CS663 Course Project Work Proposal

Badri Vishal Kasuba - 22M2119
Abhishek Kumar Singh - 22M2104
Jahanvi Rajput - 23D0378

1 Introduction

India has rich culture and heritage in it's history. Ancient scholars have piled up their teachings in the form of manuscripts in palm-leaf and paper documents. Digitizing such documents has huge value in understanding Indian history and the knowledge that has been passed on. But Manuscript digitization is challenging because of no available system to digitize and manual digitization is very time-consuming. Deep-learning methods have recently seen a lot of boost in digitizing various documents of printed, handwritten, and scene-texts producing state-of-the-art results and near human performance. So exploring **manuscript digitization problems using DL techniques** could be a great interesting problem to work with which we wish to explore as part of our project in this course.

2 Challenges found

The Palm-leaf documents are generally old and there are various complexities in developing a system to digitize manuscripts. We downloaded some of the palm-leaf documents available online and encountered various challenges[4], some of which are

- Texts inside manuscripts are densely packed and separation of words and text detection is tough unlike other text documents
- Texts are written in ancient writing style or scripts, in which some characters do not exist in the existing Unicode framework to represent text in machine-readable form
- The manuscripts are very old and some portions of them are damaged due to water, aging, and maintenance issues leading to degraded information in images for corresponding texts, so digitizing damaged areas is challenging

3 Our Work Proposal

Taking into the account challenges that we observed through our preliminary analysis of the project work, we looked out for initial works in this area in-Academia and found Layout Segmentation work done by IIIT-Hyd on palm-leaf manuscripts which we wish to implement the inference part of their paper since the data is not publicly available. Also, annotated data was not present in other places, so we wanted to do the following things for our project.

- Implement the inference part of the research papers, SEAMFORMER and PALMIRA which does layout segmentation on top of the manuscripts that we have collected
- Performing some image processing-based methods to have the best representation of OCR the manuscripts
- Implementing and testing out current Indic languages supported OCR models to check the performance on manuscripts like Tesseract, Google OCR, etc
- Looking out for any resources that support digitizing manuscripts done by some open source or from Academia works

4 Limitations

Manuscripts are found in different languages and especially in various scripts. For the project, we wanted to limit to digitizing only Devanagari script manuscripts and if possible work on Tamil script manuscripts and also work on a limited set of manuscripts from diverse sets. Also, since there is not much annotated data to develop the system, we would explore possible processing methods to be used for better representation of manuscripts for downstream tasks usage.

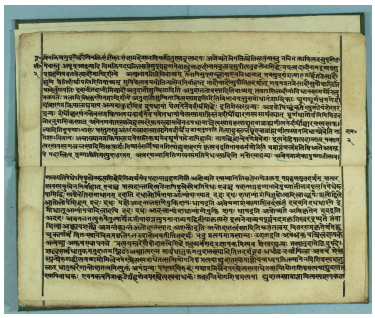

Figure 1: *
Describing the dense words present

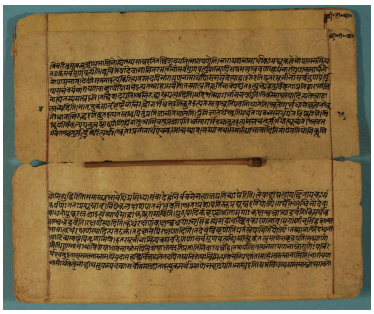

Figure 2: *
Describing the complexity of characters

Figure 3: *
Describing the lost information and clarity of text content

Figure 4: *
Describing the ancient writing style