Intelligent OCR System for Handwritten Government Document Digitization

[1]Dr. Ratna Raju Mukiri,[2]Gattupalli D. S. V Naga Sai Mani Deepika,[3]Boddu Sravani,[4]Dasari Manoj Kumar

[1]Professor, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

[2,3,4]U. G Student, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

ABSTRACT- ***The project Intelligent OCR System for Handwritten Government Document Digitization proposes a digital approach for transforming handwritten government records into structured and editable text. This traditional practice increases processing time, creates dependency on human effort, and makes searching, storing, and reusing document content difficult. To address these limitations, the proposed system introduces an OCR-driven document digitization framework that automates handwritten text extraction and organizes the processed output in digital form.The system is implemented as a web application using HTML, CSS, and JavaScript for the user interface, FastAPI for backend services, and MongoDB Atlas for persistent data storage. For handwritten text recognition, the system integrates Gemini Vision OCR, and a translation component is included to convert extracted text into English when needed. The core methods used in the project include OCR-based text recognition, image preprocessing for improving input quality, translation of extracted text, database-based record management, and document export generation. The proposed solution minimizes manual work, improves accessibility of handwritten records, and supports faster retrieval and reuse of document content.***

KEYWORDS- ***Optical Character Recognition (OCR), Handwritten Document Digitization, Gemini Vision OCR, FastAPI, MongoDB Atlas, Text Extraction, Translation, Document Management, OCR History.***

INTRODUCTION

Handwritten documents continue to play an important role in many government offices,

educational institutions, and administrative departments. Records such as applications, certificates, handwritten forms, and official notes are often preserved in paper format, which makes their management difficult and time-consuming. In the existing process, staff members must read the handwritten content, manually rewrite it into digital records, and store it for future use. This traditional approach requires significant human effort, increases the chances of transcription errors, and creates difficulties in searching, retrieving, and reusing document information efficiently.To address these challenges, the Intelligent OCR System for Handwritten Government Document Digitization is proposed as a digital solution for converting handwritten records into editable text. The system uses Optical Character Recognition (OCR) to identify and extract handwritten text from uploaded document images. It also supports translation of the extracted text into English, making the content more accessible and useful for official processing. The application is developed using HTML, CSS, and JavaScript for the frontend, FastAPI for backend services, MongoDB Atlas for storing user and OCR data, and Gemini Vision OCR for handwritten text recognition.By combining OCR, translation, history management, and file export in one platform, the proposed system reduces manual effort and improves the efficiency of document digitization and management.

RELATED WORK

Handwritten document digitization has been widely studied through the use of Optical Character Recognition (OCR) and machine learning techniques. Earlier OCR systems were mainly designed for printed text and showed limited performance on handwritten documents because of variations in writing style, document quality, and script complexity. Recent research has improved this area by introducing deep learning and transformer-based models for handwritten text recognition.Several studies reviewed handwritten OCR methods, datasets, and recognition challenges, showing how OCR has evolved from traditional image-processing approaches to AI-based systems. Models such as TrOCR demonstrated that pre-trained transformer architectures can improve text extraction accuracy for both printed and handwritten content. Other works, including archival digitization platforms like Transkribus, highlighted the practical use of OCR for large collections of handwritten historical records.These developments influenced the proposed Intelligent OCR System for Handwritten

Government Document Digitization. By combining handwritten OCR, translation, OCR history, and digital storage in a single platform, the proposed system aims to provide a practical solution for converting handwritten government records into editable and manageable digital text.

LITERATURE SURVEY

Recent studies on handwritten document digitization have focused on improving Optical Character Recognition (OCR) for documents with uneven writing styles, noisy backgrounds, and multilingual content. Memon et al. (2020) reviewed handwritten OCR research and highlighted key issues such as dataset limitations, segmentation complexity, and variation in character shapes. Li et al. (2021) proposed TrOCR, a transformer-based OCR framework that showed better performance on handwritten and printed text by using pre-trained vision and language representations. Muehlberger et al. (2019) discussed the Transkribus platform, which demonstrated the large-scale application of handwritten text recognition for archival and historical document preservation.These works show that modern OCR is moving beyond simple character matching toward context-aware recognition and practical document digitization systems. The proposed Intelligent OCR System for Handwritten Government Document Digitization builds on these ideas by combining handwritten text extraction, translation, result storage, OCR history management, and downloadable output generation in a single web-based platform.

EXISTING SYSTEM

In the existing approach, handwritten government documents are usually processed either manually or with basic OCR-based text extraction systems. Manual processing requires staff to read handwritten records and rewrite the information into digital form, which is slow, repetitive, and highly dependent on human effort. Even when OCR tools are used, many traditional systems mainly focus on extracting text and are not fully effective for handwritten documents because of variations in handwriting style, document quality, and language complexity.Most existing OCR solutions do not provide a complete workflow for handwritten government document management. They often lack features such as translation of extracted text, OCR history maintenance, structured storage of processed documents, and file export in multiple formats. As a result, users still need to perform additional manual work after OCR

processing, such as correcting text, organizing extracted content, and saving records separately.These limitations reduce the practical usefulness of traditional OCR systems for government offices and similar organizations. Therefore, there is a need for an integrated handwritten document digitization system that combines OCR, translation, document storage, history management, and downloadable output in a single platform to improve efficiency and accessibility.

PROPOSED SYSTEM

The proposed system, Intelligent OCR System for Handwritten Government Document Digitization, is designed to provide a complete digital solution for processing handwritten government records. Unlike traditional OCR systems that focus only on text extraction, the proposed method integrates handwritten text recognition, translation, storage, OCR history management, and downloadable output generation into a single web-based platform.In this method, the user first registers and logs into the application, then uploads a handwritten document image for processing. The uploaded document is passed to Gemini Vision OCR, which extracts the handwritten content and converts it into editable digital text. To improve accessibility, the system also provides a translation module that converts the extracted text into English whenever required. The processed OCR result, translated content, and document details are then stored in MongoDB Atlas, allowing users to access their OCR history at any time.The proposed system also includes an export module through which users can download the extracted content in TXT, DOCX, and PDF formats. By combining OCR, translation, digital storage, and document management in one platform, the proposed method reduces manual effort, improves accessibility of handwritten records, and provides a practical solution for efficient government document digitization.

SYSTEM ARCHITECTURE

The Intelligent OCR System for Handwritten Government Document Digitization follows a web-based client–server architecture. The user accesses the system through a web interface to register, log in, upload handwritten documents, view OCR results, and download output files. The uploaded document is sent to the FastAPI backend, which handles authentication, file processing, and communication with other modules. The document is then processed using Gemini

Vision OCR to extract handwritten text. If required, the extracted text is translated into English through the translation module. Finally, the OCR result, translated text, and document details are stored in MongoDB Atlas, enabling OCR history management and future retrieval of processed records.

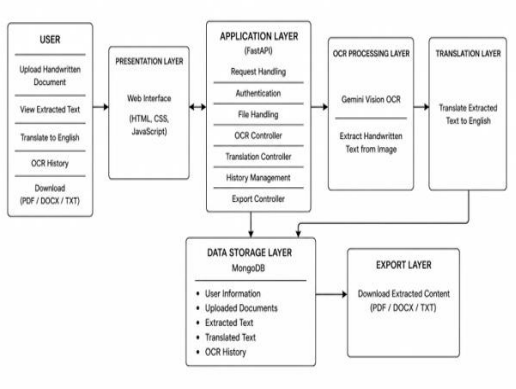

Fig 1: System Architecture

METHODLOGY DESCRIPTION

User Authentication Module

This module manages user registration, login, and logout. It ensures that only authorized users can access the OCR system and use its features securely.

Document Upload Module

This module allows users to **upload handwritten document images** into the system. The uploaded files act as the input for OCR processing.

OCR Processing Module

This is the core module of the system. It uses **Gemini Vision OCR** to analyze the uploaded handwritten document and extract text from it in digital form.

Translation Module

This module converts the extracted handwritten text into **English** when translation is required, improving readability and accessibility.

OCR Result Management Module

This module displays the extracted OCR text to the user and manages the final processed output before saving or downloading.

OCR History Module

This module stores and retrieves previously processed OCR documents, allowing users to view their earlier digitized records.

Export / Download Module

This module generates output files in TXT, DOCX, and PDF formats so that users can download and use the extracted content.

Database Management Module

This module handles storage of user details, OCR results, translated text, and OCR history in MongoDB Atlas.

RESULTS AND DISCUSSION

Home Page

This figure shows the home page of the IntelliOCR system along with its features section. It introduces the platform's main objective of extracting and translating document text using AI, and highlights key features such as AI grammar correction and export of extracted text into Word format.

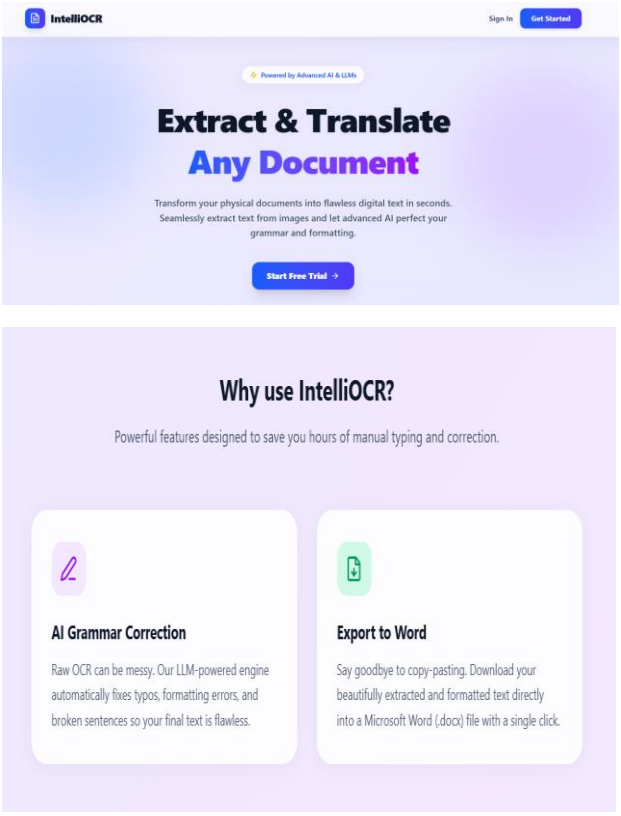

Fig: Home Page

Register Page

This figure shows the account creation page where new users can register by entering their username, email, and password.

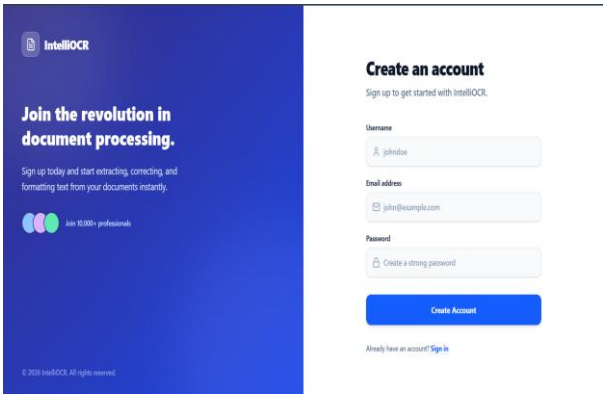

Fig: Register Page

Login Page

This figure shows the login interface where existing users can enter their username and password to access the system.

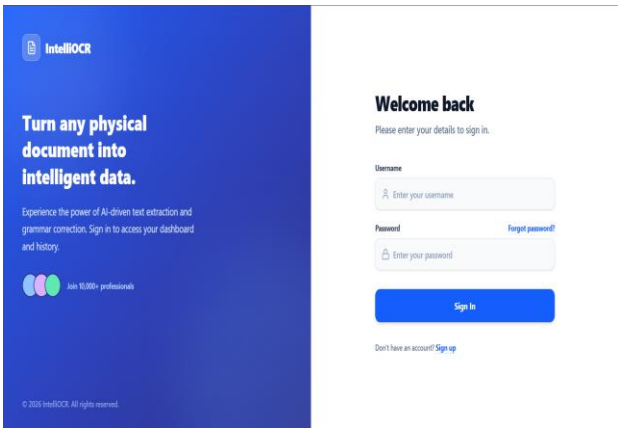

Fig: Login Page

Dashboard Page

This figure shows the user dashboard displaying overview details such as total processed documents, average accuracy, and saved documents.

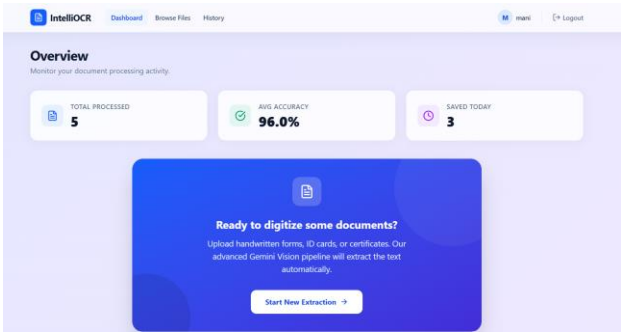

Fig: Dashboard Page

Browse Files / Upload Page

This figure shows the document upload interface where users can browse and upload handwritten files for OCR extraction.

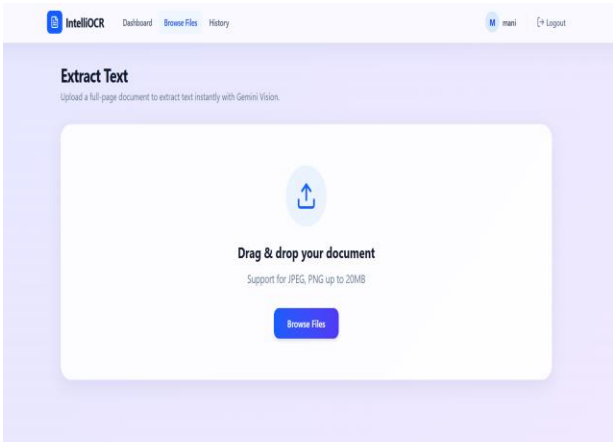

Fig: Browse Files / Upload Page

Document Upload and Language Selection

This figure shows the uploaded handwritten document preview in the IntelliOCR system. It also provides the option to select the document language before starting OCR text extraction.

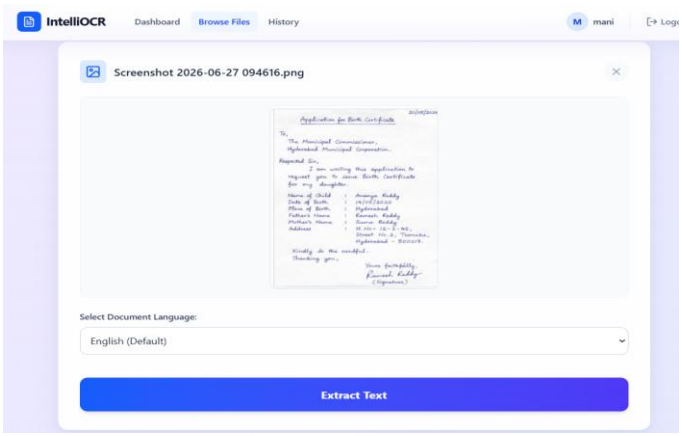

Fig: Document Upload and Language Selection

OCR Processing Stage

This figure shows the OCR processing stage where the uploaded handwritten document is being analyzed by the Gemini Vision OCR engine. The system extracts the handwritten text from the document during this step.

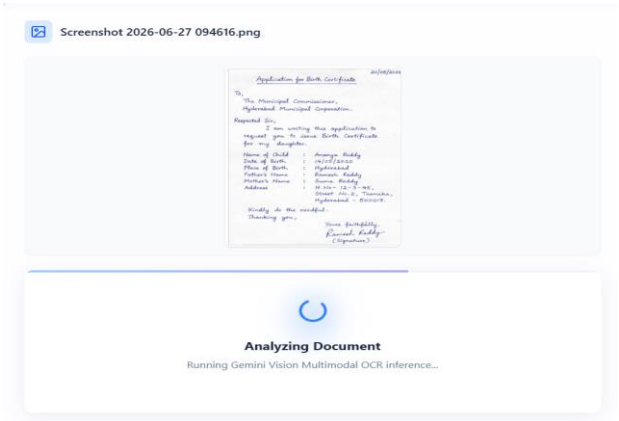

Fig: OCR Processing Stage

Extracted OCR Output

This figure shows the extracted text output displayed alongside the original uploaded document. It also provides options to copy

the text, export it, and save the extracted result to the database.

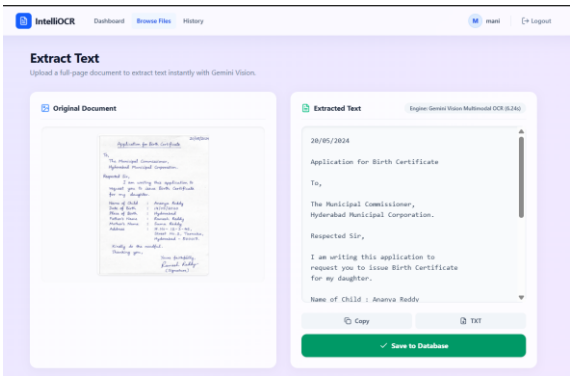

Fig: Extracted OCR Output

Save to Database Confirmation

This figure shows the confirmation message displayed after the extracted OCR result is successfully saved into the database. It confirms that the processed document has been stored for future access in the history section.

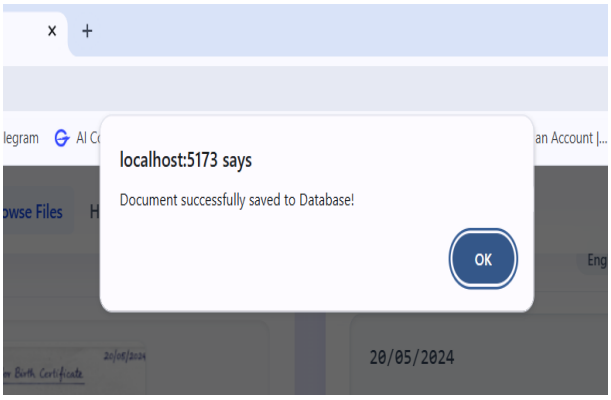

Fig: Save to Database Confirmation

Document History

This figure shows the OCR history page where users can view previously processed documents and their saved records.

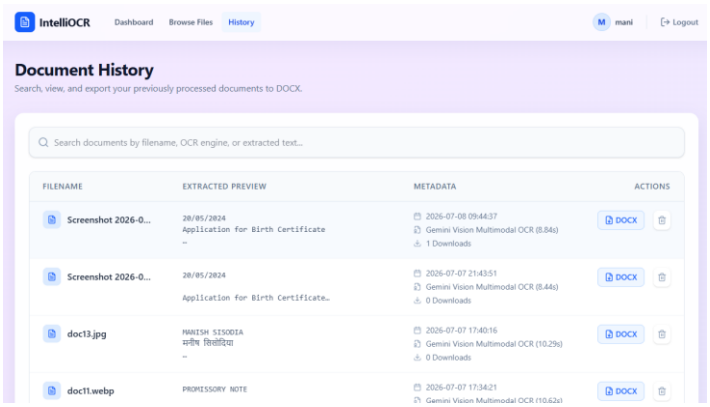

Fig: Document History

Delete Confirmation Dialog

This figure shows the delete confirmation popup displayed before removing a document from OCR history.

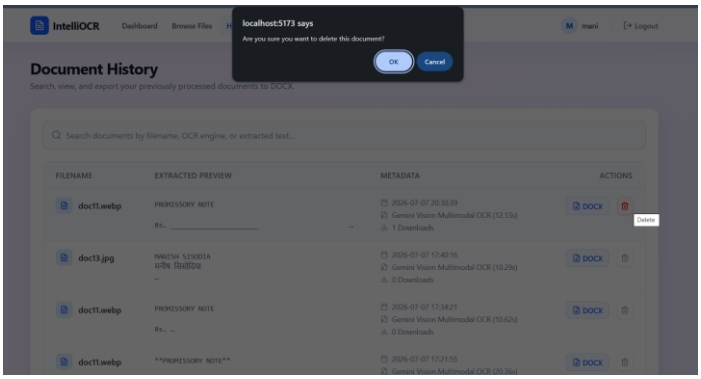

Fig: Document History

CONCLUSION

The Intelligent OCR System for Handwritten Government Document Digitization provides an effective solution for converting handwritten documents into editable digital text. By integrating OCR, translation, database storage, OCR history management, and file export, the system reduces manual effort and improves the accessibility of handwritten records. It helps users process, store, retrieve, and download document

content more efficiently. The project demonstrates how modern OCR-based technology can simplify document digitization and support better management of handwritten government records in a structured digital environment.

FUTURE SCOPE

The proposed system can be enhanced further by improving OCR accuracy for unclear and low-quality handwritten documents. Support for more regional and multilingual handwritten scripts can be added to make the system more useful in diverse government environments. Future versions may include a mobile application for easier document upload and access, as well as AI-based grammar correction and formatting of extracted text. Additional features such as keyword-based search, cloud backup, role-based access, and document classification can also be integrated to make the system more scalable, secure, and suitable for real-world document management applications.

REFERENCES

[1] D. P. Harini, "Electronic Health Record Stage Identification using Principal Component Analysis," *Machine Intelligence Research,* vol. 18, no. 01, pp. 864- 873, 17 8 2024.

[2] D. P. Harini, "Two Level Intrusion Detection For Detecting Intruders in Multitier Web Applications," *International Journal of Engineering & Science Research,* vol. 3, no. 9, pp. 472-478, 2013.

[3] D. P. Harini, "Analyze and Predict of Human Cyber Attackers using Artificial Neural Network," *Journal of Basic Science and Engineering,* vol. 21, no. 1, pp. 1529 - 1536, 8 7 2024.

[4] W. AlKendi, J. S. Tejedor, and others, "Advancements and Challenges in Handwritten Text Recognition," Sensors / Survey Article, 2024.

[5] R. Sharma and M. Blumenstein, "Offline recognition of handwritten Indic scripts: A survey," Computer Science Review, vol. 38, 2020.

[6] R. Sinha and others, "Digitization of Document and Information Extraction using OCR," arXiv preprint arXiv:2506.11156, 2025.

[7] D. Chang and others, "DLoRA-TrOCR: Mixed Text Mode Optical Character Recognition," arXiv preprint arXiv:2404.12734, 2024.

[8] F. Lauar and others, "Leveraging Transfer Learning for Language Adaptation in TrOCR," arXiv preprint arXiv:2407.06950, 2024.

[9] C. Garrido-Muñoz and others, "Handwritten Text Recognition: A Survey," arXiv preprint arXiv:2502.08417, 2025.

[10] Y. Sun, D. Zhou, C. Lin, C. He, W. Ouyang, and H.-S. Zhong, "LOCR: Location-Guided Transformer for Optical Character Recognition," arXiv preprint arXiv:2403.02127, 2024.

[11] L. Beerens and D. J. Higham, "Vulnerability Analysis of Transformer-based Optical Character Recognition to Adversarial Attacks," arXiv preprint arXiv:2311.17128, 2023.

[12] C. A. Romein and others, "Assessing advanced handwritten text recognition engines for historical documents," Journal/Conference Article, 2025.

[13] S. K. Nigam and others, "Handwritten Legal Document Translation with OCR and Vision-Language Models," arXiv preprint arXiv:2512.18004, 2025.

[14] "Transkribus: Handwritten Text Recognition technology for historical documents," READ-COOP / Transkribus Platform Documentation, 2019.

[15] "Google Cloud Vision API: Handwriting Detection with OCR," Google Cloud Documentation, 2025.