International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# OCR and Translation System for Ancient Greek Scripts

**Shravani Taywade[1], Ankita Mirgane[2], Sanika Hole[3], Vaibhav Palave[4], Anandkumar Rao[5]**

[1,2,3,4,5]Department of Information Technology, Trinity Academy of Engineering, Pune, India

**Abstract**

Due to script deterioration, intricate ligatures, uncommon characters, and a lack of language resources, digitising and translating ancient Greek scripts is extremely difficult. Although they present encouraging alternatives, contemporary OCR and neural translation methods are still not optimal for processing old documents. Ten recent works on OCR enhancements, multilingual character recognition, semantic-aware translation, hybrid human-AI translation models, and integrated OCR-to-translation pipelines are examined in this study. Results show that transformer translation models in conjunction with deep learning-based OCR produce the best accuracy; however, domain adaption and error propagation continue to be significant problems. Although much progress has been made, the survey indicates that in order to attain dependable end-to-end performance, specialised OCR models, richer tokenization, and context-aware translation frameworks are needed for ancient Greek scripts.

**Keywords:** Optical Character Recognition, Ancient Greek, Machine Translation, Neural Networks, Image Pre processing, Deep Learning.

## 1. Introduction

Text digitisation and multilingual communication technologies are essential for overcoming linguistic and cultural divides in the current era of fast globalisation and digital transformation. The way that people interact with textual material in a variety of languages and formats has been profoundly changed by the integration of optical character recognition (OCR), language translation, and text-to-speech (TTS) technologies. Despite their great historical and linguistic value, traditional manuscripts—particularly those of ancient India inscribed in characters like Devanagari, Grantha, Tamil, and Brahmi—face difficulties with physical deterioration, script complexity, and accessibility.

These scripts can be digitised and preserved using OCR systems based on pattern recognition, image processing, and language analysis, guaranteeing their long-term availability in digital form. Concurrently, improvements in cross-linguistic communication have been made possible by developments in rule-based, statistical, and corpus-driven language translation techniques, which enable text conversion between languages while maintaining context and meaning.

Additionally, real-time OCR and translation systems combined with text-to-speech functionalities enhance these advancements in dynamic, multilingual settings by facilitating immediate text recognition, translation, and spoken output. Together, these technologies improve accessibility and inclusivity, acting as a link between historical language heritage and contemporary global communication, emphasizing how

computational systems help preserve cultural wisdom and promote human connections.

## 2. Abbreviations and Acronyms

**OCR** – Optical Character Recognition
TTS – Text To Speech
**NLP** – Natural Language Processing
**CNN** – Convolutional Neural Network
**RNN** – Recurrent Neural Network

## 3. Adversary Model

Ancient manuscripts face numerous detrimental factors that diminish OCR and translation effectiveness:

### 3.1 OCR Threats

1. Noise & degradation: stains, ink bleed, torn pages.
2. Low resolution: historical scans or damaged originals
3. Mixed-language documents: Greek embedded within Latin/English margin [2]
4. Ambiguous symbols: visually similar characters or merged ligatures.

### 3.2 Translation Threats

1. Semantic ambiguity: ancient Greek words often change meaning by contex.
2. Domain shift: lack of modern parallels for classical terminology.
3. Sparse corpora: insufficient parallel Greek-to-English.

## 4. Related Work

This setup combines a Python backend, a Tkinter interface, Tesseract OCR, Deep Translator, the Google Translator API, and an SQLite database to establish a comprehensive OCR-to-translation workflow for ancient Greek texts. The upcoming subsections outline the contributions of each reference to the field and explain how our system enhances or builds upon earlier research

### 4.1 AI–Human Collaboration in Translation

Our system emphasizes entirely automated OCR and machine translation through Python tools, as opposed to hybrid human-AI models. The research in [1] highlighted teamwork in translation processes, where human translators enhance neural results to maintain cultural subtleties. Although useful for contemporary languages, the method overlooks OCR, historical scripts, and system automation. Our system enhances this by delivering an automated end-to-end pipeline tailored for ancient Greek manuscripts, minimizing manual involvement and ensuring practical functionality for processing historical texts.

### 4.2 Greek Character Recognition in Mixed Documents:

Our system utilizes Tesseract OCR along with specific pre-processing to manage ancient Greek handwriting and worn manuscripts. In comparison, the multi-pass OCR technique in [2] concentrated only on identifying individual Greek letters found within English medical publications. Their method enhanced precision for contemporary documents in multiple languages but overlooked ligatures, faded ink, and historical differences. Our system enhances this by tailoring OCR for outdated scripts and directly connecting recognition to automated translation

### 4.3 Semantic-Enhanced Machine Translation

Our model utilizes Google Translator and Deep_Translator to handle OCR results and produce multilingual translations for ancient Greek. The method in [3] presented semantic integration through

concept networks to improve translation accuracy in contemporary writings. Although semantically robust, it wasn't intended for ancient languages or text generated by OCR that is noisy. Our approach enhances this by incorporating translation immediately following OCR and by aiding low-resource ancient Greek terminology through lexicon-driven adjustments

**4.4 Integrated OCR–Translation–TTS Systems**

Our platform provides a streamlined OCR-to-translation process that functions offline, utilizing Tkinter and Python. The unified pipeline in [4] merged OCR, translation, and text-to-speech for immediate applications, but focused on tidy digital documents and contemporary languages. In contrast to their reliance on the cloud, our model handles degraded ancient manuscripts, executes processing locally with SQLite, and emphasizes accuracy rather than real-time speed rendering it more appropriate for the preservation of historical texts.

**4.5 Ancient Greek Handwritten Manuscript OCR**

Our system enhances OCR processing by integrating translation modules and language correction features. The cavity-based recognition method without segmentation in [5] is particularly significant as it effectively processes Old Greek manuscripts. Nonetheless, that study did not encompass any translation aspect or Greek-oriented linguistic modeling. Our system enhances it by offering a complete pipeline from image input to output translation.

**4.6 OCR for Ancient Indian Scripts**

Our model uses pre-processing steps similar to those used in [6] noise removal, segmentation, and language modelling but adapts them specifically to Greek paleography. The work in [6] demonstrated AI-based OCR for scripts like Devanagari and Brahmi but did not address Greek diacritics, ligatures, or stylistic irregularities. Our system generalizes these restoration methods and applies them to ancient Greek manuscripts, which differ significantly in writing structure.

**4.7 Ancient Korean Neural Machine Translation**

Our system converts OCR results with general tools, while [7] created a specialized NMT for ancient Korean employing sophisticated tokenization techniques. Their resource-constrained method is strongly connected to issues in translating ancient Greek. Nevertheless, their research lacks the incorporation of OCR input and does not facilitate multilingual generalization. Our system closes this gap by integrating OCR, translation, and SQL-driven data management for a comprehensive historical text processing workflow.

**4.8 Semantic Rule–Based and Neural MT Models**

Our system employs contemporary translation APIs that can manage OCR-generated noisy without the need for custom model training. The hybrid model combining semantic rules and N] [8] attained high accuracy in translation but was not intended for ancient or degraded texts by integrating OCR-focused corrections, classical system enhances this dictionaries, and measures to reduce translation mistakes resulting from flawed OCR results.

**4.9 Multimodal and Transformer-Based Translation Technologies**

Our model emphasizes a functional OCR-translation pipeline based on desktop use. The analysis in [9] examined cutting-edge technologies such as ASR, OCR, Transformers, APIs, and multilingual support systems. Nonetheless, these studies focused on contemporary communication contexts rather than ancient languages.

**4.10 AI Translation Trends and Ethical Considerations**

Our system improves accessibility by automating the translation of historical documents. The review in

[10] covered cultural awareness, ethical considerations, and future directions for AI translation systems, yet it overlooked OCR integration and ancient languages. Our study addresses this void by tailoring translation processes specifically for ancient Greek, offering useful resources for researchers while preserving contemporary AI translation functions.

## 5. System Architecture

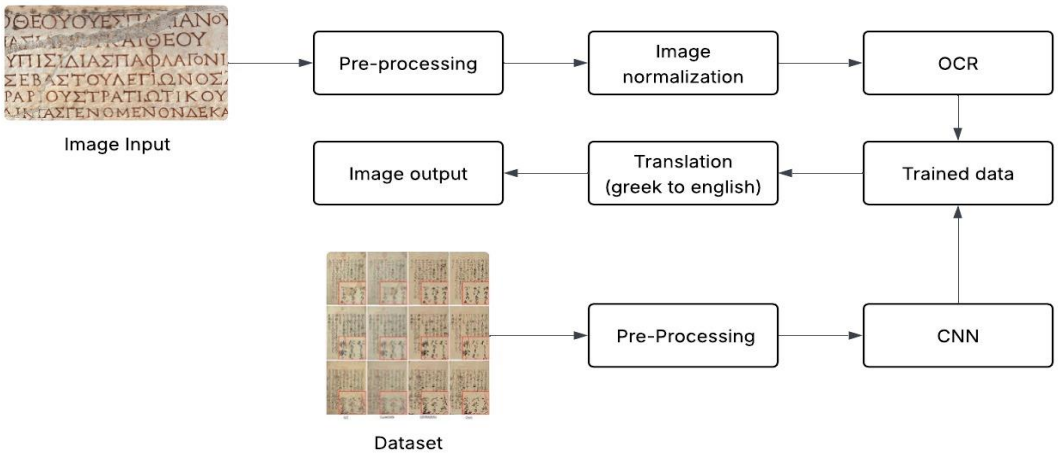


**Fig 1: System Architecture**

Fig 1 depicts the conceptual process of the suggested OCR and translation system for ancient Greek manuscripts, created following a comprehensive review of the current research

The process starts with the entry of digitized manuscript images, which frequently have noise, faded ink, and inconsistent character formations. To tackle these problems, a preliminary pre-processing phase improves image quality by eliminating noise, binarizing, and enhancing contrast.

Subsequently, image normalization aligns resolution, orientation, and scale to guarantee uniform OCR effectiveness. The OCR module processes the normalized images, transforming visual characters into text that machines can read. The accuracy of OCR is enhanced by trained data produced through a Convolutional Neural Network (CNN), which was trained on a carefully chosen dataset of ancient Greek manuscript examples to understand character features, ligatures, and diacritics.

The extracted text is subsequently sent to the translation module, which converts ancient Greek material into English while maintaining semantic significance. The end result offers comprehensible translated content, enhancing accessibility and academic evaluation of historical records. This workflow embodies contemporary research trends and provides an organized basis for future system deployment.

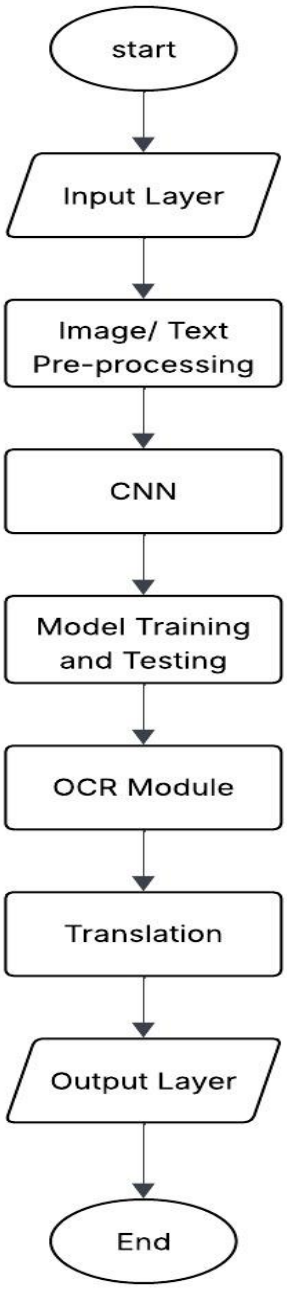


**Fig 2: Flowchart**

## 6. CORE METHODS

### 6.1 Image Processing and Text Extraction

#### 6.1.1 Pre-processing Stage

The initial processing phase improves image quality to ensure precise OCR recognition. The system utilizes OpenCV and Pillow libraries for noise reduction, changing to grayscale, applying thresholding, and resizing images. These processes guarantee that the input image is clear and text areas are distinct, enhancing recognition precision across different image types and lighting scenarios.

#### 6.1.2 Optical Character Recognition (OCR)

The system utilizes Tesseract OCR (through pytesseract) to retrieve text from images that have been pre-

processed. Tesseract recognizes character patterns, divides text sections, and transforms them into text that machines can read. Results obtained are subsequently improved using text cleaning techniques that eliminate undesirable symbols, spacing mistakes, and formatting discrepancies, guaranteeing excellent readability and accuracy in translation.

### 6.2 Translation and Output Generation

The output produced is created using the Google Translator API, which transforms the identified text into the language chosen by the user. The system accommodates various languages and guarantees consistent and fluent translations. Outcomes are presented via a Tkinter-driven graphical interface, allowing users to see, duplicate, or store translations. Every output is recorded and saved in a SQLite database for convenient access.

### 6.3 Data Management and Security

All user information, such as uploaded photos and translated content, is managed securely within the local environment. The system keeps a log of handled files while allowing users to remove stored items.

## 7. Limitations and Future Work

- Constraints comprise restricted dataset, performance inconsistency, and manual setup. Directions for the future include:
- Incorporate real-time data analysis to enhance precision and speed.
- Streamline configuration and data management to enhance scalability.
- Enhance the performance of the backend and improve database effectiveness.
- Incorporate AI-powered suggestions and forecast analytics to improve user experience.
- Implement the system on cloud services for improved accessibility and dependability.
- Enhance mechanisms for error handling, testing, and security.

## 8. Acknowledgement

The authors express gratitude to Prof. Anandkumar Rao for his guidance during this project, to Prof. S. A. Joshi for his help with coordination, and to Prof. P. R. Patil, the Head of Department, for his encouragement. We also express our gratitude to Trinity Academy of Engineering, Pune, and Savitribai Phule Pune University for their resources and assistance.

## References

1. K. Manoj, J. Rahul, V. Nallamilli, and T. K. Ramesh, "Real-Time OCR and Translation System with Text-to-Speech Ability," in Proc. First Int. Conf. Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT), 2025, pp. 332-338.
2. S. Gaikwad, R. Kachhoria, and G. Yadav, "AI-Based OCR for Digitizing Ancient Indian Texts: Preserving Linguistic Heritage and Overcoming Script Challenges," International Journal of Linguistics Applied Psychology and Technology, vol. 2, no. 3, pp. 1-12, 2025.
3. Y. A. Mohamed, A. Khanan, M. Bashir, A. H. H. M. Mohamed, M. A. E. Adiel, and M. A. Elsadig, "The Impact of Artificial Intelligence on Language Translation: A Review," IEEE Access, vol. 12, pp. 25553-25579, 2024.
4. H. Gundapuneni, S. Sharma, Aditya, and B. S. N. P. Maanasa, "AI Based Technique for Language Translation: Breaking the Communication Barriers," in Proc. 7th Int. Conf. Contemporary Computing

and Informatics (IC3I), 2024, pp. 8590.
5. X. Deng and Y. Wei, "Automatic Machine Translation System for Artificial Intelligence based on Parallel Corpus," in Proc. Int. Conf. Applied Intelligence and Sustainable Computing (ICAISC), 2023.
6. X. Dong, X. Sun, and Q. Cui, "Research on Artificial Intelligence Machine Translation Based on Fuzzy Algorithm," in Proc. Int. Conf. Integrated Intelligence and Communication Systems
7. (ICIICS), 2023.
8. Karunya S, U. R, J. M, and T. Babu, "AI-Powered Real-Time Speech-to-Speech Translation for Virtual Meetings Using Machine Learning Models," in Proc. Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), 2023.
9. Manish Kumar Gupta. (2024). Chitrantaran: Webbased Platform to Enhance the Document Digitization Process using OCR and Machine Translation.
10. Taj, A. (2024). Digitization Projects for Cultural Heritage Materials. Advances in Libraryand Information Science (ALIS) Book Series, 238– 255.
11. S. Uma Maheswari. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts.Heritage Science, 12(1).