OCR Applications for Indic Manuscript Decipherment, Preservation and Transliteration : Leveraging BHASHINI

Abstract

India's manuscript heritage spans millennia, encoded in diverse scripts ranging from widely used Indic scripts to undeciphered systems such as Indus, Gilgit, and Sankha. However, the lack of scalable digitization pipelines, noisy OCR outputs, and transliteration gaps limit accessibility. This paper presents an integrated AI-powered OCR framework aligned with the BHASHINI multilingual stack to address three critical problem statements: **(1) Project Lekhya: Scalable AI Frameworks and script decipherment and OCR of Indic manuscripts**, **(2) Transliteration of Modi script into Devanagari**, and **(3) Decipherment of ancient Indian scripts such as Indus, Gilgit, and Sankha**. By combining modular OCR, scholar-in-the-loop validation, and cross-lingual pipelines, BHASHINI can bridge gaps in digitization, preservation, and dissemination of India's manuscript knowledge systems.

1. Introduction

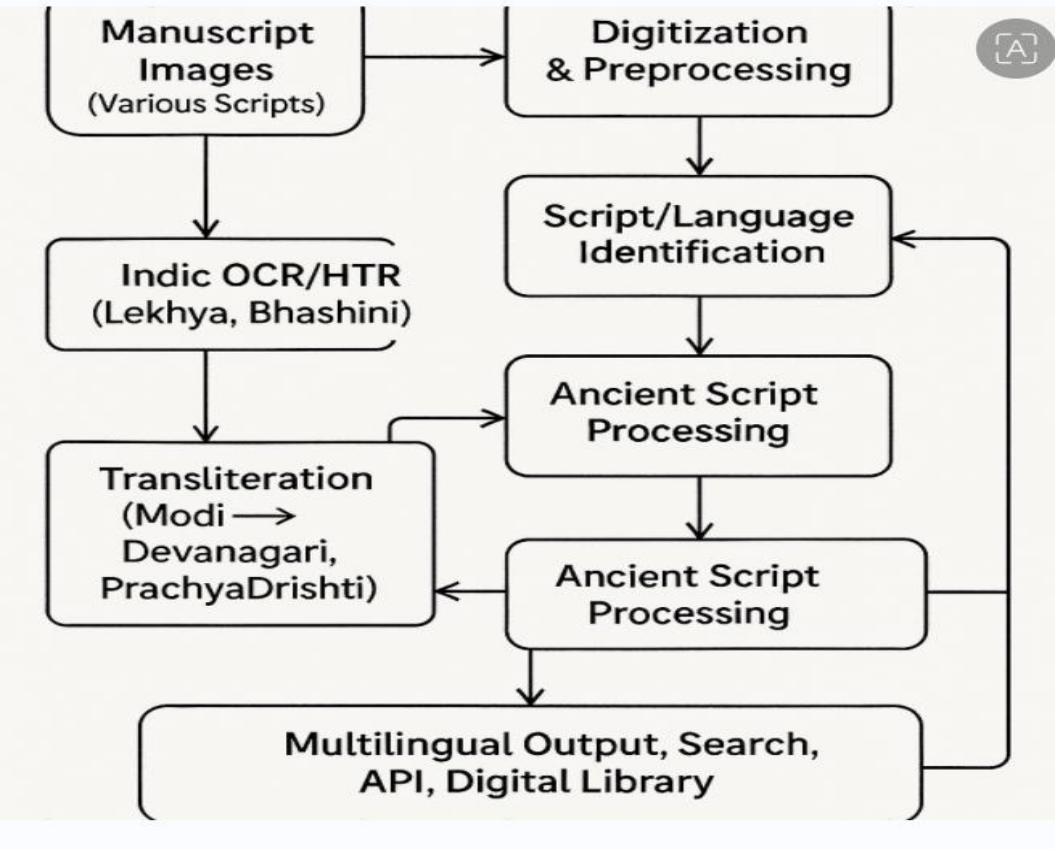
India possesses millions of manuscripts dispersed across archives, princely libraries, and private collections. These manuscripts, often in fragile states, encode knowledge across philosophy, science, medicine, and linguistics. current challenges include:

- Poor OCR/HTR performance on Indic and low-resource scripts.
- Script transliteration gaps, e.g., Modi to Devanagari.
- Undeciphered scripts (Indus, Gilgit, Sankha) lacking annotated corpora.
- Metadata linking & accessibility gaps limiting public and scholarly use.

Projects like Lekhya AI, PrachyaDrishti, and Vidhvanika demonstrate the transformative role of OCR and AI in tackling these challenges. By integrating with BHASHINI's multilingual ASR/NLP stack, a scalable, interoperable manuscript digitization ecosystem can be realized.

2. System Architecture

The proposed BHASHINI-powered OCR framework follows a modular architecture:



1. **Data Acquisition & Digitization**

   - High-resolution scanning of manuscripts (Indic, Modi, Indus, Gilgit, Sankha).
   - Metadata cataloguing aligned with IIIF standards.

2. **OCR & Script Processing**
    - Indic OCR/HTR models trained on annotated corpora (Project Lekhya).
    - Transliteration pipelines (Modi → Devanagari via MoScNet & Student-Teacher model, Project PrachyaDrishti).
    - Ancient script processing (Indus/Gilgit/Sankha) using symbol recognition, restoration, and pre-training with scholar-validated datasets.
3. **Scholar-in-the-Loop Validation**
    - Human experts refine OCR/transliteration outputs to ensure accuracy.
    - Crowdsourcing platforms integrated with BHASHINI.
4. **Knowledge Dissemination**
    - Multilingual search and cross-script linking.
    - APIs for academic access and integration into digital libraries.

3. Evaluation Metrics

To measure system effectiveness, the following metrics are proposed:

- **OCR Accuracy (CER/WER):** Character and word error rates across scripts.
- **Transliteration Accuracy:** BLEU/ChrF scores for Modi → Devanagari mapping.
- **Symbol Recognition Precision:** For Indus/Gilgit/Sankha decipherment tasks.
- **Human Validation Rate:** Percentage of AI outputs approved by scholars.
- **Efficiency:** Inference time, resource consumption, and scalability across cloud/BHASHINI infrastructure.

4. Where Bhashini OCR Can Fill the Gaps

1. **Static Bitmap Images of Manuscripts**

   - *Problem*: Millions of manuscript and artifact images are currently preserved only as static, non-searchable bitmap images.

   - *Bhashini OCR Role*: Converts these into editable, searchable text across scripts, enabling indexing, search, and scholarly use.

2. **Inefficient OCR/HTR for Ancient & Low-Resource Scripts**

   - *Problem*: Contemporary OCR/HTR systems are noisy for ancient Indic scripts like Indus, Gilgit, Sankha, or Modi.

   - *Bhashini OCR Role*: Provides script-specific OCR models, fine-tuned on scholar-validated datasets, improving recognition accuracy.

3. **Cross-Script Transliteration Challenges**

   - *Problem*: Transliteration between historical scripts (e.g., Modi → Devanagari) faces low accuracy due to lack of standardized datasets.

   - *Bhashini OCR Role*: Works as a bridge technology—first digitizing manuscripts through OCR, then integrating with transliteration frameworks (like PrachyaDrishti).

4. **Low-Resource Custodial Tooling**

   - *Problem*: Many institutions lack tools for handling fragile manuscripts.

   - *Bhashini OCR Role*: Offers cloud-based, lightweight OCR services accessible even in low-resource environments.

5. **Gaps in Metadata**

   - *Problem*: Metadata enrichment is incomplete due to unreadable or degraded texts.
   - *Bhashini OCR Role*: Extracts text for metadata generation, linking manuscripts with digital libraries.

- **Summary of OCR Application in Indic Manuscript**

Problem Statement	OCR Application
1. Project Lekhya: AI Frameworks for Script Decipherment and OCR of Indic Manuscripts	Converts fragile manuscripts into editable, searchable text using script-specific OCR models.
2. Transliteration of Modi Script into Devanagari	OCR extracts text from Modi manuscripts before transliteration.
3. Decipherment of Ancient Scripts (Indus, Gilgit, Sankha)	OCR recognizes symbols and patterns from degraded manuscripts and inscriptions.

5. Future Directions

1. Scalability to Other Scripts: Expanding OCR/transliteration pipelines to other endangered Indic scripts.

2. Cross-lingual AI Integration: Seamless translation across modern Indian languages using BHASHINI NLP.

3. AI-Enhanced Restoration: Image enhancement and denoising for degraded manuscripts.

4. Public Access Libraries: BHASHINI-powered National Indic Manuscript Digital Library with multilingual interfaces.

5. Interdisciplinary Research: Integration with archaeology, history, and linguistics for holistic cultural study.

6. Ethical and Cultural Considerations

- Authenticity: Scholar validation must remain central to avoid misinterpretation of cultural heritage.

- Cultural Sensitivity: Manuscripts with sacred/religious significance require ethical guidelines for digitization and public access.

- Open Access: Balance between public digital libraries and respecting custodial rights of traditional institutions.

- Bias Reduction: Avoid Western-centric AI training biases by prioritizing Indic corpora.

7. Citations

1. Sh. Bharath Rao – *A Cryptanalytic Decipherment of the Indus Script.*

2. Dr. Nisha Yadav – *Understanding the Indus Script: How Far Have We Come.*

3. Shri M.D. Srinivasa – *The Untapped Wealth of Indian Manuscripts on Bhāratīya Jñāna Paramparā.*

4. Dr. Shiv Shankar – *Decipherment of Indus Script and Symbols in the Light of Tribal Culture of Bastar.*

5. Sh. Gautam Raj Anand - *A Structural Grammar Decipherment of the Indus Script: Rule-Based Decoding Without Linguistic Assumption.*

Thank You

Divyansh Rawal