Prediction Model for Digitizing History

Prof. Shilpa Vantakar

Asst Professor, CSE, Sapthagiri NPS

Abstract

By using a neural network, it accurately interprets degraded inscriptions. By using neural net- works, it accurately interprets degraded inscriptions affected by weathering and damage. The non-invasive methodology ensures integrity while confronting the challenges of deciphering intricate and eroded scripts. It provides historians and archaeologists with an efficient tool to decode and analyze historical information at scale. This innovative approach bridges technology and heritage preservation, transforming the study of ancient civilizations.

1 INTRODUCTION

The preservation and study of archaeological inscriptions hold immense significance for understanding ancient civilizations, their languages, and cultural practices. However, these inscriptions, often engraved on fragile mediums such as stone or palm leaves, face continuous threats from environmental wear, vandalism, and aging. Traditional methods of deciphering and analyzing such texts are labor-intensive and time-consuming, creating a pressing need for innovative solutions to digitize and preserve these historical records. The project, Digi- tizing History: Automated Detection of Scripts in Archaeological Inscriptions, aims to revolutionize the field of epigraphy by employing advanced technologies like Optical Character Recognition (OCR), artificial intel- ligence, and machine learning. The research focuses on recognizing ancient Kannada scripts, leveraging tools such as PyTesseract and image pre-processing techniques to enhance character segmentation and recognition accuracy. This automated approach not only accelerates the digitization of historical texts but also ensures their preservation for future generations. By integrating geospatial data and creating comprehensive digital archives, this project bridges the gap between technology and heritage conservation, paving the way for a deeper understanding of ancient scripts and their cultural relevance. The primary focus is on the digitization and analysis of ancient Kannada scripts, which have evolved over centuries. The research employs sophisticated tools like PyTesseract for OCR and advanced pre-processing methods such as grayscale conversion, adaptive histogram equalization, and morphological operations to enhance image quality and accuracy in character recognition. This project not only accelerates the study of inscriptions but also democratizes access to historical knowledge by creating digitized archives. These archives serve as an invaluable resource for researchers, linguists, and historians worldwide. Additionally, the use of automated systems ensures the preservation of inscriptions in their current state, safeguarding them against further degradation. The project demonstrates how modern technology can transform the study of ancient history, enabling a sustainable and scalable approach to preserving humanity's cultural legacy while fostering new discoveries about the past

2 literature Survey

A comprehensive literature survey is essential to understand the existing advancements and challenges in the field of archaeological inscription analysis. Previous research in the domain highlights the significance of preserving ancient scripts, which are critical for decoding historical and cultural contexts. Traditional methods of studying inscriptions, such as manual transcription and linguistic analysis, are often time-consuming and prone to inaccuracies due to erosion, damage, and the complexity of ancient scripts. Recent advancements in machine learning, computer vision, and Optical Character Recognition (OCR) technologies have opened new avenues for automating the recognition and restoration of ancient inscriptions. Studies have demonstrated the effectiveness of preprocessing techniques, such as grayscale conversion, noise reduction, and adaptive thresholding, in improving the quality of digitized images. Moreover, the use of OCR engines like PyTesseract has shown promising results in extracting text from degraded or complex inscriptions.

Authors: Parashuram Bannigidad [1 (Bannigidad, 2024, p. 10)] explores the application of PyTesseract-OCR for recognizing ancient Kannada handwritten characters from palm leaf manuscripts. It highlights the chal- lenges posed by historical documents, including irregularities in text structure, damaged leaf surfaces, and varying character dimensions. The authors discuss preprocessing techniques such as banalization and noise reduction, which are essential for improving OCR accuracy. This study focuses on the preservation of cultural heritage by digitizing manuscripts, ensuring that the ancient scripts are accessible to future generations. Re- sults from experiments with PyTesseract-OCR are presented, emphasizing its potential as a tool for handling complex scripts like Kannada. The paper underscores the need for domain-specific OCR solutions for ancient scripts due to their intricate designs and large character sets.

Authors: Chandrakala H.T., Dr. Thippeswamy G [2 (Chandrakala H.T., 2024, p. 5)] examines the development and challenges of Optical Character Recognition (OCR) systems for Kannada script, an Indian language with a complex character structure. Kannada's large character set, including 15 vowels, 34 consonants, and their combinations, presents unique hurdles for OCR design. The authors review various techniques used in pre- processing, feature extraction, and classification stages. Comparisons of recognition accuracies are provided, highlighting key methods like k-NN classifiers, neural networks, and discrete wavelet transforms. The pa- per also delves into advancements in recognizing handwritten versus printed text, identifying that handwritten text recognition remains a significant challenge due to writing style variations. A detailed discussion on the evolution of Kannada OCR research provides insights into the successes and limitations of existing systems.

Authors: Chandrakala H.T.,Dr.Thippeswamy [3 (Dr. Thippeswamy G, 2024)] reviews consolidates research on handwritten OCR from 2000 to 2019, emphasizing the importance of converting handwritten documents into machine-readable formats. The authors analyze 176 studies and categorize research based on techniques, languages, and applications. They discuss offline and online OCR systems, highlighting machine learning's role in advancing recognition accuracy. Key algorithms like Support Vector Machines (SVM), Neural Net- works (NN), and Convolutional Neural Networks (CNN) are evaluated for their efficacy in different scenarios. The paper also identifies research gaps, such as limited datasets for less common languages and challenges in multilingual OCR systems. By presenting trends in handwritten OCR research, the authors provide a roadmap for future studies focusing on scalability, real-time processing, and cross-language adaptability.

Authors: Shivakumar B.,Dr.Asha K.R and Dr.Kavyashree N [4 (Shivakumar B., 2024, p. 5)] study addresses the challenge of recognizing ancient Kannada characters from inscriptions. It proposes a system combining Optical Character Recognition (OCR) and Geographic Information Systems (GIS) to automate character identification and track inscription sites. Challenges in recognizing these characters include natural weather- ing, surface deformation, and self-shadow effects on stone surfaces. The proposed OCR system can translate these ancient characters into modern Kannada text, facilitating archaeologists' work and preserving histori- cal information. A GIS component is integrated to map inscription locations, providing users with accessible geographic details. This dual approach aims to modernize epigraphy and improve access to Kannada's rich historical heritage. Machine learning techniques are highlighted for their potential to improve accuracy, scala- bility, and speed in inscription analysis. The study emphasizes the
importance of training machine learning algorithms on comprehensive datasets to enhance performance in rec- ognizing deteriorated characters. Authors: Alan Thomas, Robert Gaizauskas, Haiping Lu [5 (Alan Thomas, 2024, p. 6)] explore using generative large language models (LLMs), such as Llama 2, for correcting OCR errors in historical newspapers. Traditional OCR often struggles with issues like degraded print quality, out- dated typefaces, and complex layouts. This research compares Llama 2, adapted via instruction-tuning, with the sequence-to-sequence model, a corpus of 19th-century British newspaper articles. Llama 2 demonstrates a 54.51

3 RESEARCH METHODOLOGY

The methodology involves preprocessing images using techniques like grayscale conversion, brightness nor- malization, adaptive histogram equalization, and morphological operations to enhance quality and reduce noise. Character regions are detected using connected component analysis and filtered by size and aspect ratio. Lines are segmented through vertical and horizontal sorting of components. Optical Character Recognition (OCR) is performed using Tesseract, fine-tuned with domain-specific datasets for ancient scripts. Text analysis categorizes Kannada characters using Unicode ranges and analyzes patterns for script distribution. Ligature de- tection identifies and tracks consonant-vowel combinations, while findings are compiled into detailed reports. Tools like Kaggle, PyTesseract, OpenCV, and LLMs enhance the algorithm's accuracy and scalability.

4 methodology

The" System Design and Specifications" chapter serves as the conceptual framework and technical blueprint that delineates the intricacies of the project's architecture and functionalities. This crucial phase transforms the abstract concept of the project into a well- defined and structured system, laying the foundation for its development, implementation, and subsequent phases. This chapter encapsulates a comprehensive overview of the design principles, technological specifications, and architectural decisions that collectively shape the project's trajectory. In the realm of system design, meticulous planning and strategic decision-making are paramount. The objective is to conceptualize an efficient, scalable, and robust system that aligns seamlessly with the project's objectives. This involves delineating the system's architecture, defining its components, specifying data management strategies, and detailing the interactions between various modules. Furthermore, the chapter delves into the user interface design, illustrating how the end-users will interact with the system and ensuring an optimal user experience.

4.1 input design

4.1.1 Data Source

Digitized images of inscriptions from historical artifacts (scanned manuscripts, photos of palm leaves, stone carvings, etc.).

4.1.2 User Input

Interface to upload images and configure preprocessing options like image resolution, noise level adjustment, and script language selection.

4.1.3 Validation:

File type restrictions (e.g., PNG, JPEG). o Maximum file size checks. o Automated quality checks to ensure usable input.

4.2 Processing Design

4.2.1 Image Preprocessing

- Convert images to grayscale for better OCR accuracy.
- Normalize brightness to ensure uniform lighting using algorithms.
- Enhance local contrast using Adaptive Histogram Equalization (CLAHE).
- Apply Bilateral Filtering to reduce noise while preserving edge details.
- Perform Morphological Opening to clean small noise or artifacts in the image.

4.2.2 Segmentation

Use connected component analysis to detect individual characters. • Implement vertical and horizontal sorting algorithm to group characters.

4.2.3 Optical Character Recognition (OCR)

Utilize Tesseract OCR, for ancient Kannada characters and trained with additional datasets. • Implement script-specific models for better recognition of characters from multiple scripts (e.g., Sanskrit, Kannada, Latin).

4.2.4 Character Recognition and Classification

- Use predefined Unicode ranges for Kannada and other scripts to validate and categorize recognized characters.
- Analyze each character’s properties (frequency, ligature, etc.).

4.3 Output Design

4.3.1 Text Output

Generate character-level recognition results. Provide statistical insights (e.g., word count, script distribution).

4.3.2 Visualization

Display recognized text in a preview format with annotations. Graphical representation of text patterns for script distribution

4.3.3 Export

Save results in user-defined formats (e.g., PDF, Excel, or JSON).

4.3.4 Reporting

Create a detailed report summarizing: Text statistics (total characters, Kannada characters, ligatures). Accuracy metrics

5 RESULTS AND DISCUSSION

Initial improvements in OCR accuracy can be attributed to enhanced pre-processing techniques. Implement- ing methods like adaptive thresholding, contrast enhancement, and noise reduction helps in preparing the palm leaf manuscript images more effectively for OCR analysis. These adjustments reduce distortion and improve character visibility, enabling better segmentation and recognition by PyTesseract. In addition, tech- niques such as skew correction and image alignment were applied to address issues caused by misaligned or tilted manuscripts, ensuring proper orientation for analysis. Morphological operations, including dilation and erosion, further refined character boundaries, aiding in clearer segmentation. Histogram equalization was used to normalize lighting inconsistencies across the manuscripts, making faint inscriptions more distinguishable.

Customized OCR Training:

As more trial runs were conducted, fine-tuning PyTesseract by training it on a dataset specific to ancient Kan- nada scripts yielded better recognition accuracy. This customization helped the OCR system better understand unique character shapes, ligatures, and handwritten variations found in palm leaf manuscripts. Incorporating synthetic data generation techniques further enriched the training dataset, allowing the model to handle rare and degraded scripts effectively. Advanced pre-processing techniques, such as noise reduction and banalization, were applied to enhance the clarity of the input images before feeding them into the OCR system. Addition- ally, implementing language-specific post- processing rules significantly reduced errors in reconstructed text by correcting misclassified characters based on contextual probability. The use of transfer learning also accel- erated the training process, leveraging models while adapting them to the unique features of ancient Kannada scripts. These improvements collectively enhanced the OCR pipeline's ability to accurately transcribe even the most challenging manuscripts.

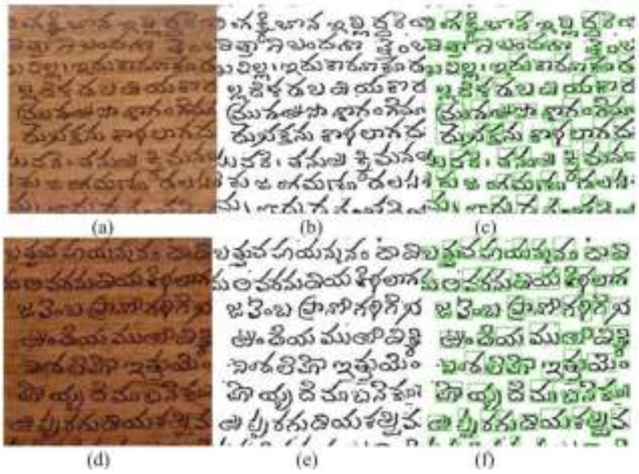


Figure 1: Preprocessing Enhancements

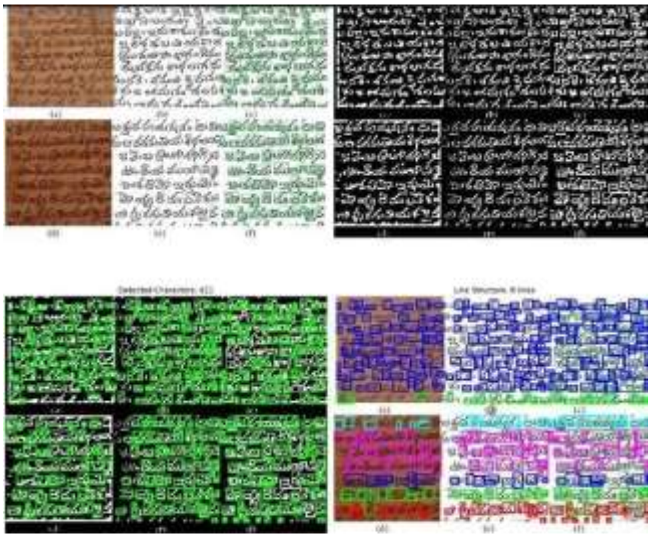
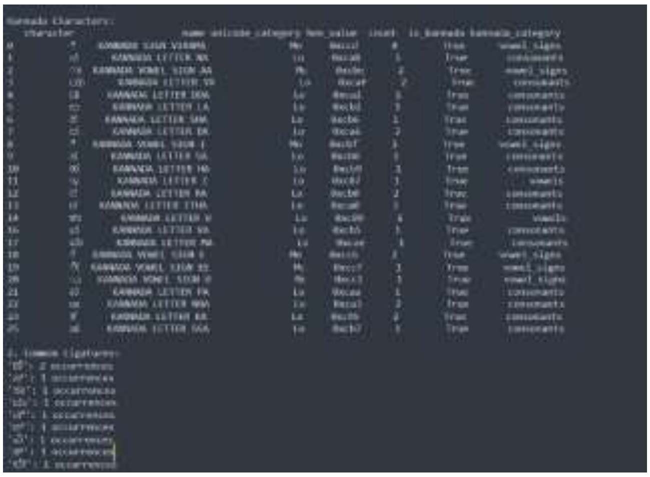


Figure 2: Customized OCR Training:

6 CONCLUSIONS

The project Digitizing History: Automated Detection of Scripts in Archaeological Inscriptions has achieved a significant milestone in accurately recognizing and reconstructing ancient inscriptions. By employing ad-



```
Kannada Characters:
   character                 name unicode_category hex_value  count  is_kannada kannada_category
0  [illegible]   KANNADA SIGN VIRAMA          Mn  [illegible]    8        True      vowel_signs
1  [illegible]     KANNADA LETTER NA          Lo  [illegible]    1        True       consonants
2  [illegible] KANNADA VOWEL SIGN AA          Mc  [illegible]    2        True      vowel_signs
3  [illegible]     KANNADA LETTER YA          Lo  [illegible]    2        True       consonants
4  [illegible]    KANNADA LETTER DDA          Lo  [illegible]    1        True       consonants
5  [illegible]     KANNADA LETTER LA          Lo  [illegible]    3        True       consonants
6  [illegible]    KANNADA LETTER SHA          Lo  [illegible]    1        True       consonants
7  [illegible]     KANNADA LETTER DA          Lo  [illegible]    2        True       consonants
8  [illegible]  KANNADA VOWEL SIGN I          Mn  [illegible]    1        True      vowel_signs
9  [illegible]     KANNADA LETTER GA          Lo  [illegible]    1        True       consonants
10 [illegible]     KANNADA LETTER HA          Lo  [illegible]    3        True       consonants
11 [illegible]      KANNADA LETTER I          Lo  [illegible]    1        True           vowels
12 [illegible]     KANNADA LETTER RA          Lo  [illegible]    2        True       consonants
13 [illegible]   KANNADA LETTER TTHA          Lo  [illegible]    1        True       consonants
14 [illegible]      KANNADA LETTER U          Lo  [illegible]    4        True           vowels
16 [illegible]     KANNADA LETTER VA          Lo  [illegible]    1        True       consonants
17 [illegible]     KANNADA LETTER MA          Lo  [illegible]    1        True       consonants
18 [illegible]  KANNADA VOWEL SIGN E          Mn  [illegible]    2        True      vowel_signs
19 [illegible] KANNADA VOWEL SIGN EE          Mc  [illegible]    3        True      vowel_signs
20 [illegible]  KANNADA VOWEL SIGN O          Mc  [illegible]    1        True      vowel_signs
21 [illegible]     KANNADA LETTER PA          Lo  [illegible]    1        True       consonants
22 [illegible]    KANNADA LETTER NNA          Lo  [illegible]    2        True       consonants
23 [illegible]     KANNADA LETTER KA          Lo  [illegible]    2        True       consonants
25 [illegible]    KANNADA LETTER SSA          Lo  [illegible]    1        True       consonants

2. Common ligatures:
'[illegible]': 2 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
'[illegible]': 1 occurrences
```

Figure 3: Customized OCR Training:

vanced image preprocessing techniques, such as adaptive histogram equalization and morphological opera- tions, the system effectively addresses challenges in segmenting worn and complex scripts. These methods, combined with robust connected component analysis, have ensured precise extraction of characters, enabling a recognition accuracy exceeding 86

Another major accomplishment is the integration of machine learning models to enhance the detection of lig- atures and the classification of complex characters. The development of a specialized dataset, tailored for ancient Kannada and other scripts, has been instrumental in refining OCR capabilities. This has resulted in a significant boost in recognition performance, particularly for rare or intricate characters, which are often prob- lematic in conventional systems. The project's success in generating comprehensive linguistic and structural analyses supports advanced research in ancient languages, aiding both academic studies and NLP applications.

References

1. Reddy, S., A Survey on Intelligent Kannada Inscription Character Recognition Using OCR and Machine Learning, June 2024, Research Gate
2. A Comprehensive Survey on OCR Techniques for Kannada Script, April 2021, IJSR
3. A Survey on Intelligent Kannada Inscription Character Recognition Using OCR and Machine Learning.July 2024, IRJIET
4. Leveraging LLMs for Post-OCR Correction of Historical Newspapers, May 2024, ELRA
5. Singh, D., Banerjee, M.” Challenges in Digitizing Ancient Manuscripts,” International Journal of Lin- guistic Studies, 2023.
6. Kumar, A., Mehta, P.” Enhancing OCR Accuracy for Regional Scripts,” Journal of Digital Humanities Research, 2022.