iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5157

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

OCR Text Detector and Audio Convertor

Himank Dave[1], Aryaman Gobse[2], Aryika Goel[3], Swati Bairagi[4]
[1, 2, 3]*Student, EXTC,* [4]*Assistant Professor, MPSTME, NMIMS University*

Abstract: ***In recent years, text recognition and extraction from different image documents and its conversion into audio is one of the foremost widely studied topics in Computer Vision, Image processing and Optical Character Recognition. Many complex text detection systems, like FAST and East algorithms, are used to infer text from a picture. During this article, a mixture of easy filters and detection systems are used. We also present a geometrical rectification framework which is crucial to revive the frontal-flat view of a document from one camera captured image. Our approach for geometric rectification is predicated on an estimation of the image homography matrix. We've incorporated a deep learning-based recognition engine, Tesseract, to extend the accuracy of our OCR system. Here, Tesseract is implementing an extended Short-Term Memory (LSTM) based recognition engine which may be a quite Recurrent Neural Network (RNN). Then we've converted our OCR text output into an Audio output using gTTS, a screen reader application developed to convert text into speech for our OCR system. This approach is incredibly fast. It is used to detect multilingual handwritten script and convert them into speech. The precision of this method is approximately 85% or higher. Last, this approach leads to uncomplicated and accurate text detection from document images and converts them into speech.***

Keywords-OCR (Optical Character Recognition), geometric rectification, image homography, LSTM (Long Short-Term Memory), gTTS, image documents, handwritten documents.

I. INTRODUCTION

Optical Character Recognition emanates from technologies involving telegraphy and creating reading devices for the blind. OCR is defined as the electronic or mechanical conversion of handwritten script or printed text images into machine-encoded text. Over the years, OCR has gone under many changes that supported its efficiency to perform and accuracy to induce promising outputs. To realize this many techniques and algorithms are used for text extraction in document images. It's observed while using a sophisticated complex approach like FAST and EAST text detectors on handwritten text, images fail to acknowledge the handwritten text and find you producing less efficient output. Here we'll use a very simple approach supported combination of filters, canny text detector, and contours. We've used a geometrical rectification framework for reinstating the frontal-flat view of a document from one camera captured image. Our approach for geometric rectification is predicated on an estimation of the image homography matrix. Firstly, a handwritten text image is given as an input, which if geometrically not aligned is rectified using a homography matrix whose output is given as input to our combination of filters and detectors for text detection. To extend the accuracy of our OCR system, we use Tesseract that contains a Deep Learning based recognition engine that converts our detected text from the image into machine-encoded text. Lastly, we convert this machine-encoded text into speech using gTTS. This approach is incredibly fast and often used to detect various languages, handwritten scripts and even images with a lot of noise.

II. PROPPSED METHOD

Accuracy of an OCR system is very plagued by the standard of noise removal and text detection. Ideally these systems should be highly accurate and fast. Geometric rectification, pre-processing, image detection, text extraction, text to speech conversion are the key sub-elements that outline the system's efficiency.

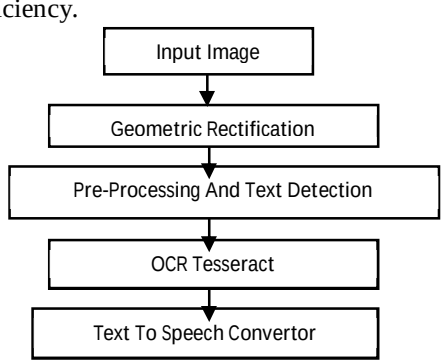


Fig1. Flow Chart Explaining the motion of our proposed OCR system

A. Geometric Rectification

Here we use a planar image homography approach to attain geometric rectified images. A Homography could be a transformation that maps the points in one image to the corresponding points within the other image. It's a linear geometric transformation between two spaces in N dimensions [2].

Planar homography, in computer vision is employed to model geometric transformations between two views. Mathematically it's expressed as:

$$x' = Hx$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

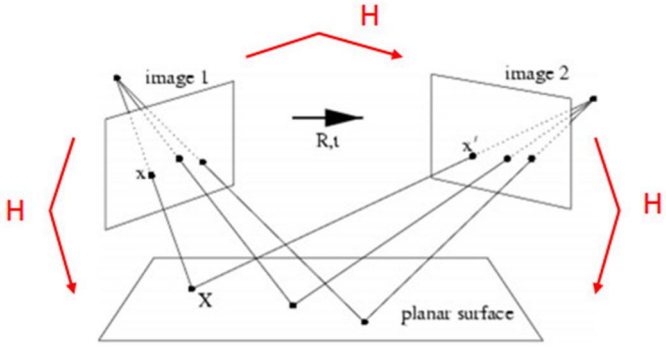


Fig2.General Diagram of Planar image homography

We have purposely selected this approach to reveal the likelihood of solving adjustment problems either using nonhomogeneous or using homogenous statistical procedures. To estimate and calculate the homography matrix H, we want some correspondences features like points, lines and curves. The planar homography matrix is said to the transformation between two planes, which may be seen as 2D projective transformation. It contains 9 elements, but it's defined up to a scale. This suggests that product of H and an arbitrary nonzero won't bring any change in the projective transformation. Therefore, the amount of degrees of freedom of such transformation is 8. Each corresponding 2D point or line generates two constraints and hence a minimum of 4 correspondences features are sufficient to compute the H matrix.[2]

Thus, the homography matrix could be a 3X3 matrix but with 8 DoF (degrees of freedom) because it is estimated up to a scale [3] [4]. It's generally normalized with

H_{33}=1 or $H_{11}^2 + H_{12}^2 + H_{13}^2 + H_{21}^2 + H_{22}^2 + H_{23}^2 + H_{31}^2 + H_{32}^2 + H_{33}^2$

Therefore, to revive the frontal-flat view of a document from one camera captured image. We estimated the image homography matrix by detecting unique features between the same images of various alignments; this is called the formation of matches. We then removed 'not so good matches' to extend the uniqueness of the image and plotted top matches using the estimated Homography matrix.

Fig3. (a). Example of Matching in Planar Image Homography

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Fig3. (b). Final Geometric rectified Image

A major challenge in image homography is perspective distortion. In our article we've corrected perspective distortion by creating an estimated homography matrix from two images of various alignments giving us an accuracy of 75-80%. But two more approaches may yield more accurate results which may be done by specifying the horizon line and parallel line for estimation of homography matrix [1] [2]. We also observed the presence of noise Deteriorates or damages within the end product thus reducing the accuracy of rectification.

B. Preprocessing and image detection

A captured image might contain noise. This noise can lead to less efficient OCR outputs. The OCR might even omit the region where the noise too much. To avoid these problems before processing, we use filters.

We have chosen a nonlinear median filter in our project. The reason we did not select a linear filter was that it tends to blur highly defined images.

Median filters are used to reduce the intensity variation between two pixels. We use the simple mathematics method of calculating the median. The pixels are first arranged in an ascending order. If the number of pixels is odd, the middle value is chosen to replace the pixel being calculated. Whereas if the number of pixels is even, the average of the middle two values is chosen for replacement.

$$\begin{bmatrix} 3 & 5 & 8 & 10 & 11 \\ 5 & 3 & 5 & 11 & 10 \\ 8 & 6 & 60 & 3 & 5 \\ 20 & 22 & 3 & 5 & 7 \\ 11 & 14 & 9 & 8 & 9 \end{bmatrix}$$

Now considering our Ax,y=60 Xmed calculated=11.

Median (Central value 60 is replaced by 11).[6][7]

We have used a canny edge detector for its main purpose of recognizing the edges of different characteristics.

We improvise the canny detector with the help of thresholding which reciprocally gives us highly efficient results. This approach proves to be detecting text from real lifetime images with more high accuracy than other simple text detectors. In our approach, with help of filtering framework we've removed image noise and use canny text detectors and have two output images one is canny and another is a mask or a binarized image this is done to provide options to OCR Tesseract to extract text and convert into machine-encoded text. To increase efficiency, we used contours and mapper algorithms which helped us in mapping of text images.

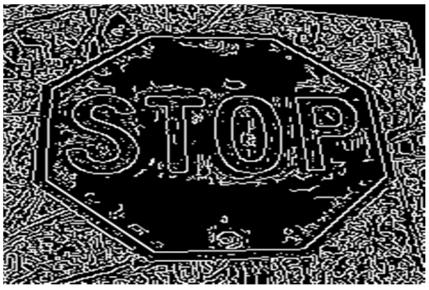


Fig4. (a). Canny detected Image with noise

Fig4. (b). Canny detected Image with Gaussian filter

Fig4. (c). Canny detected Image with framework of Gaussian, median and Wiener filter

Fig4. (d). Binarized Mask Image

C. OCR Tesseract

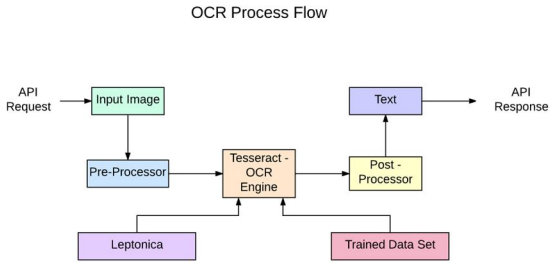


Fig5.Flowchart of OCR Tesseract [8]

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Optical Character Recognition (OCR) technology improved over the last ten years due to more advanced machine learning methods, more CPU power, and more elaborated algorithms. However, achieving 99% or higher OCR accuracy levels continues to be difficult. There are two measuring ways for the reliability of OCR. One is based on accuracy on a character level and the other are based on accuracy on a word level. Mostly it's preferred to measure the accuracy upon the character level; this relies on how often a character is recognized correctly versus how often a character is recognized incorrectly.

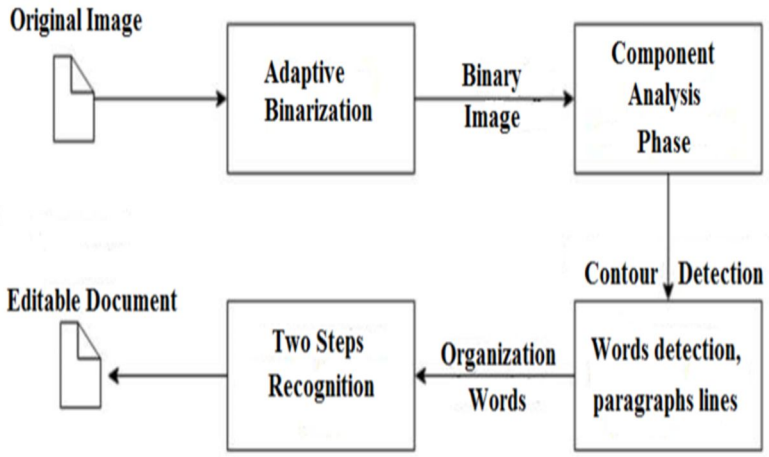


Fig6.Internal Architecture of OCR Tesseract [8]

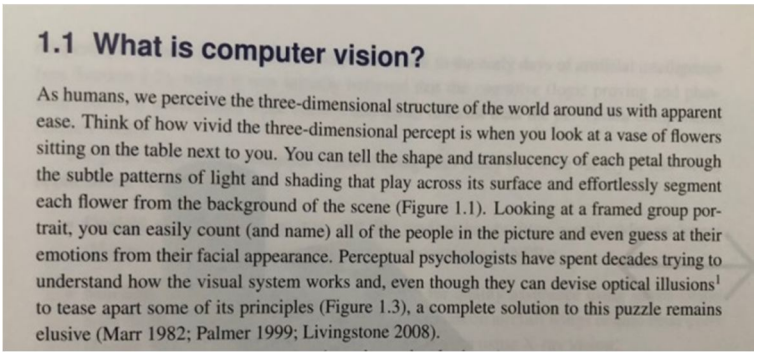
Version 4 has a dataset knowledge of hundred and sixteen languages.
Version 4 Tesseract has employed a Long Short-Term Memory (LSTM). LSTM is a Recurrent Neural Network (RNN).[8]
The language in our example is English. An option -1 is used by the Tesseract to specify the 'eng' language.
OCR Engine Mode (OEM); OEM has four functioning modes: (0) Legacy engine only, (1) Neural nets LSTM engine only, (2) Legacy and LSTM engines, (3) Default. We have selected mode (2). [9]
To change your page segmentation mode, change the --psm argument in your custom config string to any of the above-mentioned mode codes [9]
Tesseract works best when there's a clean segmentation of the foreground text from the background.[10] In practice, it is often extremely challenging to ensure these sorts of setup.
Even though Tesseract OCR is kind of powerful, but it also has limitations like it cannot comprehend the images with imprecise perspective and unclear backdrops. It also has problems when it comes to reading the order of the page. For example, it can fail to recognize a document that contains a table where it can join the text across the columns.
If the scanned output is of poor quality, it may produce OCR which is not of good quality either. Hence, it might fail to acknowledge the text's font family. That means it must be trained for various fonts, without any proper training there's a high probability it'll fail to acknowledge text given as input. If text contains language which isn't mentioned within the language arguments, it'll fail to read that text. However, overall OCR Tesseract proves to convey results with high accuracy. It is also proposed that with advance image processing OCR Tesseract can overcome above-mentioned limitations.

D. Text to Speech Conversion

We have used gTTs for this process. We've used it as a CLI tool and Python library to interface with Google Translates text-to-speech API [11]. Which then writes spoken mp3 data to a file, a file-like object (byte string). It features flexible pre-processing and tokenizing, also as automatic recuperation of supported languages.
gTTS is under constant change which suggests bugs that decrease its efficiency are fixed while new languages are added and making it more compatible for various environments.
In the end, after we have obtained a machine-encoded text from OCR Tesseract it is given as an input to gTTS application with the help of python libraries it converts machine-encoded text into speech. Audio output is extremely clear and may be converted into different languages depending upon the user.

III. RESULTS

Our approach is simple with precision and recall are over 80%.However, this approach can take some time in analyzing handwritten text this could occur due to lack of dataset training. It is also observed for not trained fonts system doesn't respond efficiently. Despite of these problems it is less complex and fast compared to other OCR text detection techniques and could be further improved in the future. Finally, this method turns out to be very efficient in extracting more complex layouts such as paragraphs.This method is able to identify and extract texts from latin text based languages that is most of European languages



1.1 What is computer vision?

As humans, we perceive the three-dimensional structure of the world around us with apparent ease. Think of how vivid the three-dimensional percept is when you look at a vase of flowers sitting on the table next to you. You can tell the shape and translucency of each petal through the subtle patterns of light and shading that play across its surface and effortlessly segment each flower from the background of the scene (Figure 1.1). Looking at a framed group portrait, you can easily count (and name) all of the people in the picture and even guess at their emotions from their facial appearance. Perceptual psychologists have spent decades trying to understand how the visual system works and, even though they can devise optical illusions[1] to tease apart some of its principles (Figure 1.3), a complete solution to this puzzle remains elusive (Marr 1982; Palmer 1999; Livingstone 2008).

Fig9(a). Input English Paragraph

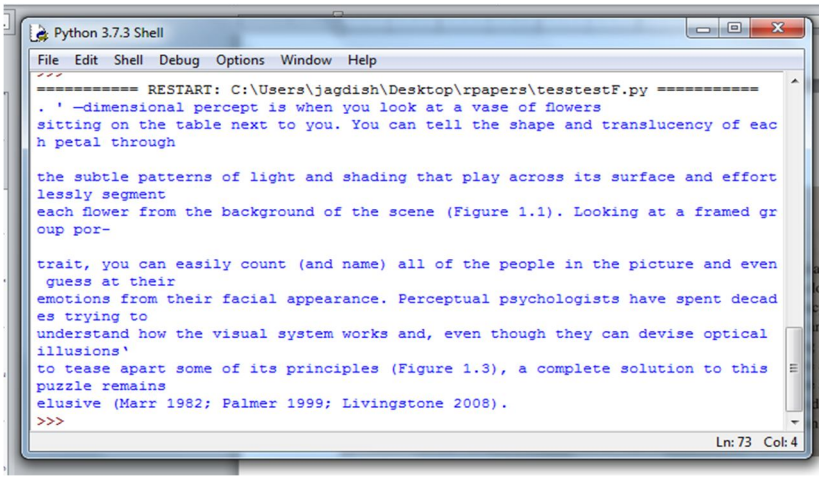
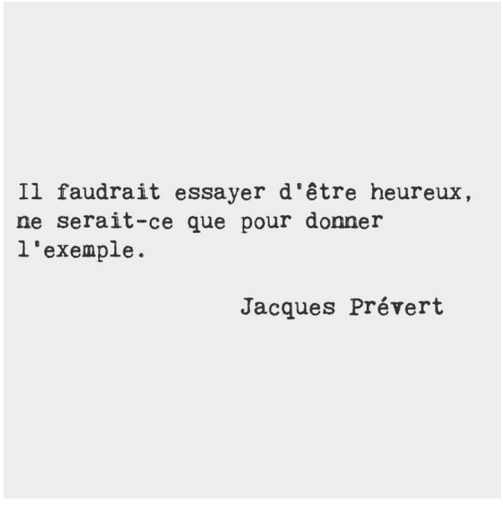


Fig9(b). Output English Text



Il faudrait essayer d'être heureux,
ne serait-ce que pour donner
l'exemple.

Jacques Prévert

Fig10(a). Input French Text Image

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

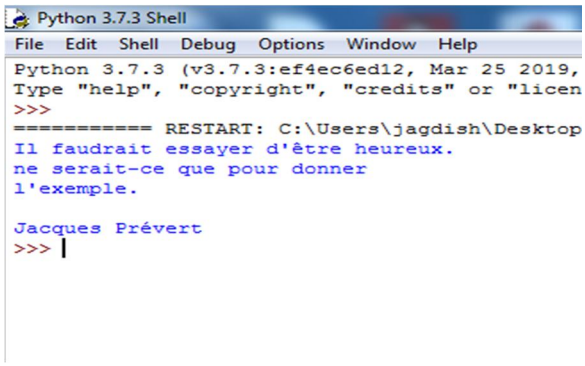


Fig10(b). Output French Text

Fig12(a).Application Oriented Receipt Input

Fig12(b).Application Oriented Receipt Output Text

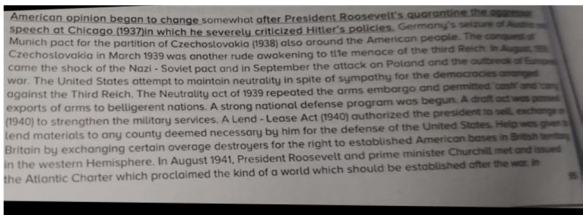


Fig12(a).Geometric Rectified image input

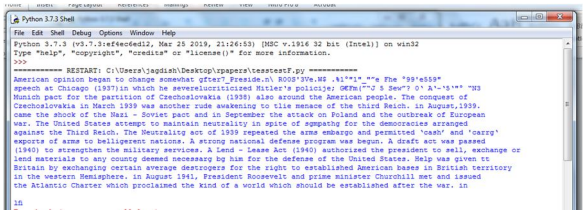


Fig12(b).Geometric Rectified text output

IV. CONCLUSION

In today's world OCR is one of the most highly used applications and plays an important role in the detection of physical documents. On analyzing camera-based documents from mobile applications, it is clear distortion introduced by incorrect planar and perspective projection plays a major challenge. We are solving this problem with an automatic rectification approach by creating an estimated homography matrix from two images of different alignment and plotting top matches to increase the uniqueness of the image. We then with help of framework we developed by a combination of different filters like Gaussian, threshold and median filter remove unwanted noise from the geometric rectified image which in return increases the accuracy of image detection. to extract text from filtered images we use a very simple algorithm that works on the combination of canny and mask text detector using principles of contour mapping. This helps in an increase in the accuracy of Tesseract OCR. Tesseract OCR then converts text detected into machine-encoded text using a Long Short-Term Memory (LSTM) based recognition engine which is a kind of Recurrent Neural Network (RNN). In the end we convert machine-encoded text into speech using gTTS.

This method is fast and less complex that it can be used to detect different languages, different fonts of handwriting and gives highly accurate results also in the presence of noise. It is observed due to the high accuracy of the sub-processes in our approach our overall average accuracy of OCR system has increased from 80 to 85%.

V. FUTURE SCOPE

The OCR developed by us in this project is still at a very beta phase and can be improved on multiple counts and basis. The utilization of an OCR is already tremendous and can be increased by implementing new features or overcoming some of the limitations in our current system. Some of the future work that we think may be implemented in our system is as follows: -

A. An Ocr That Can Take Multiple Font Input

The OCR developed in our project is sensitive to font style. With more font data we can train the system to recognize various peculiar fonts and cursive scripts. Even during scanning pictures taken in real time of different objects, fonts differ in various sizes. This may cause an error in the accuracy of the output. Thus, a wider dataset of fonts can help us train our system and make it better for the future.

B. OCR that can Support Multiple Languages

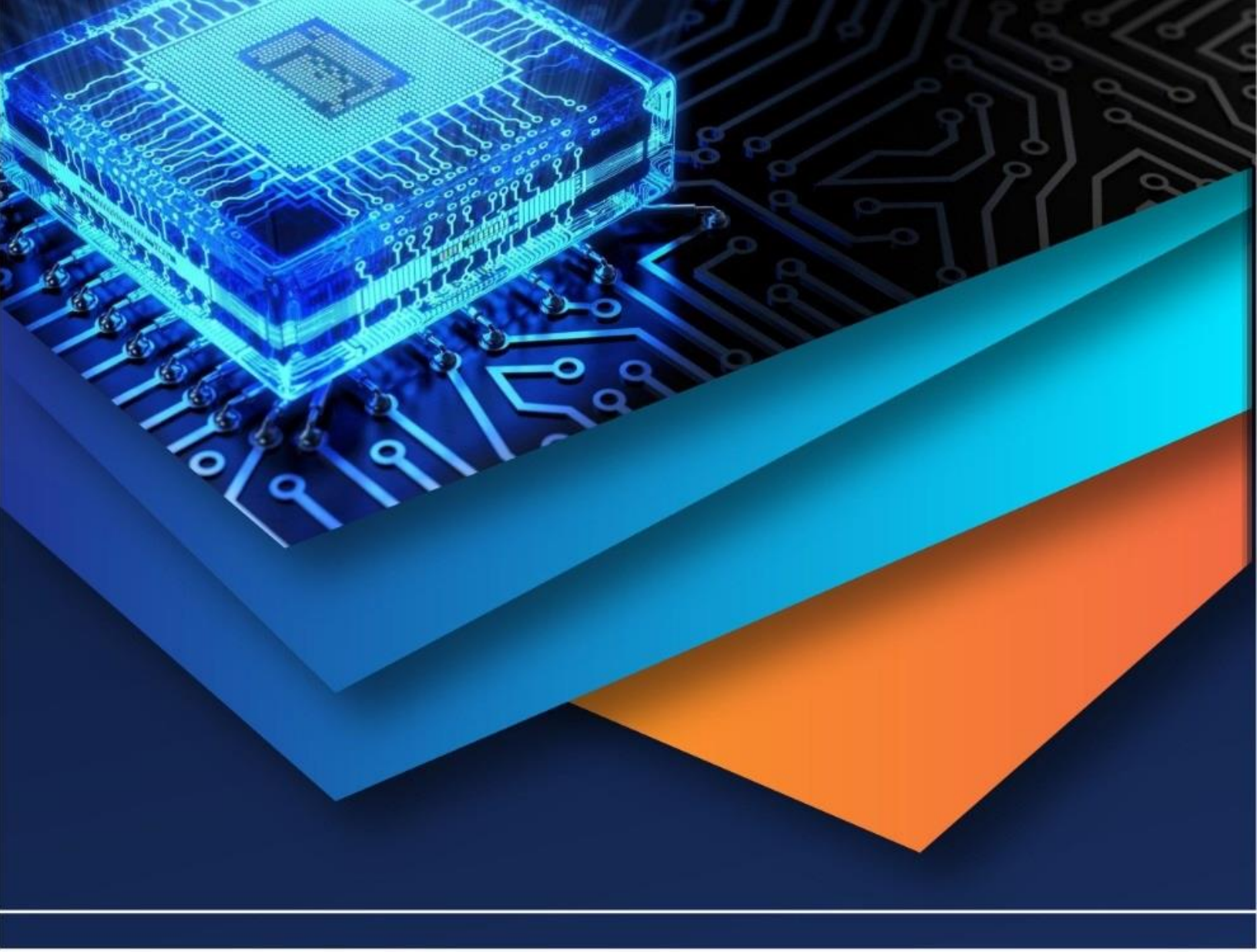
OCR is a globally useful system. In our project we have made an OCR that has the capability to recognize only European languages. OCR's can be made more useful by broadening its functions like recognition of texts from multiple languages. We can also train the OCR such that it can recognize multiple languages in a single document too, so that we don't need to segment the texts from the document and give separate inputs while feeding them into the system.

C. OCR as a Translation Device

Once we overcome the multiple languages barrier, we can implement a system in the future such that the OCR can translate from one language to the other with the help of a transliteration or translation system.

REFERENCES

[1] Jian Liang, Daniel DeMenthon, David Doermann, "Geometric Rectification of Camera-Captured Document Images", 2008 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Computing and Engineering, April 2008
[2] Bensoukehal Ali, Meche Abdelkrim,Keche Mokhtar, "Perspective Rectification Using Homography Planar: Plane Measuring", 2015,International Conference on Control System, Computing and Engineering, 2015.
[3] Sebastien Lefevre , Devis Tuia , Jan Dirk Wegner , Timothee Produit , Ahmed Samy Nassar. " Toward Seamless Multiview Scene Analysis From Satellite to Street Level", Proceedings of the IEEE , 2017
[4] R. n. Hartley, "nn defense of the eight-point algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 6, 1997.
[5] A. Zisserman, "Multiple view geometry in computer vision." , Cambridge university press, 2003.
[6] Smriti Srivastava , Sugandha Agarawal , O.P Singh. "A hybrid algorithm using transform for de-noising celestial images" , 2017 IC3TSN, 2017
[7] Image Processing Course Project: Image Filtering with Wiener Filter and Median Filter by Le-Anh 2019/04/24
[8] Ray Smith, "An Overview of the Tesseract OCR Engine", 2007 IEEE
[9] Ekraam Sabir, Stephen Rawls, Prem Natarajan, "Implicit Language Model in LSTM for OCR", 2018 IEEE International Conference on Control System, Computing and Engineering, 23 May 2018.
[10] A. S imran , S. Chanda, F.A Cheikh,K Frank , U Pal "Cursive handwritten segmentation and recognition for instructional videos",2012 Eigth international confernce on Signal Image Technology and Internet Based Systems,2012
[11] Pierre-Nick Durette, "gTTS Documentation",2020 Google Conference,26 Jan.2020



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY