Machine Learning Based OCR & text detection

Overview:

In order to digitize books you need a high tech scanner but what if you want to digitize damaged archives ? that make scanning and text detection more and more complicated.
This project, AuroraPi, start with the aim of digitization of the league of nations archives. But we thought about libraries around the world who cannot afford the price of high-tech scanner.

The aim of AuroraPi is to make a cheap scanner as well as portable with high accuracy, 98%, for carrying out text detection for every book across the world. Also it uses google cloud vision based text detection to detect hand written text

The scanner use a high level protocols of machine learning based on APIs (Google Cloud Vision), Natural Language processing algorithm and a Raspberry-Pi. The same scanner can also be used in order to generate digitized copies for notes and can be personalised as a device for individuals (Students, professors ..). Moreover, a natural language processing API that is implemented in order to get the summary of the document as well as the labels that are being identified which can be used in order to find the information with search engine requests.

Description:

AuroraPi uses a Raspberry Pi with a NoIR camera (for better low light imaging) in order to capture the image of the text in a document and uploads it to the google cloud where the vision (a machine learning based text detection API) recognizes the text, send the recognized text (En & FR) to a file with different format (.text, .pdf, .docx …) and store it
As only 60% of knowledge is available on the internet, the ultimate goal of this project is to make this technology available for everyone to digitize documents and make this digitized data accessible online in soft copies.

Development:

The project consist of two aspect, Electronic Engineering and Mechanical Engineering.

- In the first aspect, we would like to use Raspberry Pi zero instead of Raspberry Pi 3. Also implement the white and IR LED strips around the camera which gives it better focus and sharp images of the document.
- In the second aspect, we would like to design a frame to have an Auto page turning scanner. This frame would contains a plug where the AuroraPi could be detachable.

The project has been tested on several images and is successful in carrying out text detection and get the output in a .txt format.

Technologies to be used:

- **Google cloud vision text detection API -** For carrying out machine learning based OCR.
- **Google natural language processing API -** For extracting information from archive.
- **Microsoft Excel:** Making database for the trash classification device.
- **Raspberry Pi:** For clicking pictures with camera and providing automation platform.
- **Python:** For programming the Raspberry Pi.
- **Google cloud vision label detection API:** For carrying out machine learning based trash detection.
- **Microsoft Oxford vision API:** For making the results more accurate and verifying them.

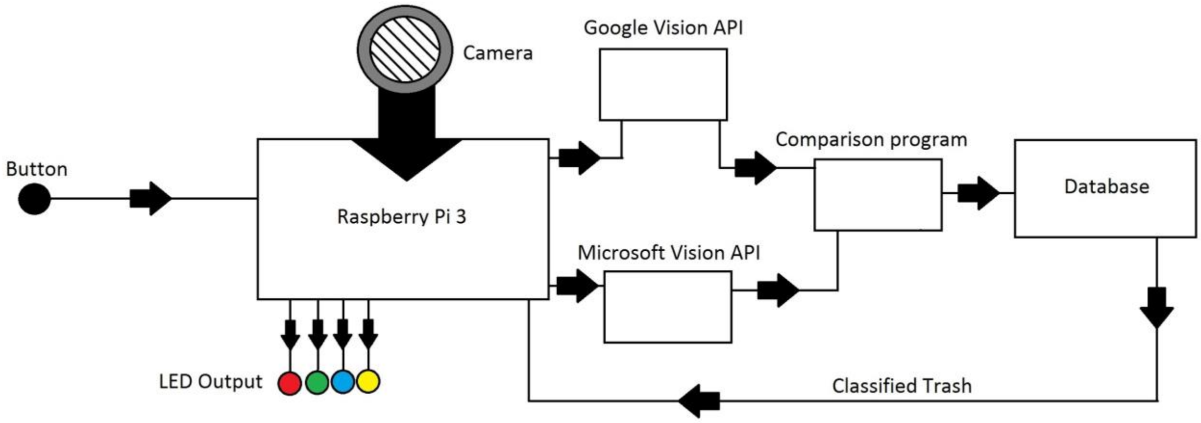

Fig 1. Architecture of Trash classifier

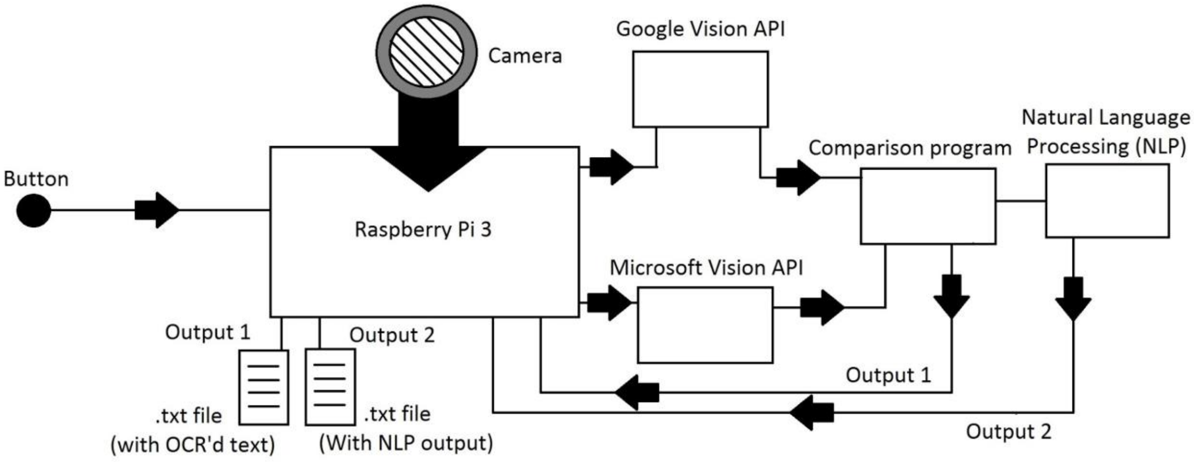

Fig 2. Architecture of Text Detector

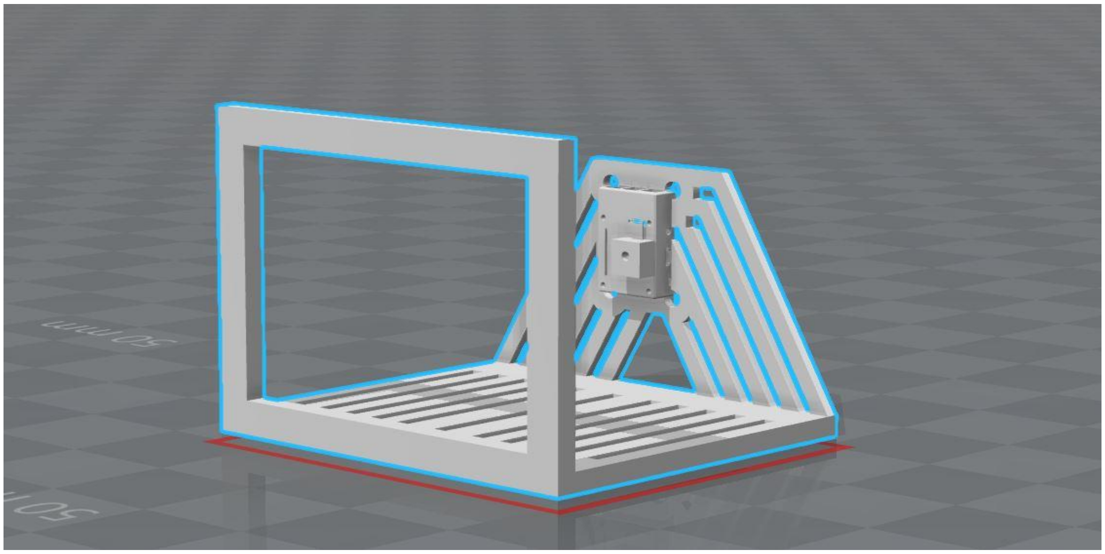

Fig 3. Structure of Stable Text detector 3D printed frame