Department of Computer Science

LIPI - Malayalam OCR System

Aarya R Shankar
Amrith M
Anand R
Sarathchandran S

1 Scope

To develop a high quality OCR of Malayalam language using deep learning and computer vision within 3 months, after preparing the dataset from 3+ scanned copies of malayalam literature.

2 Technical Feasibility of Project

Our project is completely based on open source libraries and softwares. Hence the only potential problems that may arise during the implementation of our project will be during the collection of data set and training our algorithm using collected data.

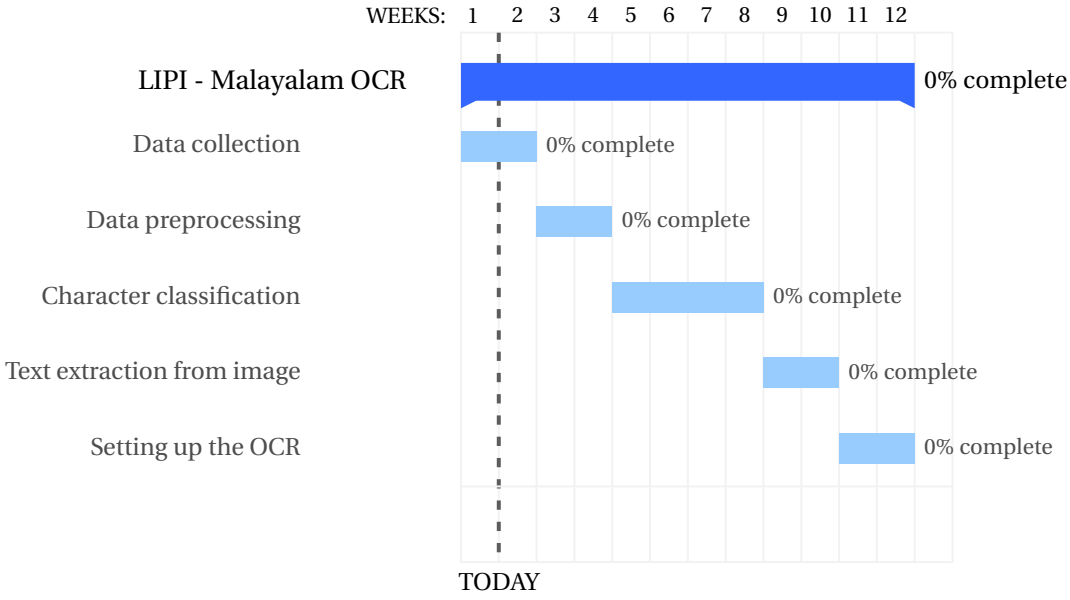
3 Project Benefits

Our projects has various benefits. Some of them are noted below:

1. Our Kerala Government has enlisted Malayalam as the language of all official documents. Hence, a malayalam OCR is necessary for digitalising them.
2. It can be used for other open source projects for organisation purposes in NGOs, etc

3. This project may be modified to read the malayalam scripts and process them using NLP (Natural Language Processing) and use a text-to-speech software as an aid to blind people.

4 Timeline



5 Resources

1. Nvidia GPU processor provided by Kerala Startup Mission
2. Keras - neural network library written in Python
3. numpy - Library for scientific computing with Python
4. matplotlib - excellent plotting and graphing libraries
5. tensorflow - Depplearning Library
6. tiano - Deeplearning Library
7. pandas - Python version of R dataframe
8. IPython - with the additional libraries required for the notebook interface.
9. openCV - Computer vision Library

6 Risks

1. Finding a large dataset to pretrain on
2. Difficult to debug in between models which is not working
3. Training a large set of dataset need a very efficient GPU
4. Updating the pretrained model when more data or better techniques becomes available