AI & OCR

Organization – Government of Andhra Pradesh

Category – Software

Domain Bucket – Software - Web App development

Problem Code – KB141

THE TEAM – VOID ABSTRACTORS

- TEAM LEADER – G Mukkesh [17BCE1128]
- TEAM MEMBER – Rohit Subramanian [17BCE1291]
- TEAM MEMBER – Amrit Krishna O [17BCE1133]
- TEAM MEMBER – Sanjana Dulam [17BCE1068]
- TEAM MEMBER – Akshay Kumar [17BCE1290]
- TEAM MEMBER – Maheshvar C [17BCE1172]

PROBLEM STATEMENT

AI & OCR – To search Telugu & Urdu words in PDF present in Unicode as well as image format.

THE GOAL

Optical Character Recognition (OCR) is a software to convert printed text and images into digitized form such that it can be manipulated by machine.

- The goal is to be able to input a pdf file using a web application and read its data.
- Data present in the pdf is a combination of English , Telugu and Urdu languages. Data can be present as either Unicode or in Image format.
- We need to be able to search the data present in the pdf in English , Telugu and Urdu.

TECHNOLOGY STACK

Tesseract OCR

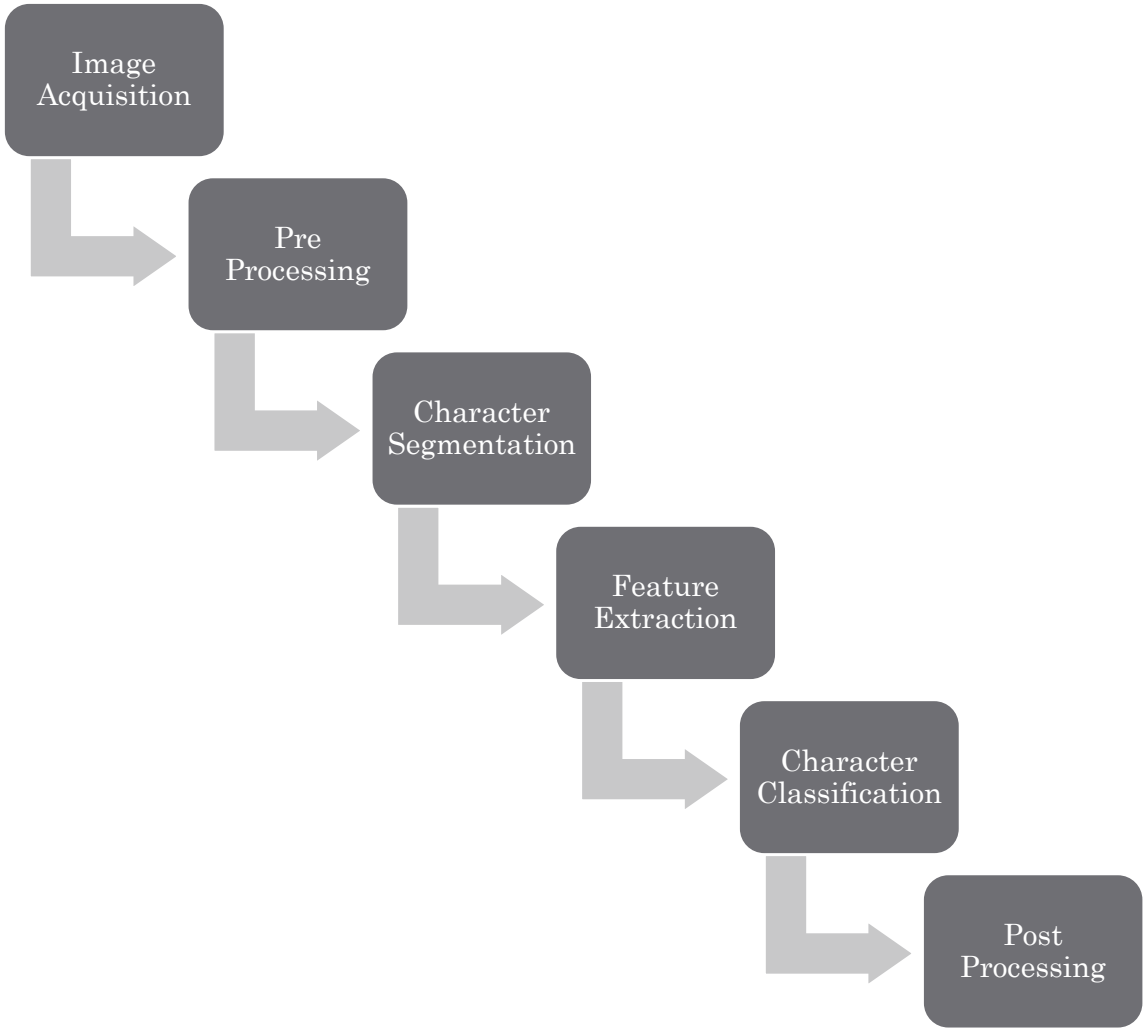
TECHNOLOGY STACK (contd.)

- Python is an interpreted, high-level, general-purpose programming language used for the development of our software.
- Python-tesseract is an optical character recognition (OCR) tool for python. It will recognize and “read” the text embedded in images.
- Python-tesseract is a wrapper for **Google’s Tesseract-OCR Engine**. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.
- Flask is a lightweight WSGI web application framework. It is designed to get started quick and easy, with the ability to scale up to complex applications.

OUR PLAN

- We are building a web application using Python and Flask framework.
- The user will be able to upload a pdf to the site . On getting the pdf document the application , converts each page of the pdf to an image (eliminating the necessity of creating multi channel input/outputs) .
- Then we use the Pytesseract - an optical character recognition (OCR) tool for python. It will recognize and “read” the text embedded in images. The result is stored in a text file . These models also have the ability to be retrained on custom fonts .
- When the user wants to search for anything in the document , the application searches the extracted text file and gives the result.
- The user is also given an option for downloading the extracted text file.

MAJOR PHASES OF OCR



USE CASE / FUNCTIONALITIES

The two main features of this application are simplicity and practicality. The algorithms used - both the connected component algorithm for segmenting words into recognizable units, and the recognition algorithm based on template matching - are simple and efficient.

The main stakeholders for this web based application are :

- AP Government
- Any consumer of Telugu, Urdu and English documents

Their interactions with the below-mentioned functionalities essentially constitute the use case.

- **Upload PDF** : The user should be able to upload a pdf from their personal computer in a quick and secure way.
- **Search PDF** : The user should be able to search the pdf document for the presence of a text be it in Telugu , English or Urdu.
- Support for multi-lingual document OCR

FEASIBILITY STUDY

For the Feasibility Study , we would like to answer 3 basic questions.

- **Problem Statement Validity** : Recognition of characters from document images is at the heart of any document image understanding system. The activities and results on Indian language OCR are grossly insufficient even today. It is our view that an OCR engine has a much better impact as part of real-life applications than along with word processors.
- **Social and Ethical Concerns** : This will help the society in a great deal as it bridges the gap between the different languages and removes barriers .
- **Organization Capabilities** : Our team of 6 members are highly talented and capable and will be able to develop this software efficiently .

LITERATURE REVIEW

We have done some literature review before starting this project :

- Recognition of characters from document images is at the heart of any document image understanding system. The activities and results on Indian language OCR are grossly insufficient even today. Character recognition is not a new problem but its roots can be traced back to systems before the inventions of computers. The earliest OCR systems were not computers but mechanical devices that were able to recognize characters, but very slow speed and low accuracy.
- Telugu is the language spoken by more than 100 million people of South India. Telugu has a complex orthography with a large number of distinct character shapes (estimated to be of the order of 10,000) composed of simple and compound characters formed from 16 vowels (called achchus) and 36 consonants (called hallus).
- Mostly research has been done for Urdu OCR is with respect to scripts, fonts and text environment which are other obstacles in the way of making complete OCR. So we have research and moderately implements online and offline OCR system which is irrespective of Urdu scripts and fonts.

LITERATURE REVIEW

These are the papers we have referred for the literature review :

- A Survey on Optical Character Recognition System [Noman Islam, Zeeshan Islam, Nazia Noor]
- An Overview of the Tesseract OCR Engine [R. Smith]
- A Bilingual OCR for Hindi-Telugu Documents and its Applications [C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran]
- An OCR system for Telugu [A. Negi ; C. Bhagvati ; B. Krishna]