WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Based OCR for Digitizing and Translating Historical Handwritten Records

Mrs. K. Srisathiya[1], Vithya D[2], Jeevitha M[3], Dhanishwar B[4], Lenin Kumar R [5]

[1]*Assistant Professor, Department of Artificial Intelligence and Data Science, Dhirajlal Gandhi College of Technology (Autonomous), Salem - 636309*
[2345] *Department of Artificial Intelligence and Data Science, Dhirajlal Gandhi College of Technology (Autonomous), Salem – 63630*
[1]srisathiya.aids@dgct.ac.in, [2]vithya.d23@gmail.com, [3]jeevithaaids@gmail.com, [4]shridhanu1969@gmail.com, [5] vijaylenin@gmail.com
[1]srisathiya.aids@dgct.ac.in, [2]gokulavani.ai.ds@gmail.com, [3]mehala0609@gmail.com, [4]ramyakuberan56@gmail.com, [5]snehaammu364@gmail.com

ABSTRACT

This project presents a deep learning-based approach for recognizing ancient Tamil inscriptions and converting them into modern Tamil characters for improved readability and digital preservation. The dataset consists of inscription images sourced from two archives containing stone-engraved Tamil scripts. To process these images, a multi-stage pipeline is developed comprising image preprocessing, character segmentation, character recognition, and translation into contemporary Tamil script.

Initially, the input images are cleaned and enhanced using techniques like binarization, noise reduction, and rotation correction. The pre-processed images are then subjected to character segmentation using bounding box algorithms to isolate individual characters. Each segmented character is classified using a trained Convolutional Neural Network (CNN) model built using labelled historical characters data.

The project handles both single and multi-part Tamil characters using specialised datasets and recognition models. Once identified, the recognised characters are detected object and the alert status. This smart surveillance solution enhances security in remote or sensitive areas, such.

The project handles both single and multi-part Tamil characters using specialized datasets and recognition models. Once identified, the recognized characters are matched against their corresponding modern Tamil equivalents to reconstruct the complete text. The entire system is implemented in Python using Jupyter. Notebooks and TensorFlow, with all processing done in the Tamil language. This system enables automated digitization and understanding of heritage Tamil inscriptions, making them accessible to researchers, linguists, and the public.

1. INTRODUCTION

1.1. GENERAL:

Tamil is one of the world's oldest living languages, with a rich history documented in stone inscriptions across temples, monuments, and archaeological sites. These inscriptions, carved centuries ago, contain immense historical, cultural, and linguistic significance.

This system leverages traditional image processing techniques alongside deep learning models, specifically Convolutional Neural Networks (CNNs), to detect, segment, and classify historical characters extracted from stone-carved images.

The system is trained using labelled datasets that include single and multipart characters, allowing for accurate recognition of a diverse set of Tamil glyphs. The character recognition process is followed by mapping these characters to their modern equivalents, ensuring a comprehensive transliteration of the original text.

1.2 AUTOMATED INSCRIPTION TRANSLITERATION SYSTEM:

1.2.1 CHARACTER RECOGNITION:

Character recognition involves extracting text from ancient stone inscriptions through image preprocessing and deep learning-based classification. The system identifies and distinguishes historical Tamil characters by employing CNN models, ensuring accurate recognition despite script variations and erosion.

1.2.2 MAPPING TO MODERN SCRIPT:

After character identification, the recognized ancient Tamil characters are mapped to their modern equivalents. This transliteration process ensures historical texts remain comprehensible in contemporary Tamil, facilitating academic research and cultural preservation.

1.2.4 APPLICATION & RECENT TRENDS:

The application of Tamil script recognition using deep learning models extends to historical research, digital archives, educational tools, museum exhibits, and linguistic studies, enabling researchers and historians to analyses ancient Tamil texts more effectively while preserving them for future generations. Recent trends in this field include AI-powered Optical Character Recognition (OCR) systems, integration with augmented reality for real-time translations, crowdsourced data labelling to enhance model accuracy, cross-language comparisons for broader linguistic research, and blockchain technology for secure digital preservation of Tamil inscriptions.

2. LITERATURE SURVEY:

1. Title: Analysis of Various Deep Learning Algorithms for Tamil Character Recognition

Authors: Ashok Kumar L, Shalini J, Karthika Renuka D

Abstract: This study evaluates multiple deep learning models, including Convolutional Neural Networks (CNNs), YOLO, and ResNet, for Tamil character recognition. The research focuses on identifying the most effective model in terms of accuracy and computational efficiency for recognizing Tamil characters.

Published In: Proceedings of the First International Conference on Combinatorial and Optimization (ICCAP), December 2021.

2. Title: Effective Tamil Character Recognition Using Supervised Machine Learning Algorithms

Authors: Dr. S. Suriya, S. Nivetha, P. Pavithran, Ajay Venkat S., Sashwath K. G., Elakkiya G.

Abstract: The paper presents a supervised learning approach for Tamil character recognition, employing Convolutional Neural Networks (CNNs) to handle challenges such as distortions and variations in handwriting. The proposed system demonstrates high accuracy in recognizing complex Tamil characters.

Published In: EAI Endorsed Transactions on e-Learning, February 2023.

3. Title: A Novel Approach to OCR Using Image Recognition-Based Classification for Ancient Tamil Inscriptions in Temples

Authors: Lalitha Giridhar, Aishwarya Dharani, Velmathi Guruviah

Abstract: This research addresses the

challenges in recognizing ancient Tamil scripts from temple inscriptions dating between the 7th and 12th centuries. Utilizing Otsu thresholding for image binarization and a 2D CNN for classification, the system integrates with Tesseract OCR and Google's text-to-speech engine, achieving an accuracy of 77.7%.

Published In: arXiv preprint, July 2019.

4. Title: Handwritten Tamil Character Recognition and Digitization Using Deep Learning

Authors: N. Sasipriyaa, K. Abirami, G. Banupriya, S. Dhivya

Abstract: Focusing on the digitalization of handwritten Tamil characters, this study employs deep learning techniques to overcome challenges posed by the script's complexity and variability in handwriting.

Published In: International Journal of Advanced Science and Technology, March 2020

5. Title: Ancient Tamil Character Recognition Based on Edge Mapping Pointed Multi-Perspective Neural Network for Enhanced Font Definition

Author: P. Selvakumar

Abstract: This paper introduces a novel neural network architecture that utilises edge mapping and multi-perspective analysis to enhance the recognition of ancient Tamil characters. The approach addresses the challenges of font degradation and variability in historical inscriptions.

Published In: International Journal of Intelligent Systems and Applications in Engineering, March 2024.

3. EXISTING SYSTEM:

The existing systems for Tamil inscription recognition primarily rely on traditional OCR (Optical Character Recognition) methods and manual transcription. These systems process printed or handwritten Tamil characters using basic image processing techniques such as thresholding, contour detection, and segmentation. For historical inscriptions, techniques include preprocessing the images using grayscale conversion and noise removal, followed by character extraction. Some projects have used Tesseract OCR or template matching methods for recognition. These tools perform relatively well with modern printed Tamil texts, but they struggle with ancient inscriptions due to stone-engraved scripts, erosion, irregular spacing, and stylistic variations over centuries.

3.1 Disadvantages:

- Low accuracy for ancient and damaged inscriptions.
- Manual transcription is labor-intensive and slow.
- Inability to handle complex, stylized, or eroded characters.
- Poor generalization to various ancient script formats.
- Limited use of AI, lacking robust automated translation to modern Tamil**.**

4. PROPOSED SYSTEM:

The proposed system aims to automate the recognition and translation of ancient Tamil inscriptions into modern Tamil using advanced deep learning techniques. Initially, the input stone inscription images undergo preprocessing including noise removal, rotation correction, and enhancement. Then, character segmentation is applied to extract individual characters from the inscription. These segmented characters are passed through a Convolutional Neural Network (CNN)-based model trained on labeled datasets of ancient Tamil scripts. The model predicts the closest matching modern Tamil character. For multipart or composite characters, a separate classifier is employed to handle segmentation and combination accurately. Finally, recognized characters are concatenated and post-processed to form coherent modern Tamil sentences.

4.1 Advantages:

- Automated recognition of ancient Tamil characters.
- High accuracy using CNN-based deep learning models.
- Effective character segmentation, even for multipart characters.
- Fast and scalable compared to manual transcription.
- Preserves heritage by digitizing and translating ancient scripts.

5. SYSTEM REQUIREMENTS

5.1.1 Hardware Requirements:

- Processor: Intel Core i5 or higher
- RAM: Minimum 8 GB
- Storage: 10 GB (for datasets and model files)
- GPU (optional): NVIDIA GPU with CUDA support (for accelerated training and inference)

5.1.2 CCC:

- Operating System: Windows 10 / Ubuntu 20.04
- Python Version: 3.7 or higher
- Jupyter Notebook: For executing .ipynb files

5.1.3 Libraries/Frameworks:

- TensorFlow / Keras (CNN model development)

- OpenCV (image preprocessing and segmentation)
- NumPy (numerical operations)
- Pandas (data handling)
- Matplotlib (visualization)
- Scikit-learn (evaluation and clustering, if required)

5.2 Hardware Requirements:

In the realm of machine learning and deep learning projects, hardware plays a critical role in ensuring the efficient execution of complex models, especially those involving image processing, character segmentation, and neural network-based recognition. For the specific project of converting ancient Tamil inscriptions into modern readable Tamil text, we deal with high-resolution images, character extraction, and convolutional neural networks (CNNs), all of which demand adequate hardware.

5.2.1 Processor: Intel Core i5 or Higher

Why is a powerful processor important?

The processor, also known as the Central Processing Unit (CPU), is the brain of any computer system. It carries out the instructions of a computer program by performing basic arithmetic, logic, control, and input/output operations.

Minimum Requirement: Intel Core i5 or Higher

The Intel Core i5 is a mid-range processor from Intel that offers a good balance between performance and power consumption. It usually comes with 4 to 6 cores, which are beneficial for parallel processing.

Higher versions like Intel Core i7 or i9, or AMD Ryzen 7/9 can further improve speed and efficiency, especially during image processing and model training.

5.2.2 Tasks Handled by CPU in This Project:

- Reading and loading high-resolution images of Tamil inscriptions
- Performing image preprocessing operations such as rotation correction, resizing, thresholding, and contour detection using OpenCV
- Running character segmentation algorithms
- Managing dataset input/output operations
- Triggering and managing training and inference processes of deep learning models when a GPU is not available
- If the CPU is underpowered, image processing tasks and notebook execution may become sluggish, making development inefficient.

5.2.3 RAM: 8 GB Minimum

Why is RAM important?

RAM (Random Access Memory) temporarily stores data that the CPU needs to access quickly. In machine learning and deep learning, especially in image processing and model training, data often gets loaded into memory for rapid access.

Why 8 GB as the minimum?

In this project, you deal with multiple image files, sometimes as high-resolution grayscale or color images, which need to be processed in batches.

Character segmentation generates multiple image slices that must be temporarily held in memory for further classification.

During model training, batches of images, labels, and prediction results are constantly being loaded into memory.

Libraries like TensorFlow, Keras, and OpenCV are memory-intensive. Without at least 8 GB RAM, your system may freeze or crash.

5.2.4 Benefits of Higher RAM (16 GB or more):

- Allows you to run multiple Jupyter notebooks and programs in parallel
- Enables smoother model training and real-time monitoring
- Supports background tasks (like saving weights, visualizing outputs, or running Flask-based demos)

5.2.5 Storage: 10 GB Minimum

Why is storage essential?

Storage, typically a Hard Disk Drive (HDD) or Solid-State Drive (SSD), is used to store all files permanently, including datasets, models, scripts, logs, and reports.

Why is 10 GB considering the minimum for this project?

Raw image dataset: 66 images of inscriptions may seem small, but when expanded into segmented characters, labeled data, and augmented images, they multiply in size.

Model files: The CNN.model file along with intermediate training checkpoints and logs can occupy hundreds of megabytes.

5.2.6 GPU: NVIDIA with CUDA Support (Optional but Recommended)

Why is GPU important in deep learning?

A GPU (Graphics Processing Unit) is specially designed to handle complex matrix and vector operations. Since deep learning models, especially CNNs, rely heavily on such operations (convolutions, pooling, activations, etc.), a GPU can speed up the training process 10x to 50x compared to a CPU.

CUDA Support

NVIDIA GPUs support CUDA (Compute Unified Device Architecture), a parallel computing platform and API model that allows you to use the GPU for general-purpose processing.

Recommended GPU Models:

NVIDIA GTX 1050 Ti (entry-level)

NVIDIA GTX 1650 or 1660 (mid-range)

NVIDIA RTX 2060/3060/4060 (high performance)

Tasks where GPU accelerates this project:

Training the CNN.model on thousands of segmented characters

Performing real-time predictions or batch inference on new inscription images

Speeding up image augmentation and batch processing

What if no GPU is available?

- The model can still be trained using CPU, but it will take significantly longer.
- For instance, what takes 10 minutes on GPU might take 1 hour or more on CPU.
- You can use cloud services like Google Collab, which provide free GPU support for training models.

5.2.7 Summary Table:

5.3 Conclusion:

Component	Minimum Requirement	Role in Project
CPU	Intel Core i5 or higher	Executes image processing, model logic, and training (if no GPU)
RAM	8 GB	Loads image batches, supports training, and handles parallel processing
Storage	10 GB	Stores datasets, models, scripts, and checkpoints
GPU	NVIDIA with CUDA (Optional)	Accelerates CNN training, real-time recognition, and batch predictions

Adequate hardware is crucial for efficient execution of deep learning workflows, especially when handling high-resolution images and training CNN models on large character datasets. While basic configurations may support initial experimentation, optimal performance requires sufficient RAM and GPU acceleration. The specified hardware setup is tailored to the project's dataset size, model complexity, and software dependencies. For future scalability—such as multilingual support or mobile deployment—hardware upgrades are strongly recommended.

6. DETAILED EXPLANATION OF SOFTWARE REQUIREMENTS

In any machine learning or deep learning project—especially one that involves heavy image preprocessing, character segmentation, and training of neural network models—selecting the right software stack is essential. The choice of operating system, programming language, development environment, and supporting libraries/tools directly impacts the ease of development, performance, and maintainability of your project.

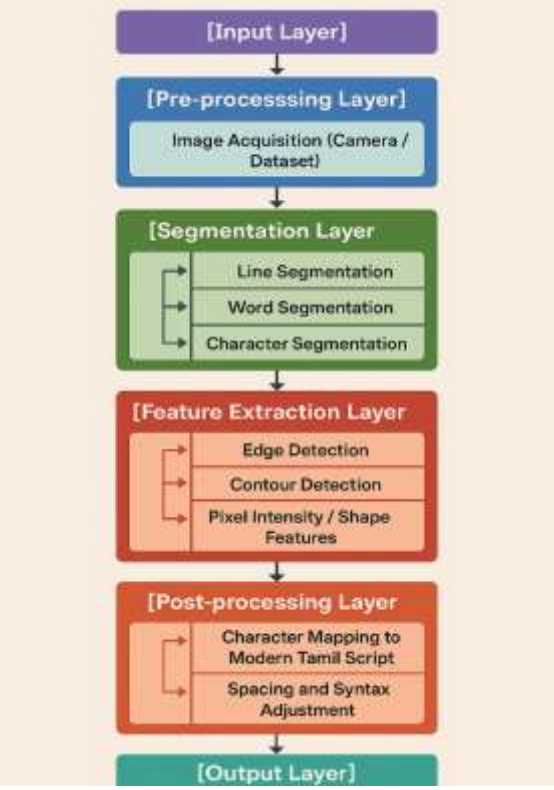
1. **Operating System: Windows 10 / Ubuntu 20.04**
2. **Python Version: 3.7 or Higher**
3. **Jupyter Notebook: Development and Execution Environment**

7.LIBRARIES AND TECHNOLOGIES USED IN THIS PROJECT:

Converting ancient Tamil inscriptions into modern readable text is a complex task that involves multiple stages: image preprocessing, character segmentation, feature extraction, and character classification. This requires a robust pipeline combining image processing and deep learning techniques. To build such a pipeline efficiently, powerful libraries like

1. TensorFlow/Keras
2. OpenCV
3. NumPy are essential.
4. Matplotlib (for Visualization)
5. Scikit-learn (for Evaluation and Clustering)
6. Pandas (for Data Manipulation)

8. ARCHITECTURE DIAGRAM:



9. ADVANTAGES:

- Preserving Ancient Tamil Heritage
- Automating the Translation Process
- Improving Accessibility to Historical Texts
- Enhancing Research in Linguistics and History
- Scaling and Expanding Tamil Script Recognition
- Enabling Real-Time Translation in Modern Applications
- Facilitating Multilingual Understanding

10. APPLICATION:

- Digital Preservation of Ancient Tamil Inscriptions
- Cultural Heritage Research and Documentation
- Automated Translation for Historical Texts
- Educational Tool for Linguistics and History Students
- Support for Archaeological and Historical Analysis

11. CONCLUSION:

This project demonstrates a significant advancement in the recognition and translation of ancient Tamil inscriptions into modern Tamil script using deep learning technologies. By combining image preprocessing, character segmentation, and Convolutional Neural Networks (CNNs), the system is able to accurately recognize and convert stone-engraved inscriptions into readable text, preserving the cultural heritage of Tamil civilization. The automation of this process eliminates the need for labor-intensive manual transcription, enhancing efficiency and scalability, especially for large archives of historical inscriptions.

The system's ability to handle complex, damaged, or stylized characters ensures high accuracy even with challenging ancient scripts. Furthermore, by employing multipart character recognition, the system can deal with the intricacies of ancient Tamil characters, offering a robust solution for digitalizing these inscriptions.

Ultimately, this project not only facilitates the preservation of historical Tamil texts but also makes them accessible for research, education, and broader public understanding. It paves the way for future advancements in the field of digital humanities and linguistic studies, enabling modern-day access to centuries-old Tamil heritage while supporting the ongoing efforts to maintain and revitalize ancient languages.

12. REFERENCES:

1. *"Analysis of Various Deep Learning Algorithms for Tamil Character"* Recognition

Authors: Ashok Kumar L, Shalini J, Karthika Renuka D

Published In: Proceedings of the First International Conference on Combinatorial and Optimization (ICCAP), December 2021.

2. "*Effective Tamil Character Recognition Using Supervised Machine Learning Algorithms"* Authors: Dr. S. Suriya, S. Nivetha, P. Pavithran, Ajay Venkat S., Sashwath K. G., Elakkiya G.

Published In: EAI Endorsed Transactions on e-Learning, February 2023.

3. *"A Novel Approach to OCR Using Image Recognition-Based Classification for Ancient Tamil Inscriptions in Temples*"

Authors: Lalitha Giridhar, Aishwarya Dharani, Velmathi Guruviah

Published In: arXiv preprint, July 2019.

4. *"Handwritten Tamil Character Recognition and Digitalization Using Deep Learning*"

Authors: N. Sasipriyaa, K. Abirami, G. Banupriya, S. Dhivya

Published In: International Journal of Advanced Science and Technology, March 2020.

5. *"Ancient Tamil Character Recognition Based on Edge Mapping Pointed Multi-Perspective Neural Network for Enhanced Font Definition"*

Author: P. Selvakumar

Published In: International Journal of Intelligent Systems and Applications in Engineering, March 2024.

6. *"Tamil Text Recognition Using Deep Learning Models: Challenges and Approaches"*

Authors: A. Subramanian, S. Ramesh, M. Venkatesh

Published In: Springer Nature, 2022.

7. *"OCR for Tamil Handwritten Characters Using Deep Convolutional Neural Networks*"

Authors: S. Kumaran, R. Sundararajan Published In: Procedia Computer Science, 2021.

8. *"Segmentation and Recognition of Tamil Characters Using Convolutional Neural Networks*"

Authors: J. Pradeep, M. Vijayakumar Published In: Journal of Computer Science and Technology, 2020.

9. *"Multi-Class Classification for Tamil Text Recognition Using Convolutional Neural Networks*"

Authors: P. Karthik, S. Kalaiselvi Published In: International Journal of Computer Science and Information Security, 2019.

10. *"Preservation and Digitization of Historical Inscriptions: A Tamil Case Study*"

Authors: R. Manoharan, K. Balamurugan

Published In: Journal of Digital Humanities, 2021.