Term Paper:
Real Time OCR for Regional Languages

Submitted by: Manish Kumar Rohilla, 11714083, K17JT

Submitted to: Professor. Sanjay Kumar

B. Tech Computer Science and Engineering, Student

kotamanishrohilla@gmail.com

Abstract- OCR means Optical Character Reader is the electronic conversion of Images which contains typed data, printed or handwritten text into machine-readable encoded text, whether from a scanned text document, a photo of a text document, a scenario photo (for example texts on signs boards and billboards) or from subtitle text super-imposed on an image.

Keywords – OCR Optical Code Reader, Tesseract, Google Text-to-Speech(GTTS), Google Speech-to-Text(GSTT), Computer Vision.

INTRODUCTION

As this era is an era of digitalization and technology each individual has access to the Internet. With ease of access to Internet people from almost every culture and diversity uses the Internet. With diversity comes differences which involves difference in their speaking style, writing, characters or language. Everyone is not linguistic to be able to read or understand every language, that will be pretty Impossible. That's where the need of OCR and Translator comes in. If u ever come across a document which is hand written and you want to know what that document actually wants to convey. First you have to learn the language the meaning of characters then translates it or you have to find someone who can read it for you but that will also have some security issues if the document is confidential. But with OCR and translator combine together its pretty easy, you just have to click a picture of the document and give to OCR to extract characters and then transfer those characters to translator and Walla! It is translated. This project's sole purpose is to show how the OCR works and how the translator is associated with OCR to make this process of reading and translation successful.

Tesseract OCR Engine

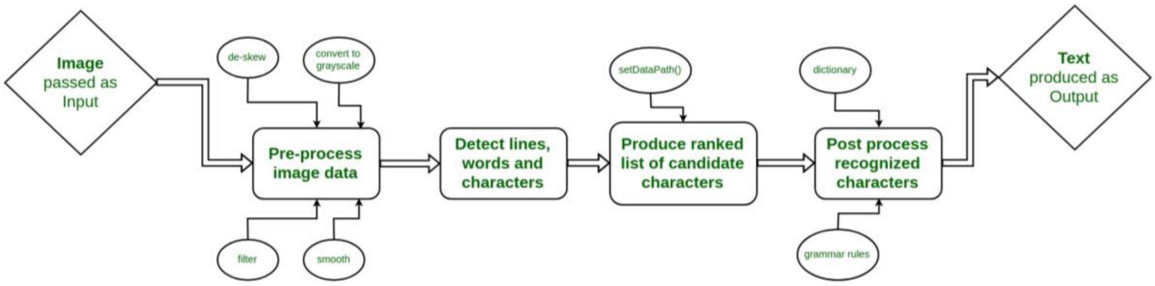
Tesseract OCR Engine is and open source OCR engine for image text recognition. It is developed by Hp Laboratories in 1985 and open sourced in 2005 later then from 2006 its is being developed by google. Tesseract has Unicode UTF-8 support and has the capability to recognize over 100 languages and can also be used for developing different scanning software also. It consists a latest neural net LSTM based OCR engine which is more focussed on line-recognition but also supports the legacy Tesseract OCR engine which also works by recognizing pattern of the character.

How OCR works?

OCR works in following steps:

- At First the Image is processed and converted into some X cross Y matrix, then each and every cell of the matrix is weighted according to Gray-scale.
- The Weighted cells of the matrix is passed through various filters like Smooth and De-Skew Filter.
- Then the lines and words will be detected
- Then the ranked list of candidate characters based on trained data set is produced. (For setting the path of trainer data the setDataPath() method is used)
- In Post-process recognized characters will choose best characters based on confidence from previously executed steps and language data. Language data includes some dictionary, grammar rules, etc.

General working of OCR



Advantages of Tesseract OCR:

The advantages of OCR are numerous, but namely:

- It increases the effectiveness and efficiency of official work.
- The capability to instantly search through the content of an image's text data is very useful, especially in an official setting that has to deal with high volume of scanning or high document inflow and outflow.

- OCR is quick in ensuring the document’s text content remains unchanged and unaffected while saving time as well.
- Workflow is increased immensely since employees no longer have to waste.

Disadvantages of Tesseract OCR:

- The OCR can only be used for language recognition.
- A lot of effort is required to make or maintain trainer data of different languages and implement that.
- Some extra efforts are also required when it comes to Image Processing as it is only part that really matters when it comes to the performance of OCR.
- After doing such a heavy amount of work, there is no OCR that can offer an accuracy of 100% and even after OCR we have to determine the unrecognized character by certain neighbouring methods of machine learning or manually correct it.
- Time on manual labour is decreased and thus work can be done quickly and more efficiently.

How to use Tesseract OCR:

- The first step is to download the tesseract API installer. Or directly type in IDE command prompt: *pip install pytesseract*
- Extract the files from the downloaded rar file.
- Open your IDE and make a new project and link the jar file like this:

  pytesseract.pytesseract.tesseract_cmd= 'C:\\Program Files\\Tesseract-OCR\\tesseract.exe'

Google Text-To-Speech

gTTS (*Google Text-to-Speech*) is a Python library and CLI equipment to interface with Google Translate's text-to-speech API module. gTTS takes spoken audio mp3 data to a file then a file object of byteString for further audio manipulation and processing or standard output is created. Or simply pre-generate Google Translate TTS request URLs to provide to an external program at redthedocs.org.

Features:

- There is an inbuilt Customizable speech-specific sentence tokenizer that makes it possible for unlimited lengths of text to be read while keeping proper intonation, abbreviations, decimals etc;
- Further this library also consists Customizable text pre-processors which can provide pronunciation corrections.
- Supported languages are retrieved automatically.

Open Computer Vision (OpenCV)

OpenCV (Open source computer vision) is a module of programming functions. The main aim of this library is real-time computer vision. It was developed by Intel; Later it was supported by Willow Garage and then by Itseez (which was owned by Intel after some time).

OpenCV mainly supports various models of certain deep learning frameworks which includes TensorFlow, PyTorch after converting to an ONNX model and according to a defined list of supported layers it also supports Caffe. It promotes and generalize Open Vision capsules which is compatible and a portable format with all types of formats.

OpenCV's application areas include:

- 2D and 3D feature toolkits
- Facial Recognition System
- Recognition of Gestures
- Interaction between computer and human
- Motion understanding
- Object identification
- Tracking of Motion

OpenCv also contains some statistical machine learning libraries which includes:

- Boosting
- Decision Tree Learning
- Gradient Descent and Boosting techniques.
- Maximization of Expectation Algorithm
- KNN (Nearest Neighbour) Algorithm
- Naïve Bayes and probabilistic Algorithms
- Support Vector Machine
- Random Forest Classifiers

This library module has more than 2500 optimized algorithms which further includes a comprehensive and effective set of both classic dynamic computer vision and machine learning algorithms. OpenCV has approximately more than 47 thousand people of user community, library is extensively used by companies, research groups and by governmental bodies. It has various programming interfaces which includes C++, Python, Java and MATLAB and it also supports various Operating Systems like Windows, Linux, and Mac OS. OpenCV aims mostly towards real-time vision applications and takes use of MMX and SSE instructions when available. Library featuring CUDA and OpenCL interfaces are being developed right now with over 500 algorithms there are about 10 times as many functions that are developed to support those algorithms which are written natively in C++ and has a interface that works with Standard Template Library containers.

Project Code Briefing

List of Libraries imported :

- Import cv2 #For computer vision
- Import pytesseract #OCR engine

  Pytesseract.pytesseract.tesseract_cmd=<full_path_to_your_tesseract_exe>'
- import googletrans #For conversion to audio format and translation
- from googletrans import Translator #for detecting language and translating
- from gtts import gTTS #gTTS google text to speech
- from playsound import playsound #to play Audio files
- import speech_recognition as sr #To use Laptops Microphone to take Audio Input
- import sys #To create and provide system permissions and exclusive command
- import os #to provide driver support and exclusive thread support from kernel to IDE

References

- GeeksforGeeks
- TesseractOCR.org
- guides.library.illinois.edu/
- https://stackoverflow.com/
- https://gtts.readthedocs.io