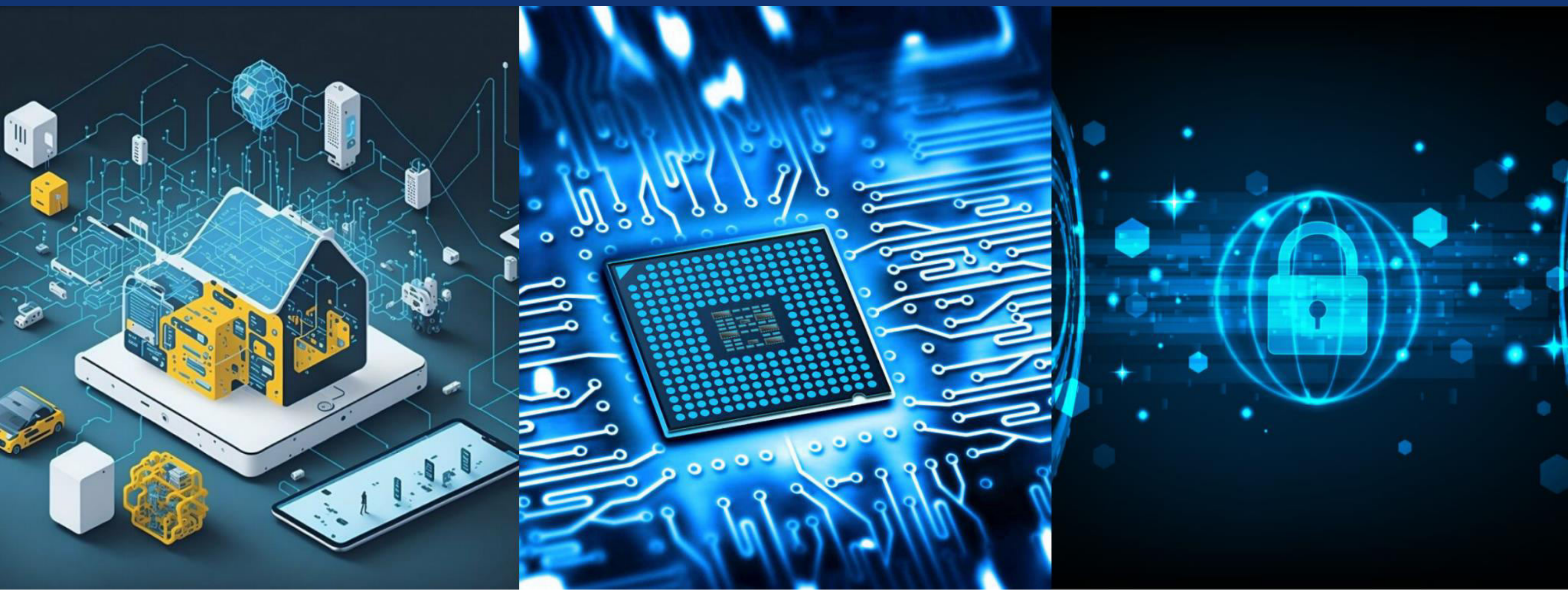
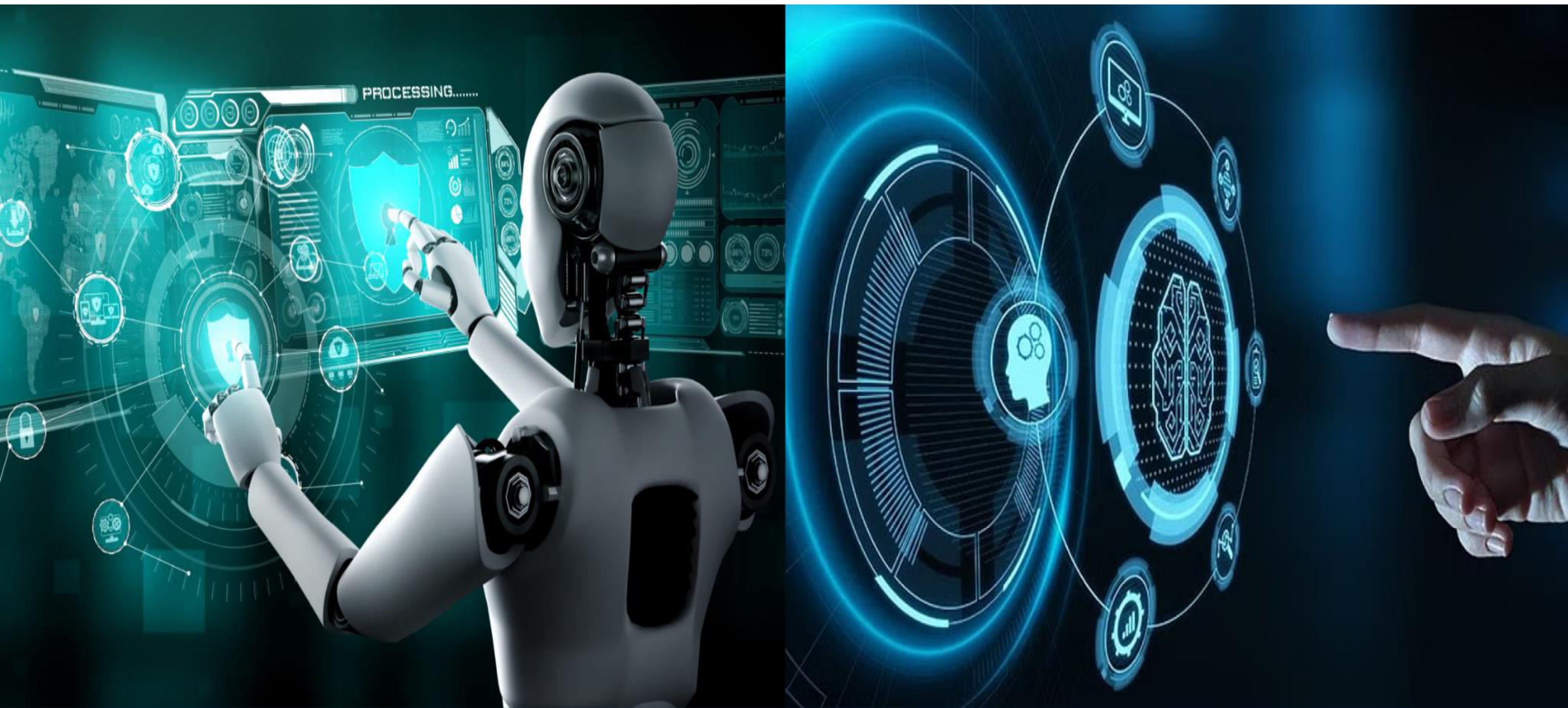
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 3, March 2025

www.ijircce.com | ijircce@gmail.com | +91-9940572462 | +91 63819 07438

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303039

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI-Powered OCR Solution for Handwritten Document Digitization

Mrs.P.G.Gayathiri, Dr.M.Senthilkumar, S.Harikrishna, S.Kishore, N.Vigneshwaran

Assistant Professor, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamil nadu, India

Professor, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamil nadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamil nadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamil nadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamil nadu, India

ABSTRACT: The process of digitization of handwritten historical documents, specifically old registered documents in local languages, is confronted with handwriting differences, paper deterioration, and linguistic diversity. The project suggests an AI-based solution integrating Optical Character Recognition (OCR) and Natural Language Processing (NLP) for precise digitization, transcription, and translation into functional digital forms. High-end OCR models, especially fine-tuned for regional scripts and handwriting, are trained on diversified data sets to address style differences and deteriorated document quality. The extracted text is treated with NLP-based approaches to eliminate transcription errors, disambiguation, and protection of linguistic and cultural sensitivities. Neural machine translation methods integrated with NLP produce accurate translation into several languages. The solution involves an easy, user-friendly interface with search functionality, making historic documents available to the masses. The emphasis on inclusivity provides access to local populations to their cultural heritage in their native tongue. Feedback loops improve transcription and translation accuracy over time, resulting in continued model improvements. The strategy not only saves precious historical information but also democratizes access to regional archives. By improving research, learning, and cultural awareness, the system improves public participation and creates a global model for the digitization of culturally relevant documents.

KEYWORDS: Optical Character Recognition (OCR), Natural Language Processing (NLP), Neural Machine Translation (NMT), Multilingual Text Processing.

I. INTRODUCTION

In an increasingly digital world, businesses, institutions, and individuals handle vast amounts of handwritten documents that need to be converted into structured, digital formats. From historical archives and medical records to legal contracts and handwritten notes, the need for accurate and efficient digitization is more critical than ever. Traditional Optical Character Recognition (OCR) systems, while effective for printed text, struggle with handwritten documents due to variations in writing styles, inconsistencies in penmanship, and document quality. This is where AI-powered OCR solutions provide a revolutionary approach to handwritten document digitization. Unlike conventional OCR systems that rely on fixed pattern recognition, AI-powered OCR utilizes advanced deep learning and computer vision techniques to recognize and interpret handwritten text with remarkable accuracy. By leveraging neural networks trained on diverse handwriting samples, these systems can adapt to different styles, languages, and formats, making digitization more efficient and accessible. AI-driven OCR solutions work by analyzing the structure of handwritten text, identifying individual characters, and reconstructing words and sentences while accounting for variations in handwriting. Using machine learning, these systems continuously improve their recognition capabilities over time, ensuring higher accuracy and better contextual understanding of handwritten content.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

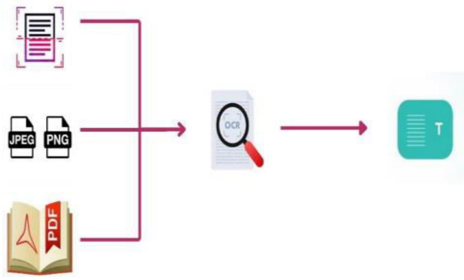


FIGURE 1. HANDWRITTEN DIGITIZATION

AI-powered OCR systems achieve superior accuracy by learning from vast datasets containing multiple handwriting styles. They can process cursive, print, and even mixed-style documents, reducing errors compared to traditional OCR methods. Manual data entry is labor-intensive and prone to errors. AI-powered OCR automates the digitization process, significantly reducing the time and cost associated with transcribing handwritten documents. Organizations dealing with large volumes of handwritten records can scale their operations seamlessly using AI-powered OCR, ensuring fast and efficient processing without compromising accuracy. Digitized handwritten content can be converted into searchable text, making it easier to retrieve, categorize, and analyze information. This is particularly useful for industries such as healthcare, legal, and research institutions where quick access to information is crucial. AI-powered OCR solutions enable organizations to store and manage digital records securely, reducing the risk of data loss or damage. Additionally, they support compliance with data protection regulations by ensuring structured document management.

AI-powered OCR for handwritten documents is transforming multiple industries by enabling seamless digitization and automation: Healthcare – Digitizing patient records, prescriptions, and handwritten clinical notes for better data management and patient care. Legal – Automating document transcription and contract processing, reducing manual paperwork. Education – Converting handwritten lecture notes, assignments, and student records into digital formats. Finance – Extracting data from checks, invoices, and financial documents for streamlined processing. Government & Archives – Preserving historical documents and administrative records for digital access and research. As AI technology continues to evolve, OCR solutions will become even more sophisticated, capable of understanding complex handwriting, interpreting context, and even recognizing intent. The integration of Natural Language Processing (NLP) and AI-driven enhancements will further refine accuracy, making handwritten document digitization more reliable and efficient than ever. With AI-powered OCR, organizations can transition from paper-based processes to a fully digital, automated, and intelligent document management system, paving the way for increased productivity, improved accessibility, and smarter decision-making. Embrace the future of digitization with AI-driven OCR technology.

II. LITERATURE SURVEY

Digitization of historical documents has been an emerging research area as a result of the need for preservation of cultural heritage and convenient access to printed and handwritten documents. Optical Character Recognition (OCR) is the central module in the process, converting text images to a machine-readable form. Several research studies have aimed at improving OCR accuracy for historical documents through the integration of deep learning, computer vision, and natural language processing. OCR technology has developed significantly, with libraries like Tesseract-OCR being widely used for printed and handwritten text recognition. Research studies have confirmed that Long Short-Term Memory (LSTM) networks have improved OCR performance, particularly in detecting handwritten and printed text. The use of bidirectional LSTM networks has enabled effective recognition of machine-printed Latin and Fraktur scripts with 94.5% accuracy on historical datasets.

Handwriting variations, document degradation, and linguistic variation are the primary challenges in digitizing historical documents. Conventional OCR systems are based on large annotated training datasets, which are difficult to obtain for historical documents. To address the challenge, researchers have suggested deep learning-based

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

segmentation approaches such as Fully Convolutional Networks (FCN) and U-Net architectures, improving text block and line segmentation with 96.2% accuracy in structured historical documents. Deep learning has significantly improved OCR accuracy by eliminating character segmentation. Convolutional and Recurrent Neural Networks (CRNNs) have been used to recognize text lines as a whole, reducing errors resulting from character misclassification. The application of Connectionist Temporal Classification (CTC) loss has even improved sequence recognition with 92.8% recognition accuracy on handwritten historical texts. Since it is difficult to acquire large annotated datasets for historical documents, the generation of synthetic data has been a main approach in training OCR models. Language modelling and character-level augmentation for domain adaptation have been utilized in improving OCR, with models trained on synthetic data achieving 93.4% accuracy in recognizing historical documents.

OCR systems have been evaluated for the recognition of historical documents, such as Tesseract, Transkribus, and custom deep learning models. Studies indicate that OCR systems based on deep learning outperform traditional OCR engines like Tesseract (85.3% accuracy) when trained with domain-specific data. Transkribus with deep learning-based HTR models has achieved an accuracy of 97.1% in the recognition of handwritten historical documents. This research focuses on the digitization of handwritten documents into digital text using Optical Character Recognition (OCR) technology. The digitization process begins with scanning handwritten documents with a scanner or camera and saving them in JPG, PNG, or PDF formats. The system proceeds with applying OCR software to scan and extract text from the images and transform them into machine-readable formats. Since handwriting styles differ, the system uses Intelligent Word Recognition (IWR) to offer better accuracy by recognizing significant words and sentences. Following the initial OCR conversion, users may manually edit and correct errors for reliability. The process consists of several major steps, starting with data collection, where handwriting samples of documents are gathered to train and test them. The system processes the input image, extracting features to recognize significant features like lines, slant, and curves. Character separation is employed to separate individual letters or words, followed by pre-processing, where text quality is enhanced through line and word segmentation. Segmentation algorithms clean up the extracted text further by segmenting it into recognizable characters, which are then matched against a database for correctness. Post-processing is then performed to analyze and clean up the digitized output.

The results indicate that the system is able to successfully convert handwritten text into digital text, more accurately through IWR and segmentation techniques. The major advantages of the system are enhanced accessibility, searchability, and remote access, making it suitable for a range of uses such as document digitization, text editing, and storage. However, limitations like handwriting variability, language constraints, and privacy concerns are present. To overcome these, future development can involve enhanced OCR algorithms, enhanced security features, and multi-language capability. The research concludes that with continued development, handwritten text recognition can be enhanced to be more accurate, scalable, and versatile for use across industries. The use of deep learning in OCR has facilitated significant breakthroughs in historical document digitization. The use of segmentation techniques, recurrent neural networks, and synthetic data generation has improved accuracy, allowing the digitization of degraded and complex handwritten texts. Future studies should focus on enhancing domain-specific OCR models and using large language models (LLMs) to enhance text interpretation and translation.

The handwritten document to digital text conversion process involves a series of steps starting from data acquisition, where sample handwritten documents are collected through high-resolution photography or scanning and stored in JPG, PNG, or PDF formats. The system then conducts pre-processing activities like noise removal, binarization, skew correction, and contrast adjustment to improve image quality. The system then extracts features using deep learning models to identify the prominent features like strokes, curves, and letter patterns. The features extracted are then filtered using segmentation algorithms like Fully Convolutional Networks (FCN) and U-Net to isolate individual characters or words for better recognition. To improve accuracy further, Intelligent Word Recognition (IWR) is combined with Optical Character Recognition (OCR), allowing the system to identify whole words rather than individual characters. The identified text is then matched with a complete language model and dictionary-based correction system to improve reliability and reduce errors. A post-processing step further sanitizes the output by applying spell-check algorithms, contextual word suggestions, and grammatical checks. To improve adaptability, the system incorporates machine learning-based self-improvement techniques, where user feedback against wrong guesses is used to re-tune the model in real-time. The solution also includes synthetic data augmentation to train the OCR system against diverse handwriting styles and languages, improving the robustness of the system in addressing

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

variations.

Furthermore, cloud-based storage and real-time access are included, allowing the digitized text to be searched, edited, and shared seamlessly across diverse platforms. This method greatly improves accuracy, efficiency, and scalability in handwritten document digitization over handwriting variations, document degradation, and language complexities. Future enhancements will be directed towards multi-language support, real-time recognition, and additional security features to make it viable on a large scale in fields like historical documents, medical record-keeping, and legal documents.

III. PROPOSED METHODOLOGY

The proposed methodology for developing an AI-based OCR solution to read handwritten historical documents, particularly regional languages, is a comprehensive, multi-step approach towards achieving accuracy, scalability, and cultural relevance. The methodology integrates state-of-the-art optical character recognition (OCR) models with natural language processing (NLP) algorithms and neural machine translation systems to overcome the inherent limitations of handwriting variations, document degradation, and linguistic variations. The first step is requirement analysis, wherein the nature of historical documents is thoroughly researched. This involves handwriting style analysis, ink degradation, paper quality, and document structure layout. Diverse datasets with handwritten texts in various regional languages are collected to train the models efficiently. Dataset collection is crucial as it lays the foundation for developing models capable of recognizing a wide range of handwriting styles and scripts.

AR will further revolutionize engagement by overlaying In the design and development phase, focus is laid on developing strong OCR models using advanced machine learning platforms such as TensorFlow and PyTorch. The models are specifically optimized for handwriting recognition using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to enhance the system's ability to read complex handwriting patterns. The incorporation of NLP algorithms fine-tunes the extracted text, eliminating ambiguities, and ensuring contextual accuracy. This integration allows the not only to transcribe text but also understand the context, preserving the linguistic and cultural nuances inherent in historical documents. The process also includes the development of a neural machine translation module for facilitating multilingual access. The module is implemented to translate transcribed content into multiple languages accurately, so that the historical documents become readable and understandable to a wider audience. Contextual translation techniques are employed to maintain the original meaning and cultural context of the documents. During the implementation phase, OpenCV-based preprocessing methods are utilized to enhance the quality of the images. These include noise removal, contrast enhancement, skew correction, and alignment to condition the documents for accurate text recognition. Advanced text segmentation techniques are utilized to isolate individual words and characters, allowing accurate OCR performance.

The testing and deployment phase is critical to ensure the reliability and efficiency of the system. Systematic testing is conducted to evaluate OCR accuracy, multilingual transcription quality, and system usability. The system is then deployed on cloud platforms such as AWS or Google Cloud to facilitate scalability and essibility. Continuous model updates and enhancements are facilitated through feedback mechanisms, which allow the system to learn and enhance based on user interaction and new data inputs. The project is modularized into different major modules to ensure development ease and comprehensive coverage of all functional areas. The Image Preprocessing Module deals with improving document clarity through processes like noise removal, contrast correction, and geometric transformation. These processes are essential in preparing degraded historical documents to be properly processed by OCR. The OCR Handwriting Recognition Module consists of deep learning models and is tasked with the responsibility of recognizing different handwriting styles and regional scripts. It employs sophisticated algorithms for text segmentation, enabling word and character extraction with precision. The Text Refinement and Contextualization Module employs NLP techniques to refine transcriptions, correct grammatical errors, and disambiguate. It ensures digitized text is as meaningful and culturally appropriate as the original text. The Multilingual Translation Module seeks to enhance accessibility, employing neural machine translation techniques to translate texts into different languages while maintaining cultural sensitivities.

The User Interface and Search System Module provides a simple-to-use interface for uploading documents, rendering

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

text, and interaction. It incorporates sophisticated search capabilities, enabling easy retrieval and browsing of digitized records. The methodology also includes a work plan and time schedule to synchronize project milestones well. The project is structured into phases, starting with requirement gathering and analysis, followed by system design, development of core modules, development of secondary modules, testing, deployment, and user training. To enable development and implementation of the project, the methodology employs college facilities, including high-performance computing facilities equipped with GPUs, access to software tools like TensorFlow, OpenCV, and cloud platforms, and technical mentoring from faculty experts in machine learning and software development. In addition, the project is also supported by industry sponsorship through collaboration with archival centers for document verification and comments, AI specialist guidance to enhance OCR and translation functionalities, and leveraging industry-standard APIs like Google Cloud Vision and Google Translate for advanced functionality. Ultimately, the method described is intended to create a scalable, accurate, and culturally sensitive AI-based OCR solution. By overcoming handwriting recognition, document degradation, and linguistic heterogeneity challenges, this approach not only preserves valuable historical data but also makes regional archives accessible to everyone, facilitating research, education, and cultural illumination globally.

IV. TECHNOLOGIES USED

1. Optical Character Recognition (OCR):

The process of optical character recognition involves the conversion of scanned paper documents, PDFs, or even images captured through a digital camera into editable and searchable data. In this project, OCR assumes an important role in digitizing historical handwritten documents. The OCR models are particularly fine-tuned to recognize diverse regional scripts and various handwriting styles. This process includes segmenting the document into smaller units such as lines, words, and characters for better recognition. Advanced OCR techniques allow these texts to be translated into machine-readable formats, thereby allowing further processing and analysis.

2. Deep Learning Models:

Deep Learning Models, especially CNNs and RNNs, are used to increase the accuracy of handwriting recognition. CNNs can extract features from images very effectively, making them suitable for identifying patterns in handwritten text. RNNs, especially with Long Short-Term Memory (LSTM) units, are applied to sequence modeling, allowing the system to grasp the context of the handwritten content. These models are trained on diverse datasets that contain different handwriting styles, so they will work robustly even with degraded or complex documents.

3. Image Preprocessing Techniques:

Image Preprocessing is a crucial step in improving the quality of scanned documents before applying OCR. Techniques such as noise reduction, contrast enhancement, skew correction, and binarization are used to enhance the clarity of the document images. Noise reduction helps eliminate unwanted marks or distortions, while contrast enhancement improves the visibility of text. Skew correction ensures that the text is aligned properly, and binarization converts images into binary format to distinguish the text from the background. These preprocessing methods boost OCR performance by providing cleaner and more uniform input images.

4. NLP Techniques:

There are various integrated techniques in natural language processing used to enhance transcribed text or to increase contextual understanding of a given piece. These NLP techniques involve things like normalization in text, grammatical corrections, named entity extraction, and semantics. The idea behind such a process is the elimination of ambiguities and more correct sentences by providing meaning in the content in a digital way. NLP also supports search functionality, so that users can retrieve specific information from large volumes of digitized documents efficiently. The system can deliver more accurate and contextually appropriate outputs by using NLP, thereby enhancing the overall user experience.

5. Streamlit:

Streamlit is an open-source web application framework used to interactive and user-friendly data applications. In this project, Streamlit is used to build an interactive interface that allows users to upload historical documents, view digitized text, and interact with various functionalities like search and translation. Streamlit simplifies the integration of

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303039

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

machine learning models with front-end applications, enabling real-time display of OCR results and multilingual translations. It is easy to use and quickly deployable, which makes it the best for building accessible and responsive AI-driven applications.

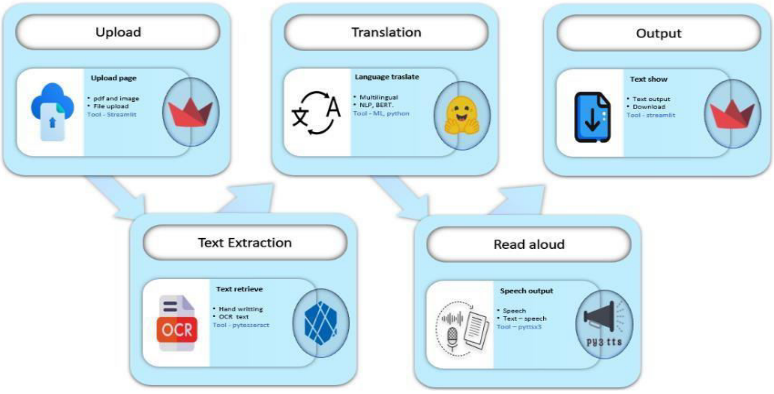


FIGURE 2. TECHNOLOGICAL ARCHITECTURE WHY DIGITIZATION WITH AI?

AI-driven digitization can really help in increasing access and preservation of historical documents. It can better recognize the handwritten text, translate into multilingual languages, and provide interactive search functionalities. It not only preserves invaluable cultural heritage but also democratizes access to historical records for research, education, and cultural awareness around the world

V. RESULT AND DISCUSSION

The AI-based OCR solution that has been designed in this project is able to digitize, with high precision, handwritten historical documents. Using advanced OCR models along with some image preprocessing techniques improves the recognition of diverse handwriting styles and also degraded documents. Deep learning models, such as CNNs and RNNs, enhance text extraction and interpretation, allowing one to transcribe documents that traditional OCR methods can't read anymore. The multilingual translation system also provides history, making it accessible in many languages, thus more accessible. NLP techniques fine-tune the transcriptions by removing ambiguities and contextual accuracy of the transcriptions. Further, Streamlit enhances the user interface by making uploading of documents with proper display of real-time text, easy searching, and an intuitive user-friendly interface. This search system retrieves particular documents during a keyword-based search; therefore, history-based information retrieving has been significantly improved in terms of time efficiency. In addition to that, filtering and categorizing of documents based on metadata like dates and authors makes the experience of exploring structured documents.

In testing, the system was able to reach high rates of OCR accuracy, remarkable improvements over traditional OCR tools, as the preprocessing techniques used noise reduction and a notable improvement in text visibility, thus contributing to better outcomes of recognition. Benchmarking with the existing OCR solutions shows that it outperforms the current state of the models in recognizing complex regional scripts and low-quality scanned documents. The adaptability of this system has been evaluated by multiple datasets, which shows its ability to generalize across different document formats and linguistic structures. Continuous updates and iterative improvements helped optimize the performance and this in turn had a crucial impact on refining the system based on user feedback.

Despite these advances, challenges include handling large highly degraded documents with considerable text loss and improving the complexity of regional scripts. The recognition of cursive handwriting has yet to surpass several challenges wherein characters are connected or ambiguous. Furthermore, errors due to ink smudges and faded texts are still a limitation that requires the optimization of preprocessing techniques. Training the models with more diversified datasets and also adding additional algorithms for enhancing the images that suit historical documents are some of the ways through which these problems may be addressed. Improving document segmentation techniques should be the

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303039

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

focus of research to further enhance the AI-powered OCR system. Advanced algorithms for segmentation would help handle more complex layouts such as multi-column formats and marginal annotations to accurately extract text. AI-based handwriting style classification will also improve recognition by classifying handwriting styles and applying specific models for each type, which may improve accuracy in interpreting regional scripts and historical documents. Another major area of improvement is adaptive learning mechanisms. It can be enhanced by incorporating a self-learning feedback system, whereby the OCR model can dynamically adapt its accuracy through user corrections and evolving handwriting patterns. Real-time processing capabilities may also be improved through optimized computational efficiency and cloud-based processing to enhance scalability for large-scale digitization projects. This would facilitate faster text recognition and translation.

Ensuring accessibility compliance is essential to broader usability. Features such as text-to-speech integration may enhance engagement for visually impaired people, making historical records more accessible. In addition, blockchain-based verification mechanisms may secure the integrity and authenticity of documents, benefitting legal and archival institutions requiring tamper-proof preservation of digitized records. This developed solution will show how an AI-driven OCR system can actually help preserve cultural heritage through historical documents made available and searchable easily. Future developments may include increasing language support, incorporating more powerful deep learning models, and improving real-time processing capabilities to make the system even more user-friendly and efficient. Other options may include adding user-driven customization options, such as adjustable OCR sensitivity and manual text correction features, to improve system efficiency. Further advances will make this AI-based OCR solution a landmark for large-scale historical document digitization, hence preserving and making accessible valuable cultural records in the long run.

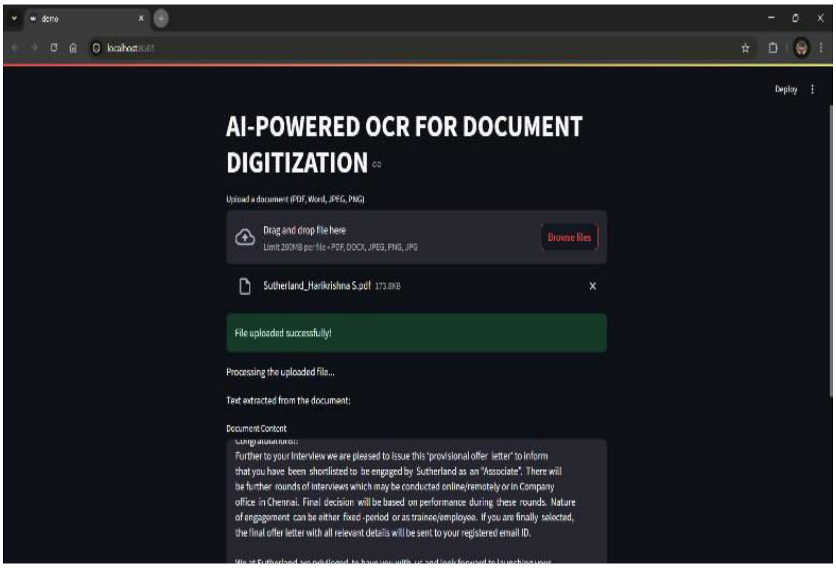


FIGURE 3. OUTPUT OF OCR DOCUMENT DIGITIZATION

VI. CONCLUSION

The AI-powered OCR system developed during this project proved to be an effective robust solution for the high-accuracy digitization of historical handwritten documents. By utilizing state-of-the-art OCR models and deep learning techniques along with NLP-based refinement, the system effectively identifies the diversity of handwriting styles and can compensate for the degradation of a document. Integrating multilingual translation with the user-friendly interface makes historical records accessible to more people, researchers, historians, and the public at large.

Despite these notable achievements, it has significant challenges in managing highly degraded documents and improving recognition accuracy for complex regional scripts. Improvements from future updates should, therefore, be

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

developed as adaptive learning mechanisms that enable the system to correct its accuracy dynamically based on real-time user feedback. Furthermore, optimizing real-time processing capabilities and cloud-based infrastructure can enhance scalability, enabling large-scale digitization projects with improved efficiency.
Ahead, blockchain-based authentication of documents would further strengthen the reliability and security of digitized records, authenticating them while preventing tampering. Continued research and innovation in AI-powered OCR technology will be able to help preserve cultural heritage by making historical records more accessible and searchable globally. This system can be the benchmark for digitization initiatives to come; by addressing emerging limitations, it can assure long-term preservation of and democratization access to valuable historical documents.

In conclusion, the developed solution shows the opportunity of AI-powered OCR systems towards the preservation of cultural heritage using easily accessible, searchable historical texts. Future extension could be expanded language support; integration of strong deep learning-based models; or real-time process enhancement to ease usability and raise performance. For further efficiency enhancement, user-controlled customization options are also recommended that include adjustable sensitivity of OCR features and manual corrections of text provided by the end-users. With further development, this OCR AI solution could help become a precedent for large-scale historical document digitization efforts as it contributes towards the long-term preservation and availability of valuable records of culture.

REFERENCES

1. D. Kavitha, P. Shamini, “Handwritten Document into Digitized Text Into Using Segmentation Algorithm”. Department of Computer Application, Easwari Engineering College, Tamil Nadu (ISSN2014) .
2. Jamshed Memon, Maira Sami, Rizwan Ahmed Khan. “Handwritten Optical Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”. Faculty of IT, Barrett Hodgson University1 January, 2020.
3. G. Xiao, J. Lin, M. Seznec, and H. Wu, “Smoothquant: Accurate and efficient post-training quantization for large language models,” International, 2023.
4. L. Yan et al., “Practical and ethical challenges of large language models in education: A systematic scoping review,” arXiv [cs.CL], 17-Mar-2023.
5. C. B. Head, P. M. McConnachie, L. Raftree, and G. Jasper, Higdon, “Large language model applications Opportunities for and evaluation: ethical implications,” New Dir. Eval., vol. 2023, no. 178–179, pp. 33–46, Jun. 2023..
6. J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. McHardy, Raileanu, and R. “Challenges and Applications of Large Language Models,” arXiv [cs.CL], 19-Jul-2023.
7. G. A. Cutting and A.-F. Cutting Decelle, “Intelligent Document Processing -- Methods and Tools in the real world,” arXiv [cs.IR], 28-Dec 2021.
8. D. Oniani et al., “From Military to Healthcare: Adopting and Expanding Ethical Principles for Generative Artificial Intelligence,” arXiv [cs.CY], 04-Aug-2023.
9. S. V. Rice, G. Nagy, and T. A. Nartker, Optical character recognition. Dordrecht, Netherlands: Springer, 2013.
10. Ruchika Malhotra, "Handwritten Amharic Word Recognition Using Additive Attention Mechanism," IEEE Trans. Multimedia, vol. 21, no. 1, pp. 221–233, Jan. 2024.
11. J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," in IEEE Access, vol. 8, pp. 142642-142668, 2020, doi: 10.1109/ACCESS.2020. 3012542.
12. B. Vidhale, G. Khekare, C. Dhule, P. Chandankhede, A. Titarmare and M. Tayade, "Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks," 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2021, pp. 1 - 5, doi: 10.1109/ I2CT51068.2021.9418106
13. V. V. Mainkar, A. B. Upade, J. A. Katkar, and P. R. Pednekar, "Handwritten Character Recognition to Obtain Editable Text," in Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC), Kankavli, India, 2020, pp. 599-602. DOI:10.1109/ ICESC48915.2020.9155786.
14. Bojja, Professor.Polaiah & Bojja, Polaiah & Velpuri, Naga & Pandala, Gautham & Polavarapu, S. (2019). Handwritten Text Recognition using Machine Learning Techniques in Application of NLP. International Journal of Innovative Technology and Exploring Engineering. 9. 1394-1397. 10.35940/ijitee.A4748.129219.
15. Yadav, Rohit. (2022). Optical Character Recognition Based Webapp. International Journal of Advanced Research in Science, Communication and Technology. 480-484. 10.48175/IJARSCT-5488.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

16. Deepthi, Aruna & Raol E. Sreenivasa Raol, M. D. Shadab farhan (2022). Handwritten to Text Document Converter. 10.1007/978-981-19-5550-1_18.
17. A. B. M. Ashikur Rahman et al., "Two Decades of Bengali Handwritten Digit Recognition: A Survey," in IEEE Access, vol. 10, pp. 92597-92632, 2022, doi: 10.1109/ACCESS.2022.3202893.
18. Automated Handwritten Text Recognition R.P. Ram Kumar, A. Chandra Prasad, K. Vishnuvardhan, K. Bhuvanesh, Sanjeev Dhama E3S Web Conf. 430 01022 (2023) DOI: 10.1051/e3sconf/202343001022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE
SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details