Progress Report

Sharada Script

April 11, 2023

Overview

Sharada is an ancient script of the Brahmic family. Among the numerous attempts made towards digitization of Sharada in the past decade , there has been a lack of an automated character recognition system. In this project we plan to fill in that particular void by bringing about an OCR system.

Goals

1. **To perform OCR on ancient scriptures written in Sharada :** Creation of an application which makes use of a deep learning model to perform OCR to decrypt the ancient scriptures
2. **Transliteration of Sharada script:** Transliteration of Sharada script to Devanagari , IAST & ITRANS.
3. **Translation of Sanskrit(in Sharada script) to English:** Performing translation of preserved scriptures to English and other Indian languages.
4. **Creation of a web-portal:** Creation of a web-portal to educate people about the language and to encourage them to learn the script.
5. **Creation of a dataset with annotated scriptures:** The annotated dataset allows further development and improvement of the previously created OCR system.

Milestones Reached

1. Transliteration System
    - A transliteration pipeline was implemented in python.
    - The pipeline allows transliteration from Sharada to Devanagari , IAST & ITRANS
2. Ongoing creation of a annotated dataset

    Scriptures of Bhagavad Gita and Shiva Stotra are being annotated manually using LabelMe.

3. Created a web-tool to perform sandhi-splitting of a sanskrit word

    We created a streamlit application to perform sandhi splitting on IAST words of sanskrit. The system could identify sandhi type and base words.

Goals for the coming month

1. OCR Model
    - We plan to make use of the annotated dataset to train a CNN model.

2. Dataset
    - We plan to create an annotated dataset which is structured in the following manner: The data is in the form of a 2D array. The first dimension indicates the indexed letter in the dataset. The second dimension is again a 1D dimensional array containing three elements- The first element is the image in array form. The second element is the corresponding class index number. The third element is the corresponding English class Annotation.

Links :

GitHub - sud0x00/Sharada: Website to perform OCR on ancient Indian script Sharada

GitHub - sud0x00/Samskritam: Projects & scripts related to samskritam.

GitHub - SharadaNLP/SharadaTransliteration: Performs transliteration from Sharada to Devanagari , IAST & ITRANS

GitHub - shakthivels300/Sanskrit-Simple-and-Compound-Character-Recognition-Using-R-CNN-: Sanskrit OCR

GitHub - avadesh02/Sanskrit-letter-dataset