International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Integrated OCR-Based Platform for Intelligent Document Processing and Text Recognition

Ms. Kavyashree K L[1], Ms. Aishwarya D Hebbar[2], Ms. Chaithra R[3], Ms. Namrata Dhapte S[4]

[1]Assistant Professor of Computer Science, MMK & SDM MMV, Mysore
[2,3,4]BCA Students, Department of Computer Science, MMK & SDM Mahila Mahavidyalaya, Mysore

Abstract
The proposed project presents a smart digital solution for converting printed and scanned documents into editable text through image processing and recognition techniques. In many workplaces and academic environments, handling document data manually still consumes significant time and often leads to inaccuracies. Existing applications mainly concentrate on simple text conversion and lack advanced analytical and user-support features. To address this shortcoming, the developed system combines text recognition with intelligent document handling capabilities in a unified environment. The application accepts both images and PDF files, improves document quality before processing, recognizes textual content accurately, and supports additional facilities such as document categorization, text searching, highlighted results, accuracy evaluation, and downloadable reports. Built using Python-based technologies and interactive frameworks, the system offers a reliable, fast, and convenient approach for modern document processing while minimizing human effort and improving overall efficiency.

Keywords: Optical Character Recognition, Document Categorization, Image Processing, PDF Processing, Text Extraction, Text Recognition.

1. Introduction

In the modern digital environment, the volume of paper-based and scanned information is increasing rapidly across educational institutions, corporate offices, and government organizations. Despite the shift toward digital systems, many users still depend on manual methods to read, store, and manage document content. This approach is not only time-consuming but also leads to reduced efficiency and higher chances of errors during information handling. Therefore, there is a growing demand for automated systems that can simplify document processing and improve accessibility of textual data. One of the widely used technologies for addressing this problem is Optical Character Recognition (OCR), which enables the conversion of visual document content into editable digital text. Although several OCR-based tools exist in the market, most of them are limited to basic conversion tasks and do not provide advanced functionalities that support intelligent document understanding. As a result, users often need separate tools for searching content, organizing files, or analyzing extracted information, which creates fragmentation in the workflow. To overcome these limitations, this project introduces an integrated system that enhances

traditional OCR capabilities with additional intelligent features. The system is designed to process both image-based and PDF documents efficiently by improving input quality before text extraction to achieve better accuracy. After extraction, it supports structured organization of content, fast text retrieval through search functions, visual highlighting of relevant data, and evaluation of extraction reliability. It also enables users to download processed outputs in a convenient format. By combining recognition, processing, and analytical functionalities into a single platform, the proposed system reduces dependency on multiple tools and minimizes manual intervention. This results in a more efficient, accurate, and user-friendly approach for handling digital documents in real-world scenarios.

2. Literature Survey

Mayank Deshmukh, Saloni Rabde, Priyanka Makode, Sourabh Jasuja, and Prof. Bhavesh Khasdev [1] presented a detailed study on text detection and extraction from images and PDF documents. Their research explained the importance of OCR technology and reviewed both traditional OCR techniques and modern deep learning-based methods used for document analysis. The study mainly focused on preprocessing, text detection, text recognition, and post-processing techniques used to improve extraction accuracy. They also discussed existing challenges such as multilingual text recognition, layout preservation, low-quality image handling, and computational complexity in OCR systems.

Prof. Anuradha Thorat, Mayur Zagade, Shivani More, Manish Pasalkar, and Anand Narute [2] presented a study on OCR-based text extraction from images and scanned documents. Their work focused on improving text recognition accuracy using preprocessing techniques such as grayscale conversion, noise reduction, and binarization. The paper also explained the use of CNN and OCR models for detecting and recognizing text from complex images. Additionally, the authors discussed system implementation, testing, and future improvements for handling handwritten and degraded text more effectively.

Jyothi E, K. Tejaswini, Lakshmi Chintalapati, and Mr. MD. Shafiulla [3] presented a system for extracting text from scanned images and documents using Optical Character Recognition (OCR) technology. Their work focused on converting image-based text into editable and searchable digital content by using preprocessing techniques such as grayscale conversion and image scaling. The authors also explained the use of Tesseract OCR with LSTMbased recognition for improving text extraction accuracy from different types of document images. In addition, the study highlighted the advantages of automated text extraction in reducing manual effort, saving time, and improving document management efficiency.

Shrinath Janvalkar, Paresh Manjrekar, Sarvesh Pawar, and Prof. Laxman Naik [4] discussed the importance of Optical Character Recognition (OCR) in converting printed text from images into editable and machine-readable formats. Their work explained different generations of OCR systems and described major OCR processes such as image acquisition, preprocessing, segmentation, feature extraction, classification, and post-processing. The paper also highlighted various real-world applications of OCR in healthcare, banking, legal industries, handwriting recognition, and automatic number plate detection. In addition, the authors emphasized that improving training data and preprocessing techniques can further enhance OCR accuracy and overall system performance.

Jyoti Wadmare, Sunita Ravindra Patil, Dakshita Kolte, Kapil Bhatia, Palak Desai, and Ganesh Wadmare [5] developed an advanced OCR tool that combines computer vision and natural language processing techniques for accurate text extraction and recognition from images. Their work focused on improving OCR performance using technologies such as OpenCV, Pytesseract, spaCy, Named Entity Recognition (NER), and edge detection algorithms. The proposed system was capable of extracting text, identifying

entities, highlighting important information, and storing the extracted content in Excel format with an overall accuracy of 98.8%. The study also discussed the importance of OCR in document digitization, automation, accessibility, and intelligent data management across various industries.

3. Methodology

The proposed system is designed to automate document digitization, text extraction, and document analysis from image and PDF files. The system integrates OCR technology, image preprocessing, text analysis, and dashboard visualization into a unified web-based platform developed using Python and Dash. Multiple libraries such as OpenCV, Pillow, NumPy, PDF2Image, Pytesseract, and Regular Expressions are used to improve processing efficiency, OCR accuracy, and document handling capabilities.

Figure 1: Block Diagram of System Architecture

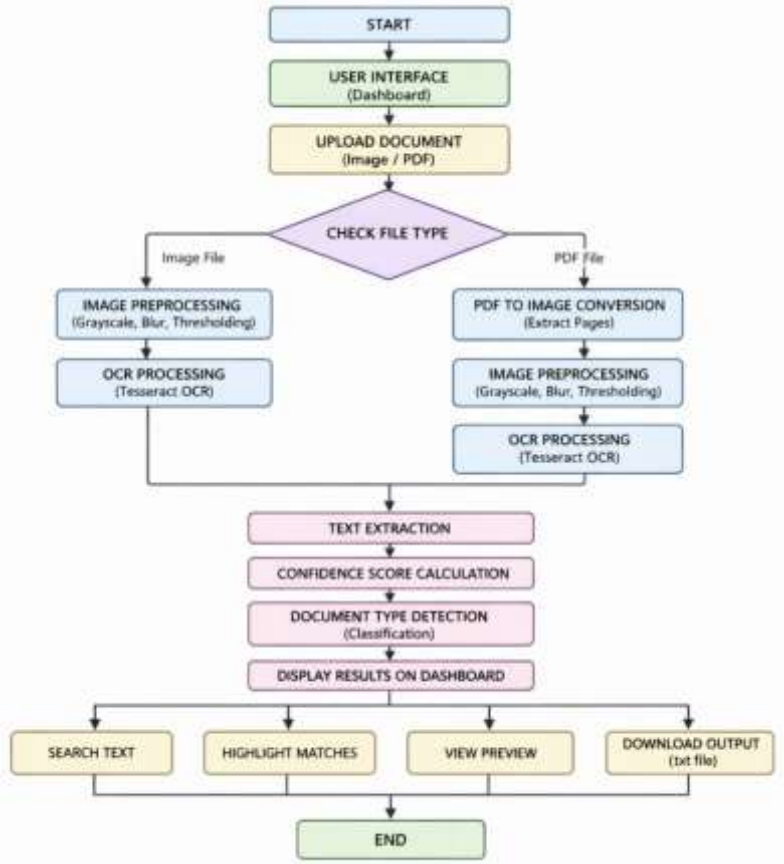

Fig: Block diagram of System Architecture

A. Document Input and Processing: The system accepts image and PDF documents uploaded through the Dash-based interactive dashboard. Supported file formats include JPG, PNG, and PDF. PDF documents are converted into image format using the PDF2Image library so that all documents can be processed uniformly during OCR execution. The uploaded files are temporarily processed in real time without permanent storage. Dash components and callback functions manage file upload operations, user interaction, and dynamic content updates within the web interface.

B. Image Preprocessing: Before text extraction, preprocessing operations are performed using OpenCV, Pillow, and NumPy libraries to improve document quality and OCR accuracy. Image preprocessing helps reduce background noise and enhances character visibility in scanned or low-quality documents.

The Preprocessing stage includes:

- Grayscale conversion for simplified image representation.
- Gaussian blur for noise reduction.
- Thresholding for separating text from background.
- Image enhancement and resizing for improved character clarity.

NumPy is used for efficient matrix and image array manipulation during preprocessing operations. These techniques improve OCR performance and reduce extraction errors.

C. **OCR-Based Text Extraction:** The preprocessed images are passed to the Tesseract OCR engine through the Pytesseract library for text recognition. The OCR module extracts textual content from documents and converts it into editable machine-readable text.

The OCR process involves:

- Character and word recognition from images.
- Line-by-line text extraction.
- Generation of OCR confidence scores.

The extracted text is formatted into structured paragraphs using Python string-processing methods to improve readability and organization.

D. **Text Analysis and Document Classification:** After OCR extraction, the system performs document analysis using keyword-based matching techniques and Regular Expressions (Regex). The extracted content is analyzed to identify document types such as invoices, certificates, reports, resumes, and identity proofs.

The analysis module also calculates:

- Word count
- Line count
- OCR confidence percentage

Regular Expressions are used for pattern matching, keyword identification, and efficient text filtering during document classification and search operations.

E. **Search and Highlighting Mechanism:** The system includes a keyword search and highlighting feature that allows users to search for specific words within the extracted text. Matching keywords are dynamically highlighted to improve visibility and accessibility of important information.

F. **Dashboard Integration and Output Generation:** All processing modules are integrated into an interactive dashboard developed using the Dash framework. HTML components are used for interface structuring, CSS styling is applied for layout and visual presentation, and JavaScript functionalities are internally handled by Dash for dynamic interaction and real-time updates. All processing modules are integrated into an interactive dashboard developed using the Dash framework. HTML components are used for interface structuring, CSS styling is applied for layout and visual presentation, and JavaScript functionalities are internally handled by Dash for dynamic interaction and real-time updates.

The dashboard supports:

- File upload and preview
- Extracted text visualization
- Confidence score display
- Search functionality

- Downloadable text output
- The final extracted output can be downloaded as a text file for future reference, storage, and document management purposes. Overall, the methodology combines OCR technology, image preprocessing, text analysis, and interactive dashboard visualization techniques to provide an efficient, accurate, and user-friendly document extraction and analysis system.

4. Result

The "OCR-Based Intelligent Text Extraction and Analysis System" was successfully tested using different image and PDF documents. The system accurately performed text extraction, document analysis, keyword search, and output generation through an interactive dashboard. The obtained results show that the system reduces manual effort and improves the efficiency of document processing and management.

A. Main Dashboard

The above result shows the main dashboard of the "OptiScan OCR" system developed using the Dash framework. The interface provides a simple and interactive environment where users can upload image or PDF documents for OCR processing. It also includes sections for extracted text display, file preview, search functionality, confidence visualization, and output downloading. The organized layout improves usability and allows users to perform document processing tasks efficiently through a single platform.

Figure 2: Main Dashboard Interface

B. File Upload

The above result demonstrates the file upload functionality of the system. When the user clicks on the upload section, the system opens a file selection window that allows images or PDF documents to be selected from local storage. After selecting the required file, the document is sent to the preprocessing and OCR modules for text extraction and analysis. This feature makes the system user-friendly and supports easy document handling for different file formats.

Figure 3: File Upload Operation for Image and PDF Document Processing

C. Result Screen Display

Figure 4: Extracted Text Display after OCR Processing and Analysis

The above result shows the extracted output after uploading the document into the system. The system processes the uploaded image or PDF file and displays the extracted text in a clear and organized format. It also shows important details such as confidence score, word count, and highlighted search results for better verification and analysis. The interface allows users to easily view the processed content and download the extracted text for future use.

D. Download Output

Figure 5: Downloadable Output File Generated After OCR Text Extraction

The above result shows the downloadable output generated after the document is processed by the system. The extracted text is displayed in a simple text file format, allowing users to save and use the content easily for future reference. The output also includes the OCR confidence score, which helps users understand the accuracy of the extracted text. This feature makes document storage, sharing, and further analysis more convenient and efficient.

5. Conclusion and Future Work

This project offers an effective method for converting scanned and image-based documents into editable digital text using recognition technology. It reduces manual effort while improving speed and accuracy in handling documents. The system also supports features like document sorting, text search, accuracy evaluation, and result export through a simple interface. Overall, it provides a convenient and efficient approach for managing and processing digital documents. As a future enhancement, the system can be extended to support multilingual OCR for extracting text from both regional and international languages, along with improved recognition of handwritten as well as printed content. It can also be upgraded to enable batch processing, allowing multiple documents to be uploaded and processed simultaneously for better efficiency. Additionally, mobile application support can be introduced so that users can capture and

process documents directly from smartphones. Stronger security features such as user login and access control can also be implemented to ensure better protection and controlled access to documents.

References

1. Mayank Deshmukh, Saloni Rabde, Priyanka Makode, Sourabh Jasuja, Prof. Bhavesh Khasdev, "A Comprehensive Study on Text Detection and Extraction from Images and PDF Documents", International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, vol. 13, no. 6, pp. 1-4, Nov-Dec 2025.
2. Prof. Anuradha Thorat, Mayur Zagade, Shivani More, Manish Pasalkar, Anand Narute, "Research Paper on Text Extraction using OCR", International Journal of Advanced Research in Science, Communication and Technology, vol. 3, no. 14, pp. 10-14, May 2023.
3. Jyothi E, K Tejaswini, Lakshmi Chintalapati, Mr. MD. Shafiulla, "Text Extraction from Images using OCR", International Journal for Research in Applied Science & Engineering Technology, vol. 8, pp. 1805-1810, May 2020.
4. Shrinath Janvalkar, Paresh Manjrekar, Sarvesh Pawar, Prof. Laxman Naik, "Text Recognition from an Image", International Journal of Engineering Research and Applications, vol. 4, no. 4, pp. 149-151, April 2014.
5. Jyoti Wadmare, Sunita Ravindra Patil, Dakshita Kolte, Kapil Bhatia, Palak Desai, Ganesh Wadmare, "Transforming images into words: optical character recognition solutions for image text Extraction", IAES International Journal of Artificial Intelligence, vol. 14, no. 4, pp. 3412-3420, August 2025.
6. Deepak Singh, "OCR-Driven Automation: A Case Study on Document Processing Using Tesseract and OpenCV", Power System Technology Journal, vol. 49, no. 1, pp. 1661-1671, March 2025.
7. Sayan Kumar Garai, Ojaswita Paul, Upayan Dey, Sayan Ghoshal, Neepa Biswas, Dr. Sandip Mondal, "A Novel Method for Image to Text Extraction Using Tesseract-OCR", American Journal of Electronics and Communication, vol. 3, no. 2, pp. 8-11, May 2023.