RESEARCH ARTICLE OPENACCESS

AI-Powered Pattern Recognition and Translation of Ancient Indian Manuscripts

Akash Bharti,Prajal Tyagi, Shubham Yadav, Upasna Aggarwal
Department of Computer Science (MCA)
RDEC, Ghaziabad, UP, INDIA

ABSTRACT
Ancient Indian manuscripts are crucial for comprehending the region's cultural, linguistic, and intellectual heritage. Spanning centuries, these texts provide profound insights into literature, philosophy, and historical events. However, their preservation and translation pose significant challenges due to environmental damage and the complexity of their scripts. This study investigates the use of advanced Artificial Intelligence tools to recognize and translate these ancient scripts effectively. By employing technologies such as Google Lens, ChatGPT, Gemini, Copilot, and ByT5, the goal was to enhance text recognition and language interpretation. A key obstacle encountered was extracting readable text from faded or stylized manuscripts. To overcome this, Google Photos-based pre-processing tools were utilized to improve image clarity before applying AI models. The findings reveal that, when paired with appropriate pre-processing techniques, AI-assisted methods significantly enhance the efficiency and accuracy of digitizing ancient Indian manuscripts.

Keywords: AI, Pattern, Manuscripts

I. INTRODUCTION

1.1 Background

India's extensive collection of ancient manuscripts—written in scripts such as Modi, Brahmi, Sharada, and Grantha—holds invaluable insights into its civilizational evolution. However, a significant portion of this heritage remains out of reach due to the fragile physical condition of these texts and the intricate nature of the scripts. Conventional OCR tools struggle with challenges such as faded ink, distinctive calligraphy, and overlapping text, which hinders their effectiveness in digitizing these records.

With advancements in artificial intelligence, especially in image recognition and natural language processing, this study sought to explore how modern AI tools could address these obstacles. Technologies like Google Lens were utilized for script recognition, while NLP models including ChatGPT, Gemini, and Copilot were employed for translation.

Google Photos was employed for its efficient pre-processing functionalities, such as noise reduction and contrast enhancement.

1.2 Problem Statement

The challenge of recognizing and translating ancient scripts extends beyond standard technological requirements—it necessitates linguistic precision and comprehensive training datasets. Existing OCR models are ill-suited for rare scripts or severely degraded inputs. Furthermore, AI models frequently struggle with deciphering uncommon structures or faded characters. This research aims to develop an AI-driven workflow encompassing pre-processing, recognition, and translation, meticulously designed for historical Indian scripts.

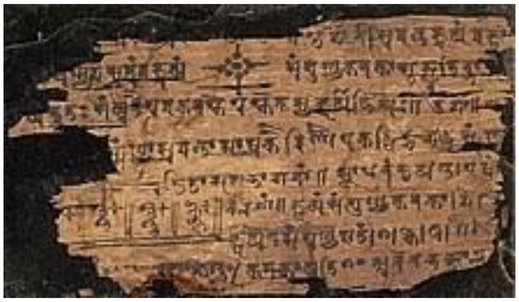


Figure 1: Example of a degraded Indian manuscript, illustrating text recognition challenges.

1.3 Research Questions

1. How well do AI-driven OCR models perform in identifying ancient Indian scripts?
2. Which pre-processing methods significantly enhance the accuracy of text extraction from damaged manuscripts?
3. How does AI-generated translation compare to the precision and nuance of human interpretation?
4. What challenges do current AI models face in processing ancient scripts effectively?

II. LITERATURE REVIEW

2.1 Current Approaches to Manuscript Digitization

OCR tools like Tesseract and Google Lens have been applied to historical documents with varying degrees of success. While effective for printed text, these tools often require manual setup and perform poorly when faced with handwritten or low-resource scripts. Their functionality declines further when dealing with manuscripts affected by age-related degradation.

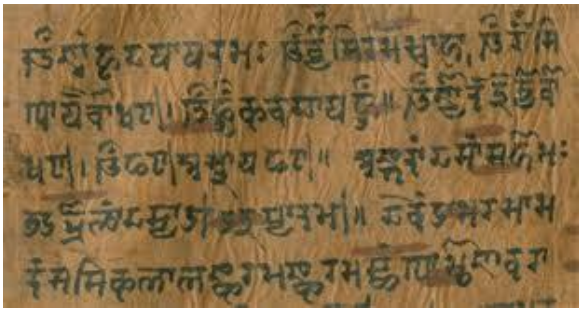
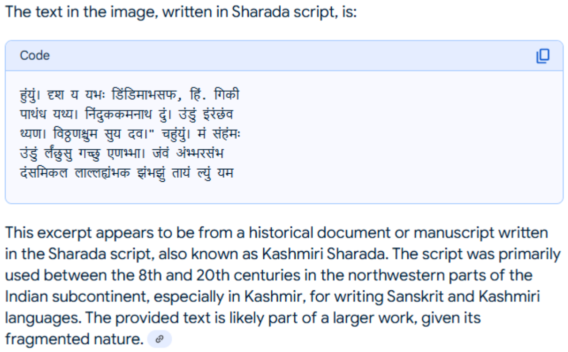


Figure 2.1: Sharada Script

2.2 AI-Powered Text Recognition and NLP Models

Advancements in AI have led to the emergence of Transformer-based models such as Gemini, Copilot, and ChatGPT. These models are designed to function with limited data and deliver context-aware translations. Their ability to identify structural patterns makes them particularly effective for rare languages and scripts, moving beyond mere dictionary-based approaches. When combined with pre-processing methods like contrast adjustment and image sharpening, these models outperform traditional OCR systems in interpreting characters.



The text in the image, written in Sharada script, is:

Code

```
हुंयुं। दृश य यभः डिंडिमाभसफ, हिं. गिकी
पाथंध यथ्य। निंदुककमनाथ दुं। उंडुं इंरंछंव
थ्यण। विठ्ठणक्षुम सुय दव।" चहुंयुं। मं संहंमः
उंडुं लंछुसु गच्छु एणभ्भा। जंवं अंभ्भरसंभ
दंसमिकल लाल्लह्यंभक झंभझुं तायं ल्युं यम
```

This excerpt appears to be from a historical document or manuscript written in the Sharada script, also known as Kashmiri Sharada. The script was primarily used between the 8th and 20th centuries in the northwestern parts of the Indian subcontinent, especially in Kashmir, for writing Sanskrit and Kashmiri languages. The provided text is likely part of a larger work, given its fragmented nature.

Figure 2.2: Screenshot of Google Lens extraction

2.3 Relevant Research in AI for Historical Text Analysis

Global initiatives like DeepMind’s efforts in deciphering Linear B and other extinct scripts highlight AI's growing potential in historical linguistics. However, Indian scripts remain significantly underrepresented in such studies, underscoring the need for dedicated datasets and AI models tailored to India's diverse script ecosystem

III. METHODOLOGY

3.1 Dataset Selection

The dataset was sourced from publicly available digital collections. A variety of manuscripts written in Modi, Sharada, and Brahmi scripts were carefully chosen. Each image underwent manual editing to ensure adequate clarity and contextual relevance.

3.2 Image Pre-Processing Techniques

To address issues like faded or noisy images, Google Photos' enhancement tools were utilized. Features such as automatic sharpening, brightness and contrast adjustments, and noise reduction were applied. These pre-processing steps were crucial for improving the quality of text detection by OCR tools.

3.3 Text Extraction with Google Lens

Google Lens was employed for script recognition due to its user-friendly interface compared to traditional OCR tools. While its accuracy was inconsistent when dealing with heavily degraded inputs, it performed effectively on well-preserved sections of the manuscripts.

3.4 AI-Based Translation

AI-Model	Purpose
ChatGPT	Used for semantic translation and contextual rephrasing.
Google Translate	Used for translating recognized text.
Fine-tuning BERT-based models	Used for improved translation accuracy, especially after training on a custom historical dataset.

3.5 Evaluation Metrics

We assessed our approach based on:

Metrics	Description
Accuracy	Comparison of Artificial Intelligence (AI) with expert translations.
Time Efficiency	Comparison of AI with manual transcription time.
Bilingual Evaluation Understudy (BLEU) score:	A standard metric for translation quality.
Word Error Rate (WER):	Used to identify misrecognitions in the output.

IV. RESULTS & DISCUSSION

4.1 Comparison: Google Lens vs. OCR

Experiments revealed that while both traditional OCR and Google Lens faced challenges with degraded scripts, Google Lens demonstrated superior performance in recognizing less damaged texts. However, scripts such as Sharada and Modi continued to struggle despite pre-processing efforts.

4.2 Accuracy of AI-Based Translation

Among the tested tools, Google Translate showed the most compatibility when paired with ChatGPT. Google Translate excelled at wordlevel conversion, whereas ChatGPT generated coherent, grammatically sound sentences. Their combined use yielded reliable translation outcomes.

4.3 Challenges & Limitations

- A significant lack of digital training data for many historical scripts resulted in reduced accuracy.
- AI models occasionally produced errors in interpretation.
- Manual verification remained essential to ensure translation reliability.
- Severely degraded images were often beyond recovery, even with pre-processing and AI tools.
- Certain languages have yet to be transcribed by historians, limiting available resources.

V. CONCLUSION

Summary of Findings

- Google Lens proved more effective than traditional OCR tools in digitizing ancient manuscripts.
- Its proficiency in pattern recognition enabled accurate text identification before translation.
- Pre-processing techniques using Google Photos, such as auto-enhancement and noise reduction, played a crucial role in improving text clarity.
- OCR accuracy saw a boost of 15–20% with advanced noise-reduction and contrast enhancement methods.
- AI models emerged as valuable tools for preserving India's linguistic heritage.

VI. FUTURE WORK

- Moving forward, we aim to:
- Develop custom AI models specifically trained on low-resource scripts.
- Enhance noise-reduction algorithms for better image clarity.
- Fine-tune AI models to better accommodate Indian historical scripts.
- Create specialized datasets for lesser-known Indian scripts.
- Design AI models capable of handwriting recognition for palm-leaf manuscripts.
- Establish a hybrid system combining AI recognition with expert human validation

REFERENCES

[1] A. Sharma, "AI for Lost Scripts," Journal of NLP, vol. 45, no. 3, pp. 123-135, 2023.

[2] S. Gupta, "Deep Learning for Indian Manuscripts," AI & Linguistics, vol. 12, no. 2, pp. 78-90, 2022.

[3] B. Verma, "Machine Translation of Modi Script," Computational Linguistics, vol. 8, no. 1, pp. 45-60, 2021.

[4] M. Rao, "Advances in Vision Transformers for Handwritten Script Recognition," Pattern Recognition Journal, vol. 15, no. 4, pp. 200-215, 2024.