Character Recognition and Language Translation with Optical Character Recognition

Dnyanada Padwal[1] Tejashree Mahtre[2] Roshan Chavan[3]
[1,2,3]Department of Information Technology
[1,2,3]Theem College of Engineering (Mumbai University), Boisar, India

Abstract— In our day to day life the people are facing many problems in understand the languages. For example, people in different states speak different languages they might not understand or speak other state language at that time this OCR Website will help them. Existing system, having a separate application for each and every process like camera, Google translator and Optical Character Recognition (OCR) text scanner. But, people expect the application consists of all the three facilities together. So this proposed web application provides a new idea to the people to translate the other language text into their known language. This application contains three steps. 1.Take a choose image of the unknown language text which you want to translate(printed material), 2.Tessaract is an open source Optical Character Recognition (OCR) technology, which is used to extract the text from the image then Google API and Bing API is used for translation of language. 3. The translated text is generated in PDF format. This paper presents details about translation in terms of a web application that accepts image document as an input, where input document is a user define image file containing text in any language available in the Python-tesseract library and does its exact translation in any supported languages using Google Translator (i. e Googletrans). Using the computational power the individual elements like text, images, and special characters can be distinguished. OCR-Optical Character Recognizer does the work.

Keywords: Tesseract, OCR, optical character recognition, character Recognition, document

I. Introduction

Character recognition, usually abbreviated to optical character recognition or shortened OCR, is the mechanical or electronic translation of images of handwritten, type written or printed text (usually captured by a scanner) into machine editable text . It is a field of research in pattern recognition, artificial intelligence and machine vision.

Though academic research in the field continues, the focus on character recognition has shifted to implementation of proven techniques. Optical character recognition technology was invented in the early 1800s, when it was patented as reading aids for the blind. Optical Character Recognition is the area of Pattern Recognition that has a topic of studies over the past some decades. Optical character recognition is technique of automatically identifying of different character from a record picture additionally provide full alphanumeric recognition of printed or handwritten characters, text numerical, letters, and symbols in to a computer process able layout including ASCII, Unicode and so forth. The OCR system uses a camera which acts like a human eye which that reads the characters from the selected input source in the form light and then the computations are done on the processors of OCR which just does the work of human brain's character recognition. The output of the OCR varies from system to system, it can be in form of text, speech or even image.

II. Literature Review

Character recognition technique has been completed through studies on different characters for example, English, Hindi, Marathi, Bangla, Tamil, Telugu and Kannada and so on. Totally, the complete method is carried out in three phase Pre-processing, Feature extraction and recognition. In this paper only cover the study has been done on English, Hindi and Marathi, Tamil, Telugu, Urdu scripture.

III. Module of OCR System

For making an OCR engine below are the steps which one can follow to make sure that the OCR meets the desired expectation of character recognition.

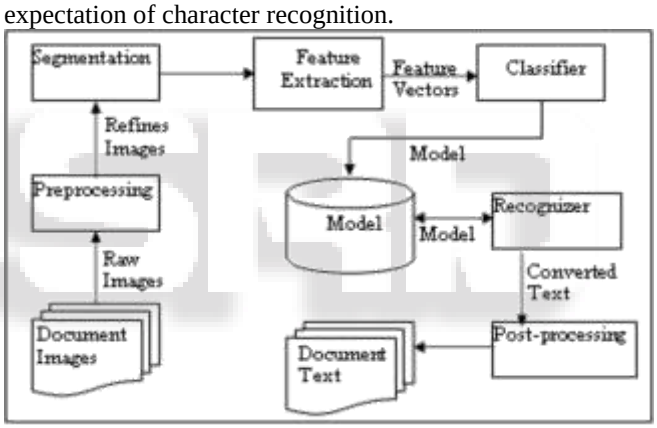

Fig. 1: Block diagram showing steps involved in OCR

A. Optical Scanning:

To start with an OCR, image can be capture by digital camera also but after seeing the challenge been faced,we need a good optical scanner. With the help of this scanner, an image of original file or document is captured. Or select a already scanned image or file.

B. Pre-Processing:

As the very first step of using OCR system the device. Select a picture or the document then converts that picture to gray-scale popularly Black and White. While doing so it removes all the extra things like images, logos, lines and dust particles or stains on the page or document if any, it converts any shades or colors of black or grey to just black and white leaving a more clear picture which just has text on it in black and white color.

C. Segmentation:

Once the pre-processing produces noise free clean character image, it's then segmented into several subcomponents. There are three steps of segmentation first line segmentation divide the character in image horizontal second word

segmentation the divide words from line sentence last character segmentation divide the characters from word. Finally we get segmented characters those character help for feature extraction and recognition.

D. Feature extraction:

This is one of the riskiest components in an OCR development. The main aim is to extract important patter from characteristics. The selected features are expected to contain pattern that differentiate one character from other and relevant information from the input data, so that the classification can be performed by using those patter extract from segmented character this instead of the complete original data.

E. Training and recognition:

Investigation of OCR's pattern recognition can be done via template matching, statistical technique, syntactic or structural techniques, and artificial neural networks. The system also has to learn in such a way that the problem associated to incomplete vocabulary is solved.

F. Post-processing:

In this final process, activities like grouping, error detection and correction take place. During grouping, symbols in the text are associated with strings. However, it's impossible to reach 100% accurate identification of characters, only some of the errors can be detected and deleted as per the context.

G. Output:

After passing through the post-processing stage the input image gets converted into a document, image or a PDF depending upon the type of OCR system used or the user's requirement. Depending upon the type of OCR used the output might or might not be in formats where editing can be done (Text document) or not.

H. Proofreading:

Once the entire process of conversion is completed the output needs to be verified as no matter how much the OCR's have advanced there is always a chance of some error. Especially when it comes to paperwork or documents it becomes necessary that it is cross-checked with human eyes.

IV. PROPOSED SYSTEM

OCR stands for "Optical Character Recognition." It is a technology that recognizes text within a digital image. It is commonly used to recognize text in scanned documents and images. It is the electronic conversion of images of printed text into machine encoded text. Images captured by a digital camera differ from scanned documents or image-only PDFs. They often have defects such as distortion at the edges and dimmed light, making it difficult for most OCR applications, to correctly recognize the text. G. R. Hemalakshmi, M. Sakthimanimala, J. Salai Ani Muthu (2017) [2].

A common problem faced by people is that of understanding unfamiliar language. Failing to understand unknown languages can lead to minor problems. It is composed of two sub-systems that perform text extraction and text translation. The extraction and translation parts are relatively well developed and there exist a large variety of software packages or web services that perform these tasks. The challenge is with extracting the exact text from the images and translating it to known language.

In a typical scenario, a user select a picture containing text, the text is extracted from the image. In image processing and computer vision, edge detection treats the localization of significant variations of a gray level image and the identification of the physical and geometrical properties of objects of the scene. The variations in the gray level image commonly include discontinuities (step edges), local extreme (line edges) and junctions. Most recent edge detectors are autonomous and multiscale then include three main processing steps smoothing, differentiation and labelling. The edge detectors vary according to these processing steps, to their goals, and to their mathematical and computational complexity. The extracted text is then translated using translation engine which contains the database of languages. Then the translated text is given as output. \Properties like color, intensity, edges etc are related in extracting the text to carry out this task we have four modules those are Image Capture, Text Identification, Language conversion, PDF generation. G. R. Hemalakshmi, M. Sakthimanimala, J. Salai Ani Muthu (2017)[2].

The text is given as input and image is got as output. The input image is first pre-processed to remove the noise present in the image. The image is converted into a gray-scale image which can then be converted into binary image. Tesseract is an Optical Character Recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006.Tesseract is considered one of the most accurate open source OCR engines currently available. The total count of support languages to over 60.It is the tool used to extract the text from an image.

After extracting the words from image by using the OCR Engine, those words are translated into known language to do this the Bing Translator API service is used. This is a free service. It provides many libraries for translation. The final step is translating the text using the language translator. The language is translated with the help of translator. After that the input text and output text is generated in PDF format. G. R. Hemalakshmi, M. Sakthimanimala, J. Salai Ani Muthu (2017)[2]

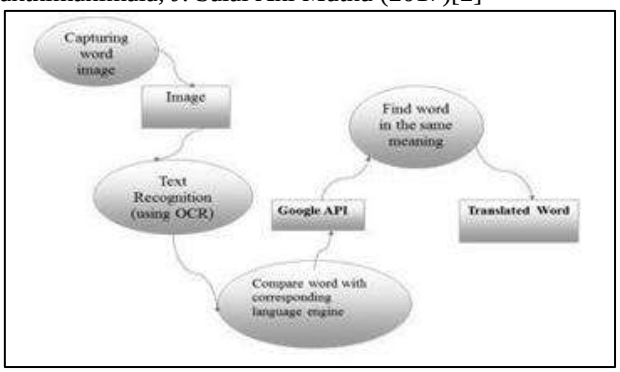

Fig. 2: Proposed system process

V. RESULT

With the help of python modules i.e. Python-tesseract, and Google API's for language translation a python script is created by which one image is taken as input and its characters are extracted and then this text is fetched to Google API's to translate this text into another language. This is further implemented as a web application in which we can browse any image and ask the application to convert the image text into the required language text.

VI. CONCLUSION

This paper is about optical character recognition techniques to translate the text from unknown language text into known language. The system has the capability to recognize characters with accuracy exceeding 85% mark. The advantage of this system is that it is easily portable and its scalability which can recognize six languages and also help in translating the text in English language. With the help of Python as the shadow of the application, we see how powerful it is in binding the latest technologies and trends to create something as founding. The main purpose of this application can be used by people all across the globe.

REFERENCES

[1] "Document Segmentation and Language Translation Using Tesseract-OCR" 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) by Sahl Thakare ,Ajay Kambale, Vishal Thengne, Mrs. U. R Kambale
[2] Extraction of Text from an Image and its Language Translation Using OCR , Volume 4, Issue 4, April 2017 ,by G. R. Hemalakshmi, M. Sakthimanimala, J. Salai Ani Muthu
[3] Novel Approach for Image Text Recognition and Translation Srinandan Komanduri, Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019),by Srinandan Komanduri, Y. Mohana Roopa, M. Madhu Bala.
[4] "Prototype Extraction and adaptive OCR" IEEE transaction on pattern analysis and machine intelligence, VOL. 21, NO. 12, December 1999, Yihong XU, Member, IEEE George nagi Senior Member, IEEE.
[5] "Extracting text from images using OCR on Android"- 27 June 2015.

