AnciDev: A Dataset for High-Accuracy Handwritten Text Recognition of Ancient Devanagari Manuscripts

Authors: Vriti Sharma, Rajat Verma and Rohit Saluja

Introduction

- Indian manuscripts contain **millennia of accumulated knowledge** in mathematics, astronomy, medicine, linguistics, philosophy, and logic.
- India's manuscripts represent **hundreds of languages and scripts** making them essential for understanding linguistic evolution and preserving endangered languages.

Figure1: Ancient Indian Pandulipi Manuscripts

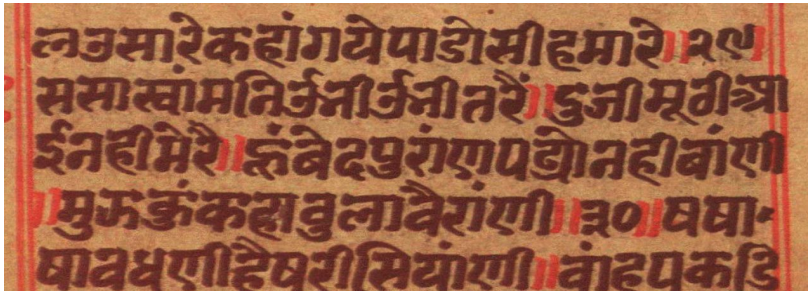
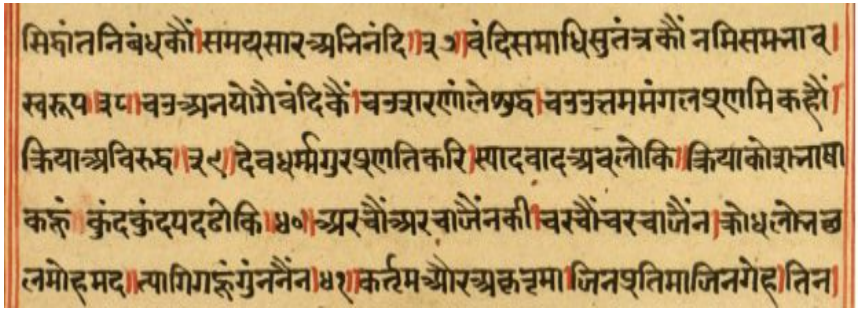
Unavailability of open-source dataset on Indian manuscripts

Goal: To propose a 500-manuscript-page dataset with approximately 3000 lines from diverse manuscripts in Devanagari script

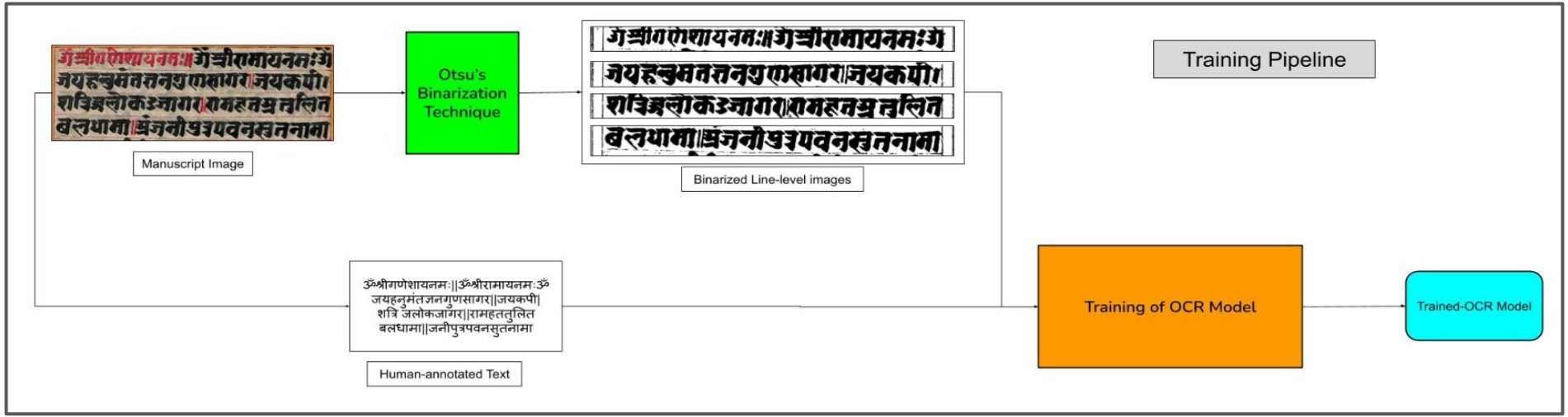
Advantages of this dataset:

- Sufficient training data: 3000 lines provides a meaningful corpus for training deep learning models, particularly for tasks like manuscript analysis or character segmentation.
- Diversity enables generalization: Drawing from diverse manuscripts creates a robust dataset that helps models generalize specific patterns or document types.
- Benchmark potential: A well-curated dataset of this size could serve as a standard benchmark for the research community.

Dataset Samples

	लनु सारे कहां गये पाडोसी हमारे ॥२९॥ ससाखामनिर्वभीर्वभीतरें दुजामूग आ ईन ही मेरे॥ झंबेदपुराण पद्यो न ही बांणी मुककुंकहावुलावैरांणी॥३०॥षषा षावधणी हैषरी सियांणी॥वांहपकडि
	सिद्धांतनिबंधकौं । समयसारअभिनंदि ॥३१॥ वंदिसमाधिसुतंत्रकौंनमिसमभाव । स्वरूप ॥३२॥ चउअनयोगेवंदिकैं । चनशरणांले श्रुछ । चउउत्तममंगलशणमिकहौं । क्रियाअविरुद्ध ॥३९॥ देवधर्मगुरप्रणतिकरि । स्यादवादअवलोकि ॥ कियाकोशभाषा करूं ॥ कुंदकुंदयददढोकि ॥४०॥ अरचौंअरचाजैंनकी । चरचौंचरचाजैंन । क्रोधलोभछ लमोहमद ॥ त्यागिगफूंगुंननैंन ।४१॥ कर्तमऔरकृत्रमा । जिनप्रतिमाजिनगेह । तिन ।

Training Pipeline



Attention-OCR and CNN-RNN Finetuning

Finetuning of Attention-OCR and CNN-RNN:

No. of Epochs	200
Batch Size	32
Total pre-training dataset size (synthetic)	820 (fonts) x 5000 per Font = 4100000 line-level dataset
Total Dataset Size	3085 lines
Training Data Size (80:20 split)	2458 lines
Test Dataset Size (80:20 split)	627 lines

Tessaract-5 Finetuning

Fine-tuning of Tessaract-5:

No. of Epochs	200
Batch Size	32
Total pre-training dataset size (synthetic)	7000 synthetic lines creating using real verses text + 3000 real lines
Total Dataset Size	3085 lines
Training Data Size (80:20 split)	2458 lines
Test Dataset Size (80:20 split)	627 lines

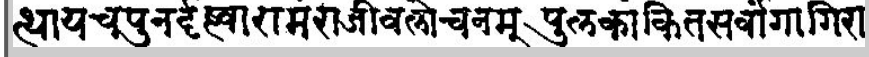
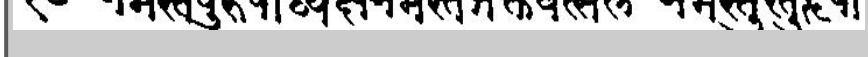
Quantitative Analysis

1. **Character Error Rate:** (Number of Incorrect characters in predicted text / total number of characters in reference text) * 100
2. **Word Error Rate:** (Number of Incorrect words in predicted text / total number of word in reference text) * 100

Model Name	Test Set	
	WER	CER
CNN-RNN	98.20	48.59
Attention-LSTM	96.73	46.33
Tessaract-5	87.42	30.06

m, s denotes manuscript and synthetic data respectively

Qualitative Analysis

Input image	Ground Truth	Tessaract-5's Predictions
	राधन तणा॥ अछैदश अधिकार॥चित्तआंणीनें आदरो जि	राधन तए ॥ थैदश प्रधिकार॥ चितआणीनें प्रादरो लि
	धत्तेस्वतंत्रःपरिपूर्णआत्मा ९० नमोस्तुतेरामतवाघ्रिपंकजं	धक्तेस्वतंत्रःपरिपूर्ण आत्मा ५० नमोस्तुतेरामतवाघ्रिपंकजं
	त्थाय चपुनर्द हवा रामं राजीवलोचनम् पुल को कि त सर्वांगा गिरा	त्थायचपुनर्टत्वारामंराजीवलोचनम् पुलकां कितसवागागिरा
	५८ नमस्तेपुरुषाध्यक्षनमस्तेभक्तवत्सल नमस्तुस्तुहषी,	५६ नमस्तेपुरुषाध्यक्षनमस्तुभक्तवतलल नमस्तुस्तु हर्षी,
	चोप बाढी सुख कवलक भूत्यै चंद्र ते कांति काढी परम बदन रंभारे	वोप बाटी सुरव कचलक भूत्पै वंद्र ते कांति काटी परम वदन एभारे
	शशि निभशुभ वदने । सिंहविराजितगम ने ही	शमि नमभुभ वदने । सिंहखिणजिचगम नही

Green and red color for correct and incorrect prediction

Conclusion and Future Work

Contributions:

- First public dataset for ancient Devanagari HTR (3K lines, 500 pages)
- Reproducible baselines with 3 HTR models
- New state-of-the-art: 30.06% CER (Tesseract-5)

Future directions:

- Dataset expansion: More manuscript types and time periods with diverse scribal hands for better generalization
- Model improvements: Advanced transformer architectures, specialized data augmentation for historical writing, language model-based post-processing
- Evaluation: Multi-annotator studies for reliability analysis

Thank you

For more info, scan the given QR code: