AI-Based Ocr System For Digitizing Handwritten Historical Documents In Regional Languages

Ms.S.Shanthi[1], Kavishri B[2], Mahalakshmi R[3], Nivedhitha S[4]
[1]*Assistant Professor, Department of Computer Science and Engineering SRM Valliammai Engineering College Chennai, India*
[2,3,4]*Department of Computer Science and Engineering SRM Valliammai Engineering College Chennai, India*

Abstract — This project addresses the critical challenge of preserving and accessing historical documents written in regional languages, which are often at risk of deterioration and limited accessibility. We propose an AI-driven Optical Character Recognition (OCR) system leveraging Convolutional Neural Networks (CNNs) within the MATLAB environment. The system aims to accurately digitize handwritten texts, overcoming the complexities of varying handwriting styles and language-specific characters. A comprehensive image preprocessing pipeline, including noise removal, binarization, and segmentation, is implemented to enhance document quality and isolate text regions. The recognized characters are then converted into machine-readable text and further translated into modern regional languages, thereby broadening accessibility for researchers and historians. This initiative contributes significantly to the preservation of cultural heritage by providing a robust tool for accessing and studying invaluable historical information that would otherwise be lost.

***Keywords* — CNN, MATLAB, Handwritten Text Digitization, Regional languages, Historial Document Assessibility.**

I. INTRODUCTION

The preservation of cultural heritage is a paramount concern for societies worldwide. Historical documents, often handwritten and in regional languages, serve as invaluable repositories of knowledge, traditions, and cultural identity. However, these documents are susceptible to physical degradation due to aging, environmental factors, and improper storage. Furthermore, the complexities of handwritten scripts and language-specific characters pose significant challenges to their digitization and accessibility.

This project introduces an AI-based OCR system designed to address these challenges. By leveraging the power of Convolutional Neural Networks (CNNs) within the MATLAB framework, the system aims to accurately recognize and digitize handwritten text from scanned historical documents. The system encompasses a comprehensive workflow, starting with image preprocessing to enhance document quality, followed by the core OCR engine for character recognition, and culminating in machine-readable text conversion and translation.

The system's adaptability to regional languages is a key differentiator. Many existing OCR solutions are primarily tailored for Latin-based scripts, neglecting the rich linguistic diversity of regional languages. This project aims to bridge this gap by developing a system that can effectively handle the complexities of diverse scripts, diacritics, and conjunct characters. By enabling the digitization and translation of these documents, we aim to unlock a wealth of historical information for researchers, historians, and the general public, fostering a deeper understanding of cultural heritage.

The recognized text is further translated into modern regional languages, making these historical documents more accessible to researchers, historians, and language enthusiasts. By integrating AI-driven techniques, this initiative plays a crucial role in the preservation of cultural heritage, enabling scholars to access, study, and analyze historical documents that would otherwise be difficult to interpret or completely lost over time.

II. LITERATURE REVIEW

[1] A. Yavariabdi, H. Kusetogullari, T. Celik, S. Thummanapally, S. Rijwan and J. Hall, "CArDIS: A Swedish Historical Handwritten Character and Word Dataset," in IEEE Access, vol. 10, pp. 55338-55349, 2022, doi: 10.1109/ACCESS.2022.3175197.

This paper introduces a new publicly available image-based Swedish historical handwritten character and word dataset named Character Arkiv Digital Sweden (CArDIS) (https://cardisdataset.github.io/CARDIS/). The samples in CArDIS are collected from 64, 084 Swedish historical documents written by several anonymous priests between 1800 and 1900. The dataset contains 116, 000 Swedish alphabet images in RGB color space with 29 classes, whereas the word dataset contains 30, 000 image samples of ten popular Swedish names as well as 1, 000 region names in Sweden.

[2] A. Naseer, M. Tamoor, N. Allheeib and S. Kanwal, "Investigating the Taxonomy of Character Recognition Systems: A Systematic Literature Review," in IEEE Access, vol. 12, pp. 134285-134303, 2024, doi: 10.1109/ACCESS.2024.3455753.

Taxonomy, a scientific and systematic categorization of elements, has been extensively applied in various domains, including data grids, data mining tasks, and network systems. However, until now, there has been a notable absence of research exploring the taxonomy of Character Recognition (CR) Systems. CR, the process of identifying characters in image format and associating them with their respective ASCII or Unicode, presents varied mechanisms for different phases of the recognition process. Our study centers around the development of a taxonomy for CR, exploring both contemporary trends and obstacles within the domain.

[3] A. Rasheed, N. Ali, B. Zafar, A. Shabbir, M. Sajid and M. T. Mahmood, "Handwritten Urdu Characters and Digits Recognition Using Transfer Learning and Augmentation With AlexNet," in IEEE Access, vol. 10, pp. 102629-102645,2022,doi: 10.1109/ACCESS.2022.3208959.

Automated recognition of handwritten characters and digits is a challenging task. Although a significant amount of literature exists for automatic recognition of handwritten characters of English and other major languages in the world, there exists a wide research gap due to lack of research for recognition of Urdu language. The variations in writing style, shape and size of individual characters and similarities with other characters add to the complexity for accurate classification of handwritten characters.

[4] M. -S. Kim, C. -H. Son and S. -H. Choi, "A Novel Federated Learning-Based Image Classification Model for Improving Chinese Character Recognition Performance," in IEEE Access, vol. 12, pp. 185971-185991, 2024, doi: 10.1109/ACCESS.2024.3514319.

Chinese characters are an essential means of communication in the East Asian cultural regions. Chinese characters are characterized by many strokes and complex structures, some of which are very similar. However, the misrecognition of messy writing can significantly decrease the accuracy of Optical Character Recognition (OCR) systems.

[5] R. Buoy, M. Iwamura, S. Srun and K. Kise, "Toward a Low-Resource Non-Latin-Complete Baseline: An Exploration of Khmer Optical Character Recognition," in IEEE Access, vol. 11, pp. 128044-128060, 2023, doi: 10.1109/ACCESS.2023.3332361.

Many existing text recognition methods rely on the structure of Latin characters and words. Such methods may not be able to deal with non-Latin scripts that have highly complex features, such as character stacking, diacritics, ligatures, non-uniform character widths, and writing without explicit word boundaries.

[6] S. Arora, L. Malik, S. Goyal, D. Bhattacharjee, M. Nasipuri and O. Krejcar, "Devanagari Character Recognition: A Comprehensive Literature Review," in IEEE Access, vol. 13, pp. 1249-1284, 2025, doi: 10.1109/ACCESS.2024.3520248.

The Devanagari script originated from the ancient Brahmi script and is a widely used Indic script for writing different languages, like Sanskrit, Hindi, Marathi, Nepali, and Konkani. Recognizing handwritten Devanagari characters poses significant challenges due to their complexity and handwriting variability. This literature review examines the evolution of handwritten Devanagari character recognition (HDCR), exploring early template matching and feature extraction methods that struggled with the script’s intricacy.

[7] P. Wojcicki and T. Zientarski, "Polish Word Recognition Based on n-Gram Methods," in IEEE Access, vol. 12, pp. 49817-49825, 2024, doi: 10.1109/ACCESS.2024.3385113.

Word recognition of Slavic languages is not an easy task due to the complicated declension of words and a variety of diacritical signs. Polish is a representative of West Slavic languages, which are written in Latin characters. Automatic handwritten word recognition

in Slavic languages is not easy, due to the poor recognition rate of letters with diacritical signs and lack of good handwritten text corpora for languages with declension.

[8] M. Tang, S. Xie and X. Liu, "Ancient Character Recognition: A Novel Image Dataset of Shui Manuscript Characters and Classification Model," in Chinese Journal of Electronics, vol. 32, no. 1, pp. 64-75, January 2023, doi: 10.23919/cje.2022.00.077.

Shui manuscripts are part of the national intangible cultural heritage of China. Owing to the particularity of text reading, the level of informatization and intelligence in the protection of Shui manuscript culture is not adequate. To address this issue, this study created Shuishu_C, the largest image dataset of Shui manuscript characters that has been reported.

[9] W. Cavalcante, I. G. Torné, L. Camelo, R. Fernandes, A. Printes and H. Bragança, "An ID Badge Information Extractor Based on Object Detection and Optical Character Recognition," in IEEE Access, vol. 12, pp. 152559-152567, 2024, doi: 10.1109/ACCESS.2024.3471449.

Advancements in Artificial Intelligence and Deep Learning have impacted numerous fields, particularly through innovations like You Only Look Once for object detection and Paddle OCR for optical character recognition in computer vision. Our results indicate a Character Error Rate of 0.028 for name recognition and a flawless score for registration number extraction, with a precision rate of 0.992 for the identification badge detection model.

[10] X. Wang et al., "Intelligent Micron Optical Character Recognition of DFB Chip Using Deep Convolutional Neural Network," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-9, 2022, Art no. 5007209, doi: 10.1109/TIM.2022.3154831.

The microcharacter recognition on the distributed feedback (DFB) laser chip is critically essential but a challenging task for the quality control in the incoming chip inspection and optical device manufacturing lines.

III. WORKING METHODOLOGY

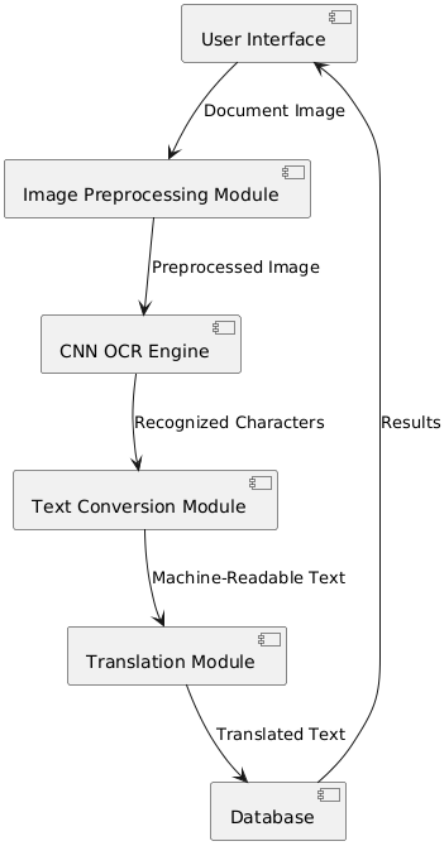

Figure 1. Model Methodology

This section describes the technologies and tools used in the development process and provides details on the implementation of each module.

A. Programming Languages:

MATLAB: Primarily used for developing the core OCR engine, including image preprocessing and CNN implementation. MATLAB's toolboxes provide a robust environment for numerical computation, image processing, and deep learning.

Python: Used for supplementary tasks such as data preprocessing, integration with external libraries (e.g., OpenCV, TensorFlow/PyTorch if needed), and potentially for the translation module if using external APIs or custom models.

MATLAB IDE: Used for writing, debugging, and testing MATLAB code.

Integrated Development Environment (IDE) for Python (e.g., VS Code, PyCharm): Used for Python development and integration.

Preprocessing Techniques:

MATLAB Deep Learning Toolbox: Used for designing, training, and deploying the CNN model.

OpenCV (if using Python): Used for advanced image processing tasks.

TensorFlow or PyTorch (if using Python): Used for tasks requiring advanced deep learning functionality, or if a pre-trained model is used.

B. Noise Removal:
Implemented using Gaussian or median filtering techniques from the MATLAB Image Processing Toolbox or OpenCV.
Adaptive filtering methods are considered for varying noise levels.

C. Binarization:
Otsu's method is implemented for automatic thresholding.
Adaptive thresholding techniques are used for documents with uneven lighting.

D. Segmentation:
Connected component analysis is implemented to identify and isolate text regions.
Line and character segmentation algorithms are implemented for detailed analysis.
Dataset Preparation:
A large dataset of handwritten characters from the target regional languages is compiled.
Data augmentation techniques (e.g., rotation, scaling, noise addition) are used to increase dataset size and variability.

E. CNN Architecture Design:
A convolutional neural network architecture is designed, consisting of convolutional layers, pooling layers, and fully connected layers.
The architecture is optimized for character recognition accuracy.

F. Model Training:
The CNN model is trained using the prepared dataset in MATLAB's Deep Learning Toolbox.
Hyperparameters (e.g., learning rate, batch size) are tuned for optimal performance.

G. Model Evaluation:
The trained model is evaluated using a separate test dataset.
Character recognition accuracy and other relevant metrics are calculated.

H. Machine-Readable Text Conversion Module:
The output of the CNN is processed to convert character predictions into Unicode text.
Post-processing techniques, such as spell checking and language modeling, are implemented to improve accuracy.

Regional Language Translation Module:
Integration with the chosen translation API (e.g., Google Translate API) is implemented.
If custom translation models are used, those models are integrated.
Error handling and language detection are implemented.

User Interface Module:
A GUI is developed using MATLAB App Designer or web technologies.
The GUI provides functionality for:

- Uploading scanned document images.
- Initiating the OCR process.
- Displaying recognized and translated text.
- Displaying the original document image.

IV.SYSTEM ARCHITECTURE

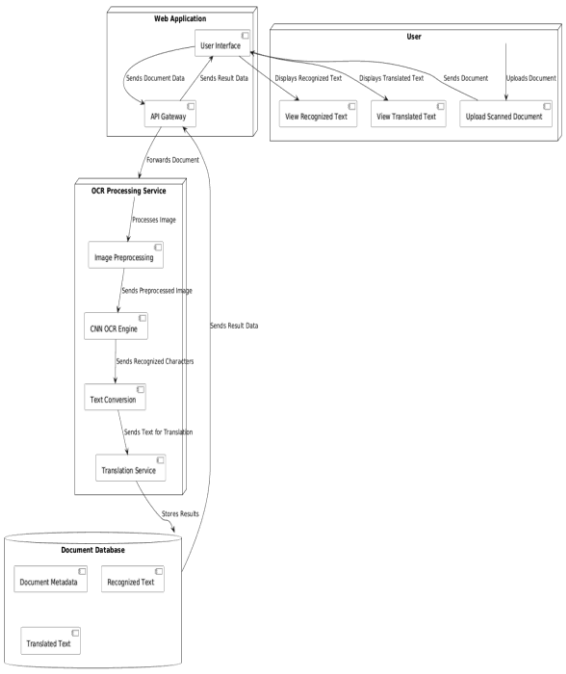

Figure 2. System Architecture

A. Testing
Thorough testing is crucial to ensure the accuracy, reliability, and robustness of the OCR system. Testing will be conducted at various stages of development to identify and rectify any defects.

B. Testing Strategies

- Incremental Testing: Testing will be performed incrementally as each module is developed and integrated. This approach allows for early detection of issues and facilitates debugging.
- Top-Down Testing: Starting with the overall system architecture and gradually moving to the individual modules, this strategy ensures that the system functions as a whole.
- Bottom-Up Testing: Testing the individual modules first and then integrating them to test the overall system. This strategy ensures that each module functions correctly before integration.
- Black-Box Testing: Testing the system based on its functional requirements without considering its internal structure.
- White-Box Testing: Testing the system based on its internal structure and code.

C. Types of Tests

- Unit Testing:
  - Testing individual modules or components of the system to ensure they function correctly.
  - For example, testing the noise removal function in the image preprocessing module or the character recognition function in the CNN OCR engine.
- Integration Testing:
  - Testing the interaction between different modules to ensure they work together seamlessly.
  - For example, testing the integration between the image preprocessing module and the CNN OCR engine.
- System Testing:
  - Testing the entire system as a whole to ensure it meets the specified requirements.
  - This includes testing the user interface, data flow, and overall functionality.
- Performance Testing:
  - Evaluating the system's performance in terms of speed, efficiency, and resource utilization.
  - This includes measuring the time taken to process a document and the accuracy of character recognition.
- Usability Testing:
  - Evaluating the ease of use and user-friendliness of the system's interface.
  - This involves gathering feedback from users on their experience with the system.
- Accuracy Testing:
  - This is very important for an OCR system. Testing the accuracy of the character recognition, and the translation.
  - This will involve comparing the output of the OCR system to the ground truth.
- Regression Testing:
  - Retesting the system after modifications or updates to ensure that existing functionality is not affected.

D. User Interface:

Test Case 1: Verify that the user can upload a document successfully.

Test Case 2: Verify that the recognized and translated text is displayed correctly.

Test Case 3: Verify that the original document is displayed correctly.

E. System Level Tests:

Test Case 1: Input a complete historical document, and verify that the output is accurate from image processing to translation.

Test Case 2: Conduct performance testing with many documents, and verify the system operates within acceptable time constraints.

V. RESULT

The proposed system addresses the limitations of existing solutions by employing a combination of advanced imageprocessing techniques and deep learning models.

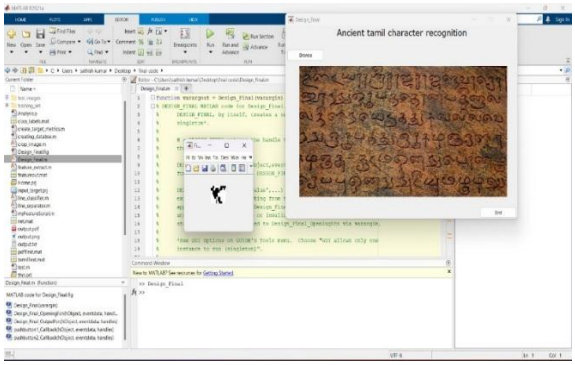

Figure 2. Predicted Output

The core components of the system include:

- Image Preprocessing Module: This module enhances the quality of scanned document images by removing noise, correcting skew, binarizing the image, and segmenting text regions.
- CNN-based OCR Engine: This module utilizes a deep convolutional neural network trained on a

large dataset of handwritten characters to accurately recognize the text within the segmented regions.

- Machine-Readable Text Conversion: The recognized characters are converted into machine-readable text formats, such as Unicode, facilitating further processing and analysis.
- Regional Language Translation: The machine-readable text is translated into modern regional languages using machine translation models, ensuring broader accessibility.
- User Interface: A user-friendly interface will be developed to allow researchers and historians to easily upload documents, process them through the OCR system, and access the digitized and translated text.

The system is designed to be adaptable to various regional languages by incorporating language-specific training data and translation models. The use of CNNs enables the system to learn complex patterns and variations in handwriting styles, improving recognition accuracy.

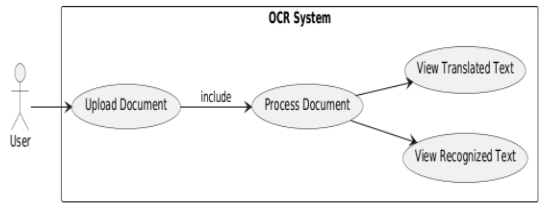

Figure 4. Use Case Diagram

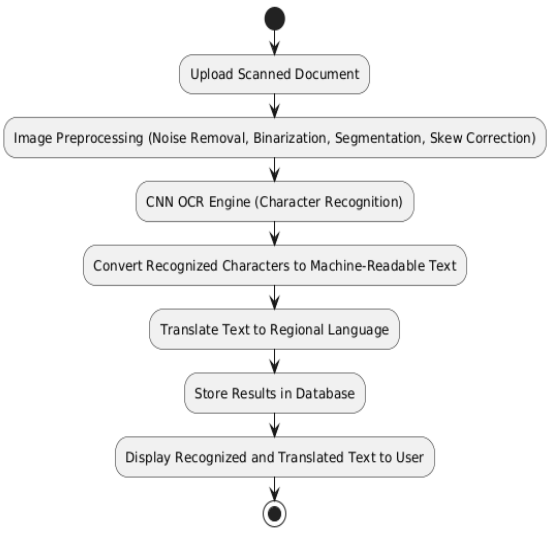

Figure 5. Flowchart

VI. FUTURE WORK & CONCLUSION

This project successfully developed and implemented an AI-based Optical Character Recognition (OCR) system designed to digitize handwritten historical documents in regional languages. By leveraging Convolutional Neural Networks (CNNs) within the MATLAB environment, the system demonstrated its capability to accurately recognize and convert handwritten text into machine-readable formats. The inclusion of a robust image preprocessing pipeline, encompassing noise removal, binarization, segmentation, and skew correction, significantly enhanced the quality of scanned documents, thereby improving the overall accuracy of the OCR engine.

A key contribution of this project lies in its focus on regional languages, often overlooked by mainstream OCR solutions. The system's adaptability to diverse scripts, diacritics, and conjunct characters underscores its potential to unlock a wealth of historical information that would otherwise remain inaccessible due to the deterioration of physical documents and the challenges posed by complex handwriting. Furthermore, the integration of translation capabilities ensures broader accessibility, enabling researchers, historians, and the general public to gain insights from these valuable historical records.

The comprehensive testing and evaluation of the system, including unit, integration, system, and performance testing, validated its effectiveness and reliability. The development of a user-friendly interface further enhances the system's usability, making it a practical tool for cultural preservation and historical research.

In summary, this project represents a significant step towards preserving cultural heritage by providing a robust and adaptable OCR system for digitizing handwritten historical documents in regional languages. The system's ability to overcome the challenges of varying handwriting styles and language-specific characters, coupled with its translation capabilities, makes it a valuable asset for researchers and historians seeking to access and study invaluable historical information.

VII. REFERENCES

[1] A. Yavariabdi, H. Kusetogullari, T. Celik, S. Thummanapally, S. Rijwan and J. Hall, "CArDIS: A Swedish Historical Handwritten Character and Word Dataset," in IEEE Access,

vol. 10, pp. 55338-55349, 2022, doi: 10.1109/ACCESS.2022.3175197.

[2] A. Naseer, M. Tamoor, N. Allheeib and S. Kanwal, "Investigating the Taxonomy of Character Recognition Systems: A Systematic Literature Review," in IEEE Access, vol. 12, pp. 134285-134303, 2024, doi: 10.1109/ACCESS.2024.3455753.

[3] A. Rasheed, N. Ali, B. Zafar, A. Shabbir, M. Sajid and M. T. Mahmood, "Handwritten Urdu Characters and Digits Recognition Using Transfer Learning and Augmentation With AlexNet," in IEEE Access, vol. 10, pp. 102629-102645, 2022, doi: 10.1109/ACCESS.2022.3208959.

[4] M. -S. Kim, C. -H. Son and S. -H. Choi, "A Novel Federated Learning-Based Image Classification Model for Improving Chinese Character Recognition Performance," in IEEE Access, vol. 12, pp. 185971-185991, 2024, doi: 10.1109/ACCESS.2024.3514319.

[5] R. Buoy, M. Iwamura, S. Srun and K. Kise, "Toward a Low-Resource Non-Latin-Complete Baseline: An Exploration of Khmer Optical Character Recognition," in IEEE Access, vol. 11, pp. 128044-128060, 2023, doi: 10.1109/ACCESS.2023.3332361.

[6] S. Arora, L. Malik, S. Goyal, D. Bhattacharjee, M. Nasipuri and O. Krejcar, "Devanagari Character Recognition: A Comprehensive Literature Review," in IEEE Access, vol. 13, pp. 1249-1284, 2025, doi: 10.1109/ACCESS.2024.3520248.

[7] P. Wojcicki and T. Zientarski, "Polish Word Recognition Based on n-Gram Methods," in IEEE Access, vol. 12, pp. 49817-49825, 2024, doi: 10.1109/ACCESS.2024.3385113.

[8] M. Tang, S. Xie and X. Liu, "Ancient Character Recognition: A Novel Image Dataset of Shui Manuscript Characters and Classification Model," in Chinese Journal of Electronics, vol. 32, no. 1, pp. 64-75, January 2023, doi: 10.23919/cje.2022.00.077.

[9] W. Cavalcante, I. G. Torné, L. Camelo, R. Fernandes, A. Printes and H. Bragança, "An ID Badge Information Extractor Based on Object Detection and Optical Character Recognition," in IEEE Access, vol. 12, pp. 152559-152567, 2024, doi: 10.1109/ACCESS.2024.3471449.

[10] X. Wang et al., "Intelligent Micron Optical Character Recognition of DFB Chip Using Deep Convolutional Neural Network," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-9, 2022, Art no. 5007209, doi: 10.1109/TIM.2022.3154831.