Machine Learning Based Handwritten Historical Text Extraction and Translation

Kavya Sundara Moorthy
Department of CSE, Easwari Engineering College, Chennai, India

PVS Deepti
Department of CSE, Easwari Engineering College, Chennai, India

Radhika Swaminathan
Department of CSE, Easwari Engineering College, Chennai, India

Dr. S. Kayalvizhi
Department of CSE, Easwari Engineering College, Chennai, India

Abstract:

The corpus of 11th century Tamil temple inscriptions offers a rich tapestry of societal, cultural, religious, and administrative insights, which are frequently obscured due to their complexity and scarcity. Comprehending these epigraphs typically requires proficiency in the ancient Tamil script, a task that poses significant challenges. This paper introduces an innovative Machine Learning framework tailored for the extraction, translation, and digital categorization of handwritten historical Tamil text from the 11th century. The system's foundation is a specialized Optical Character Recognition (OCR) pipeline that employs Neural Network architectures, specifically adapted to the intricate nature of ancient Tamil script, which often eludes conventional OCR methodologies. The system employs a multi-step process, including preprocessing for noise reduction and specialized segmentation techniques that leverage character context and sequence. To address the scarcity of annotated historical datasets, handwritten sample sets that realistically reflect ancient scripts were created, augmented to mimic inscription degradation, and further passed through transformations to synthetically generate more training data. This enables the model to learn robustly despite limited real-world data. By making ancient manuscripts and inscriptions more understandable, this scalable solution advances digital humanities, preserving cultural heritage and facilitating academic research on other low-resource historical scripts.

Keywords:

Optical Character Recognition, Text Extraction, Language Translation, Generative Models, Dataset Synthesis.