TAMILNADU SKILL DEVELOPMENT CORPORATION

NAAN MUDHALVAN NIRAL THIRUVIZHA

PROJECT PROPOSAL

Theme: Artificial Intelligence

Problem Statement:

How might we develop an AI or OCR solution to digitize and convert handwritten, old registered documents into a readable and accessible format in regional languages improving public access and readability of historical records?

College Code & College Name: 9213 - PSNA College of Engineering and Technology

Guide Name: Mrs. A. Sangeetha

Designation: Assistant Professor

Mobile No.: 8760963095

Email Id: sangeetha@psnacet.edu.in

Student Team Details:

S.No.	Student Reg. No.	Name of the Student	Branch	Mobile No.	Email Id
1	921321205029	Charankumar E G D	Information Technology	9489014033	charankumaregd21it@psnacet.edu.in
2	921321205015	Arunprasad S	Information Technology	6374824898	arunprasads21it@psnacet.edu.in
3	921321205032	Dharani Dharan R	Information Technology	7395877359	dharanidharanr21it@psnacet.edu.in

Project Summary:

The preservation and accessibility of handwritten old documents, particularly those in regional languages, pose significant challenges due to the diversity in handwriting styles, material degradation over time, and the complexities of translating text into regional languages. These challenges limit public access to valuable documents, emphasizing the need for efficient digitization solutions.

This idea presents an Optical Character Recognition (OCR)-based solution aimed at digitizing handwritten, old registered documents and translating them into regional languages. The proposed system utilizes advanced OCR technology to convert scanned images of handwritten documents into machine-readable text. Furthermore, a regional language translation model is integrated to ensure the translation of the digitized content into local languages, thereby enhancing public accessibility.

Preliminary results indicate that the OCR system efficiently handles various handwriting styles, while the translation model ensures the generation of region-specific language outputs. This approach offers a promising solution for preserving and making old documents more accessible to a broader audience.

Proposed Solution with Methodology:

The proposed solution utilizes an Optical Character Recognition (OCR)-based system integrated with a regional language translation model to digitize handwritten, old documents and make them accessible in regional languages.

The methodology begins with image preprocessing, where scanned images of handwritten documents are enhanced through noise removal and other techniques to improve input data quality. Advanced OCR algorithms are then applied to recognize and extract text from these images, designed to effectively handle diverse handwriting styles.

Once the text is extracted, a language translation model identifies the source language and translates it into the desired regional language. The outputs are validated for accuracy to ensure high-quality results. Users receive both the extracted text and its translation in a downloadable PDF format, facilitating easy access and usability.

Workplan / Time Schedule Indicating Project Milestones:

- Week 1-2: Requirement gathering, literature survey, and dataset collection.
- Week 3-4: Develop preprocessing pipelines and integrate OCR functionality.
- Week 5-6: Implement and fine-tune the regional language translation model.
- Week 7: Validate the combined system on sample handwritten documents.
- Week 8: Develop the user interface (UI) and integrate frontend and backend.
- Week 9-10: Test the system for performance, accuracy, and usability.
- Week 11-12: Finalize the prototype and prepare project documentation.

Plan of Action for Implementation:

The implementation will begin by developing a Flask-based backend to process image inputs, perform OCR, and execute translation tasks. Concurrently, a React-based frontend will be designed to enable user interactions, such as uploading images and downloading processed outputs.

The OCR model will be trained using a diverse dataset of handwritten samples to ensure robust recognition capabilities. For translation, an existing neural machine translation model will be fine-tuned to handle specific regional languages. The integration of these components will be followed by rigorous testing and optimization to ensure the system meets performance benchmarks.

List of Facilities Available in the College to Develop the Prototype:

- High-performance computing systems in the college lab for training OCR and translation models.
- Access to image processing and machine learning tools such as Python libraries (OpenCV, TensorFlow).
- Internet connectivity for research, downloading pre-trained models, and accessing cloud-based APIs if needed.
- Technical support from faculty members in computer science and linguistics.

Nature of Industry Support for the Project (if any):

Industry support could include collaboration with organizations involved in document digitization or archiving to access additional datasets of handwritten documents and regional language corpora. This support could also extend to providing guidance for refining OCR and translation models and aligning the system with industry standards. Such partnerships would enhance the quality and applicability of the final solution.

Total Cost:

The estimated total cost of the project is ₹10,000, which covers software and development expenses.

Details of Financial Assistance Required (Limited to Rs 10,000/-):

- Dataset Procurement: ₹3,000
- Cloud Storage and Computing: ₹2,000
- API and Software Tool Licenses: ₹2,000
- Hosting and Management Costs: ₹3,000

Expected Outcomes / Results:

The project is expected to deliver a robust and user-friendly solution for digitizing handwritten, old documents and translating them into regional languages. The system will efficiently extract text from scanned images of handwritten documents, regardless of handwriting style variations, using advanced OCR technology. Additionally, the integrated regional language translation model will provide accurate translations, ensuring these documents are accessible to a broader audience. The final output will include the digitized and translated text, available for download as a PDF for ease of use.

This solution will not only help preserve cultural heritage but also bridge the gap in public access to valuable old documents. Furthermore, the prototype's modular design will enable scalability to process various document types and extend support for additional languages in future implementations.