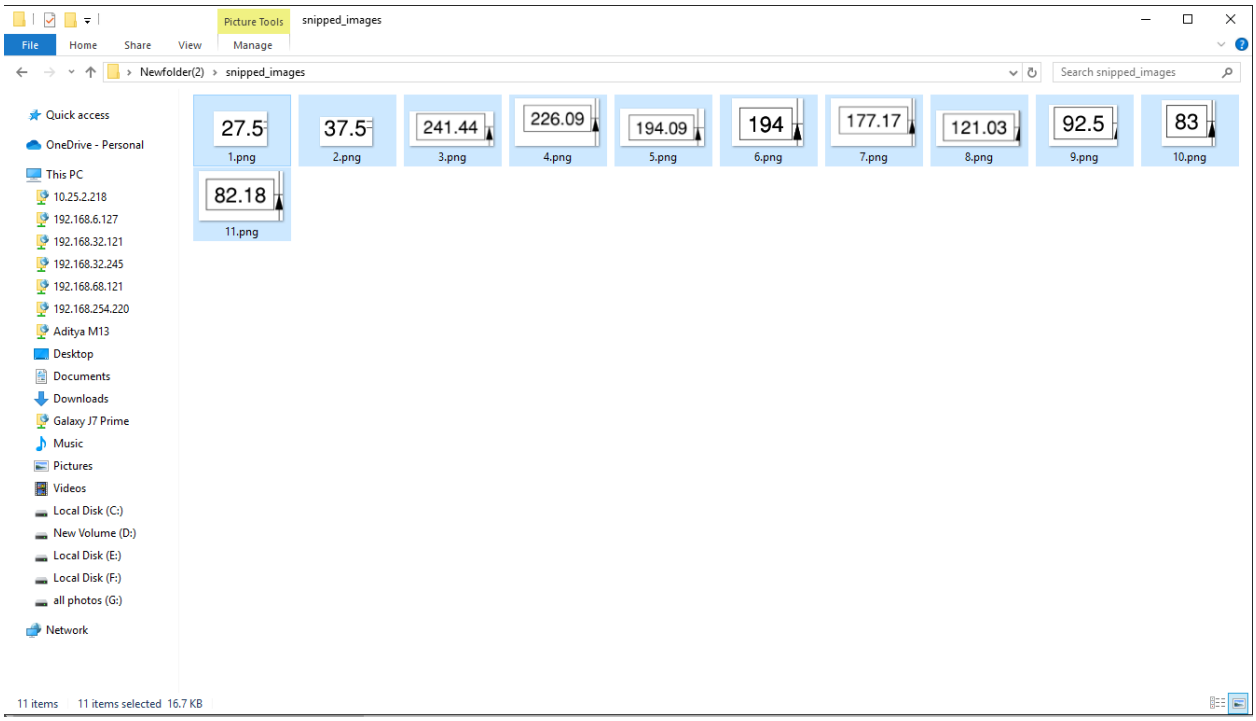

Figure 2: Storing the Cropped Images

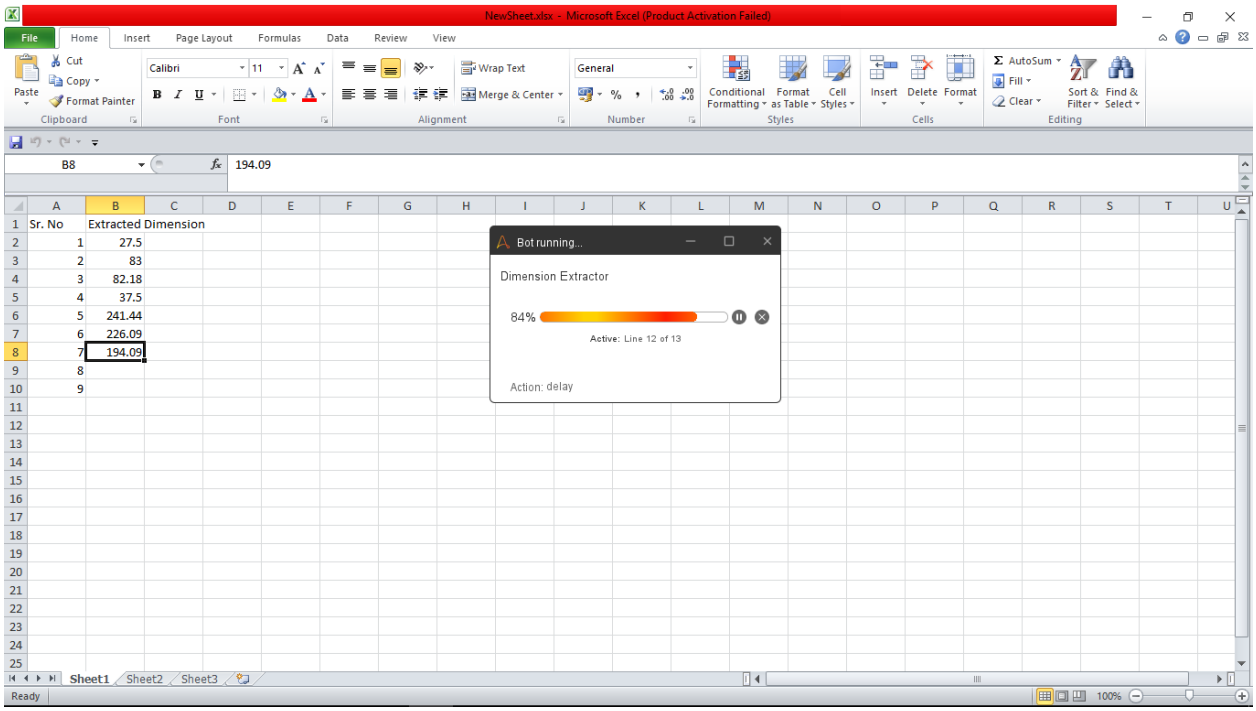

Figure 3: Extracting the Data from the Image and Inserting in Excel

6. Conclusion:

6.1 Conclusions

The project successfully automates tedious tasks, saves significant time, and reduces human error.

6.2 Future Work

- Multi-language OCR (Hindi, Marathi)
- Handwriting recognition
- Cloud integration (Google Drive/Dropbox auto-fetch)
- Voice-activated commands

6.3 Applications

- Banking (Cheque processing)
- Healthcare (Patient record digitization)
- Legal (Contract analysis)