iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XI **Month of publication:** November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47498

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: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

Text Recognition from Images

Safa Umatia[1], Aditi Varma[2], Aman Syed[3], Khushi Tiwari[4], Asst. Prof. Ms. Foram Shah[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, Thakur College of Engineering &Technology, Mumbai,*

Abstract: ***The method of extracting text from photographs is crucial in the current environment. The amount of information being stored in digital form instead than on paper has greatly increased in recent years. This makes it easier to store information and allows for easy retrieval of that information as needed. Pre-processing, segmentation, feature extraction, classification, and post processing are some of the stages in text recognition. The most important action in the pre-processing step is the conversion of a colour image into a binary image, which separates the text from the backdrop. Character separation is aided by the segmentation process. The most important data from the image can be extracted using features to help in text recognition. The classification method makes it possible to locate the text in accordance with clearly stated guidelines. After that, post processing is carried out to minimise errors. Many applications depend heavily on text recognition. This paper provides the different uses of text recognition from photographs as well as a discussion of the text recognition module. This paper also examines related literature.***

Keywords: Tesseract, OpenCV, Text recognition.

I. INTRODUCTION

This study looks for a method that may be used to extract any data from printed bills and invoices that are used every day. Later on, the information from these bills or invoices can be extensively utilised, for example, in statistical or machine learning applications. This study focuses on the extraction of final bill-amount, itinerary, date, and related information from bills and invoices because they include a wealth of information about the users' purchases, preferences, and other activities. With the use of optical character recognition (OCR) technology, printed or handwritten characters can be fully recognised from photographs. In the beginning, OpenCV was used to extract the bill or invoice from the image and remove any extraneous noise. After that, the intermediate image is processed further.

Initially, OpenCV was used to detect invoices or statements from images and remove unwanted noise from images. The intermediate images are then further processed by the Tesseract OCR engine, an optical character recognition engine. Tesseract will apply text segmentation to extract texts written in different fonts and languages.

Our methodology has proven to be highly accurate while being tested on various input images of invoices and bills. Text in Image Recognition automatically reads text from images. A field of research that seeks to develop functional computer systems. Today, there is a tremendous need to store information available in paper document form on computer disks and later reuse that information in the search process. An easy way to save the information in these paper documents to a computer system is to first scan the documents and then save them as images. However, it is very difficult to read the individual contents and search the contents of these documents line by line and word by word in order to reuse this information.

II. METHODOLOGY

Methodology:

OCR (Optical Character Recognition) is a machine that helps to recognize and obtain full printed or handwritten character-by-character alphanumeric recognition from text-based images or scanned documents. Tesseract is open Source and popular optical character recognition engine to this day. The basic processing by Tesseract is a -by-step pipeline -methodology in which connected component analysis is performed. Outlines are analysed, filtered, and blobs Collected. Blobs are processed and organized into lines of text. The process then continues by splitting the words in the lines of text. The first pass of the recognition process attempts to recognize each word in turn. Words satisfying the constraints are passed to the adaptive trainer. The Adaptive Trainer learns from the first pass and uses them in the second pass to try again to recognize words that did not satisfy all the constraints of the first pass. Fixed blurry spaces and checked small caps text. Digital text is then output.

Text image segmentation is typically used to identify objects and boundaries within an image. Text segmentation, as the term suggests, is the process of segmenting scanned text into useful elements such as words, phrases, and topics.

©IJRASET: All Rights are Reserved

Text segmentation was a prerequisite for information retrieval. The Tesseract OCR engine suffers from poor output quality if the input image is too noisy or may contain unwanted objects. Therefore, images should be pre-processed before passing to Tesseract OCR. Therefore, for better quality and more accurate output, OpenCV tool was considered for this purpose to remove unwanted noise and objects from the background.

A. OpenCV

(Open-Source Computer Vision Library) is library of programming functions intended primarily for real-time computer vision. First, an edge detection algorithm is applied to the image, then contour tracking is performed. All contours present in the image, 4-point contours are searched. The largest 4-point contour is the desired object from the photo, which is an invoice or calculation from the photo. Then the detected objects are cut out from the original image and some filters are applied. The following tools were used during the current research work.

B. Flask

Flask is a web application framework written in Python. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. Flask is considered a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Flask is commonly used with MongoDB, which provides a non-relational database.

C. Tesseract OCR

Tesseract finds templates by pixels, letters, words, and sentences. It uses a two-step approach called adaptive detection. A data phase for character recognition is required[l]. A second phase is then required to match characters that are likely to correspond to the context of the word or sentence, and characters that are not guaranteed.

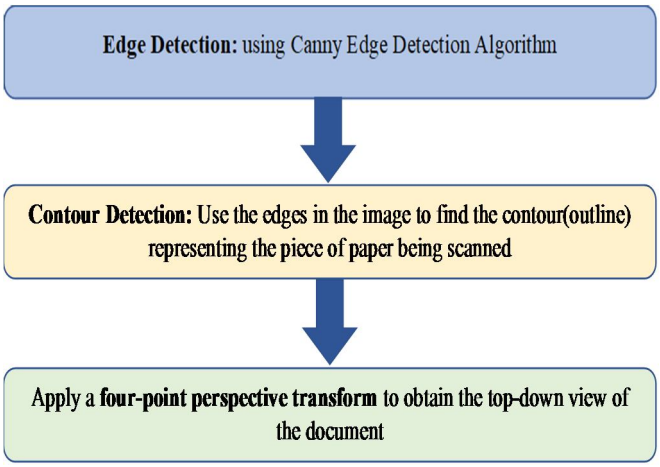


Figure 1: The methodology used in our research work

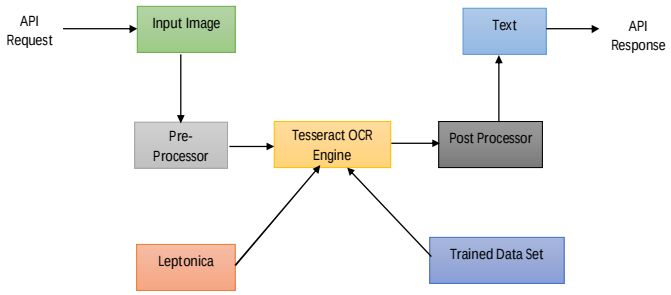


Figure 2: OCR Process Flow

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

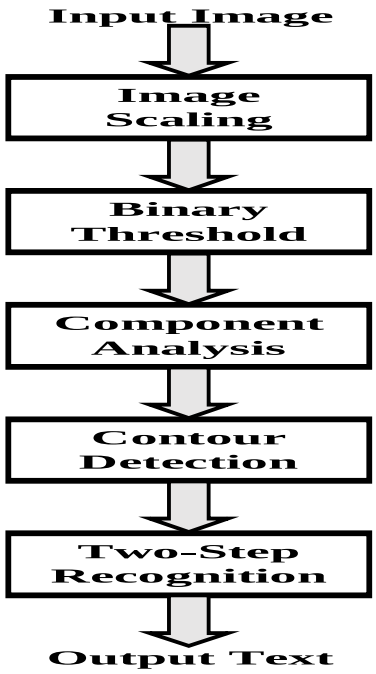


Figure 3: Block Diagram

A document is created using OpenCV technology in this project. It extracts the invoice from the user's invoice image, which may have background noise, and provides only the invoice or the transformed image of the invoice, which can improve the accuracy of text extraction. A printed invoice is a rectangular piece of paper, so you need to find the largest piece of paper. The largest rectangular object in the image, which will be the hill in the image. Consider the example image shown in Figure 1.

Canny Edge Detector is the best edge detection algorithm. As the name suggests, it is used to detect edges in images. Detect various edges in an image using a multilevel algorithm. Accuracy and simplicity of the underlying algorithms for performing line, edge or contour detection. Also known as the best detector. The main goal of this algorithm is to satisfy three main criteria:

Low error rate:

This means that only existing edges are properly detected.

Good localization:

The distance between detected edge pixels and true edge pixels should be minimized.

Minimal response:

Only one detector response per edge. Sample image output after Canny Edge detection. Contour detection is finding the contours of an image. For detecting contours, OpenCV provides a ready-made function called findcontours( ). The task of this function is to help detect contours in the provided image.

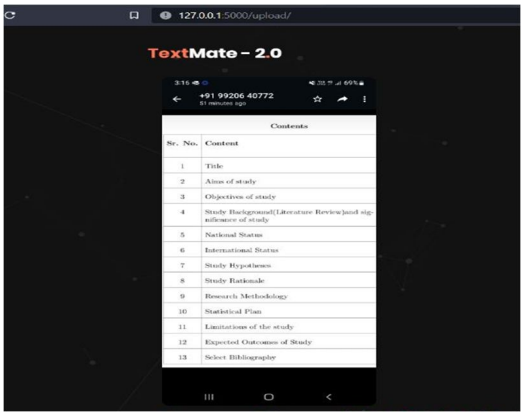


Figure 4: Input image

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

To get good output quality, the image should be processed. Since our goal is to recognize banknotes from images, we assume that banknotes are documents and sometimes rectangular objects. So exactly he can easily assume that the largest contour in the image with 4 points is ours or an invoice to scan.

1) Firstly, we need to work with grayscale images, so we convert the input image to gray.
2) A simple threshold is applied to obtain approximate contours. OPEN and CLOSE tricks are used to get smoother contours
3) A list of contours can be obtained using the FindCountour() function.
4) Sorts the retrieved contour list in non-ascending order.
5) Iterate through the sorted list of contours and see which contours have 4 points.
6) As if we were working, once we finally found the contours of the 4 points drawn in color

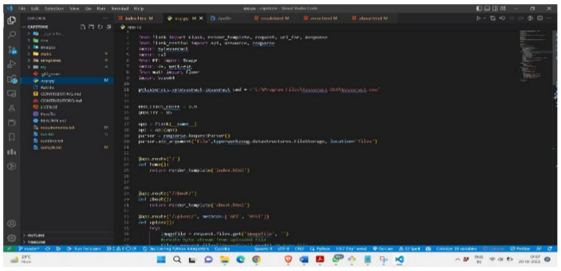
The final step in recognizing the banknote image is to crop the image along her four edges of the banknote to get the top view of the image. In addition to step 3, Adaptive Threshold Hold can be used to turn the image into a clean black and white image and give the document a paper effect.

a) *Text Extraction:* Text Extraction and Optical Character is done via the help of optical character recognition engine Tesseract and OCR. The processed image is passed to the Tesseract OCR engine and the engine helps to recognize text or extract the text from the image The engine The steps of the process of engine is demonstrated in the flow diagram as in Figure 2. The initial step is to analyze the connected elements for that the outline of the elements are saved. Though it was a costly step but it resulted in huge gain, the analysis of the outlines of the encapsulated elements, the text was identified in a much simpler way as dark element over light one., and segmentation process is applied here. Text Segmentation is the process of partitioning a digital image into multiple segments (sets of pixels). Segmentation is used for text- based Images, aimed in retrieval of specific information from the entire image.

Levels of segmentation process are as:

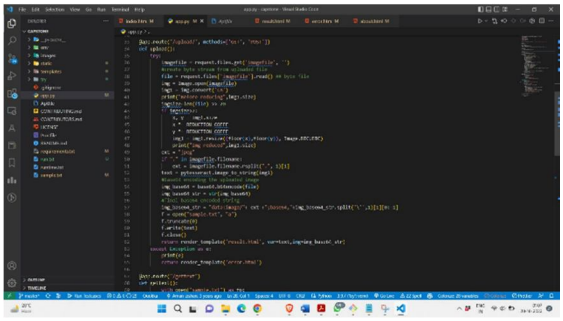
- Line Segmentation - The initial step for image segmentation is line segmentation, in which the horizontal scan of the image is done by the engine. A. Word Segmentation - The second step after horizontal scanning is vertical scanning, called word segmentation.
- Character Segmentation - Character segmentation is the final level of text-based image segmentation. This includes vertical scanning of the image. Segmentation here is basically the process of breaking the text into a series of words, then into letters, then into individual segments. To do this, we need to recognize the final signs and perform meaningful analysis. Each word determined to be meaningful is passed to the adaptive classifier as training data. The basic function of adaptive classifiers is to detect the lower side more accurately and get much better results. The adaptive classifier might discover something unique on the top page that it didn't notice before, so it makes his second check across the page for similar segmentation. Once this is done, a final check is performed to recheck for ambiguous spaces and small caps.

b) *Implementation:* Both invoice recognition and text extraction methods were implemented and tested on the Python 3.5 platform using the Python bindings for each of the above tools. The latest versions of both tools were used as OpenCV 3.3-0.10 and the latest stable version of Tesseract-OCR 3.05.01. Released on July 1, 2017. I used the open source, Toll Flask., a micro-framework for Jinja 2, a Python-based tool used to develop web apps in Python. I created a Flask app for an app that can run the tasks defined above and return a result as True or False when using the Android live web server, so for running a specific run task in an Android app can be used to test our survey, we developed an Android app that prompts users to upload an invoice or a photo of an invoice with their phone's camera. Alternatively, users can upload photos from their gallery and upload them to the web server hosting your Flask app. Perform the above tasks. Along with the photo, the user also submits the total amount paid on the invoice. I extracted the total amount from the image and matched it with the data entered by the user. If the data matched, it returned true as a response and vice versa.

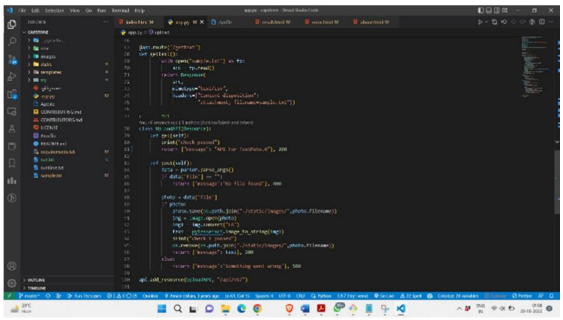


©IJRASET: All Rights are Reserved

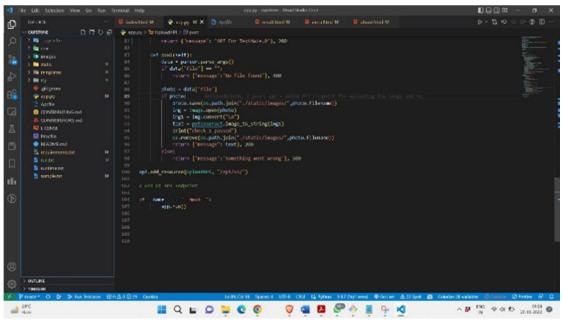
Importing libraries and initializing methods.



Here, method for uploading the image and bringing it to the desired dimensions and quality is mentioned which is further required for modelling.



Here, the tesseract model is applied on the image using an API.



The produced output of text is displayed and the methods for it are mentioned.

III. RESULT AND DISCUSSION

The text recognition system was evaluated using a variety of scanned photos with various text styles. The outcomes were incredibly positive. The suggested system pre-processes the image to get rid of the noise. From the bit map picture representation, features are extracted. Less computation is required for feature extraction, training, and testing, which is another benefit. The suggested technique produced good results for photographs with handwritten text that was written in various styles, sizes, and alignments against a variety of backgrounds. Even if there is noise in either the backdrop or the characters of the image, it labels the majority of the characters correctly. It demonstrates that, except from handling symbols, our system is really good at compressing compared to other systems.

Several text detection software has become more well-known in recent years. The most significant one is text from the scanned paper being automatically detected. The text recognition approaches have received an additional boost from recent technical developments. Numerous fields can be automated thanks to text recognition. It can be used for image tagging, scene analysis, automated signature reading on check leaves, and automatic licence plate reading at toll booths.

In the offices of the health care sector, text recognition also enables automatic storage and access of massive documents. It makes it possible to create a database where the text can be quickly searched for and located. The developed databases are easily accessible and updatable, which reduces the amount of paper labour. The visually handicapped benefit from voice aid and automated text detection. Additionally, it can be used to identify the text in the videos.

The transportation systems can also be made smarter via automated text identification. For information extraction and passport verification in airports, it is also usable. For commercial papers, text detection also enables automatic data entry. A huge selection of books are also made online thanks to text identification, enabling knowledge sharing and preservation. By facilitating automatic reading of labels and numbers, text recognition also supports automation in various industries. With the continued development of technology, text recognition has found a wide range of uses in practically every industry.

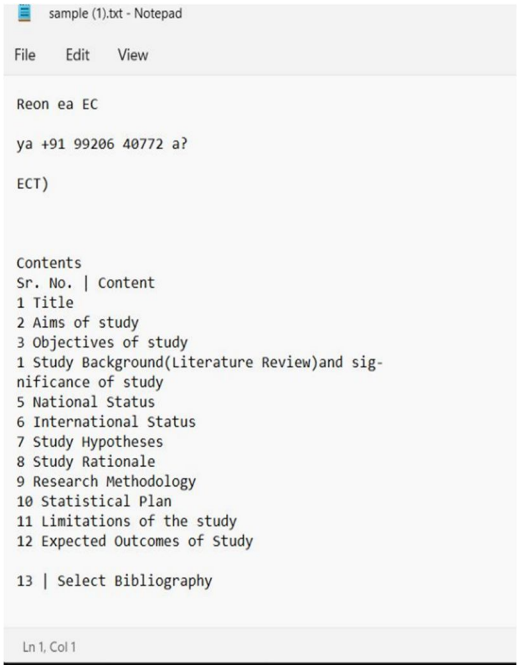


Figure 5: Recognised text from images

IV. CONCLUSION

Using a scalable feature learning approach, we developed a text detection and recognition system in this research and used it to analyse text images found in natural situations. We showed that, with larger feature banks, we can increase accuracy while maintaining peak performance that is competitive with other systems, emulating the outcomes seen in other computer vision and machine learning applications.

As a result, much research has concentrated on manually creating the models and features used in scene-text applications, our findings suggest that high performance may be achievable utilising a more automated and scalable approach.

The approaches followed here may achieve performance considerably beyond that provided by existing systems that mainly rely on hand-coded prior knowledge due to the development of more scalable and sophisticated feature learning algorithms by machine learning researchers.

Using a scalable feature learning approach, we developed a text detection and recognition system in this research and used it to analyse text images found in natural situations. We showed that, with larger feature banks, we can increase accuracy while maintaining peak performance that is competitive with other systems, emulating the outcomes seen in other computer vision and machine learning applications.

As a result, whereas much research has concentrated on manually creating the models and features used in scene-text applications, our findings suggest that high performance may be achievable utilising a more automated and scalable approach.

The approaches followed here may achieve performance considerably beyond that provided by existing systems that mainly rely on hand-coded prior knowledge due to the development of more scalable and sophisticated feature learning algorithms by machine learning researchers.

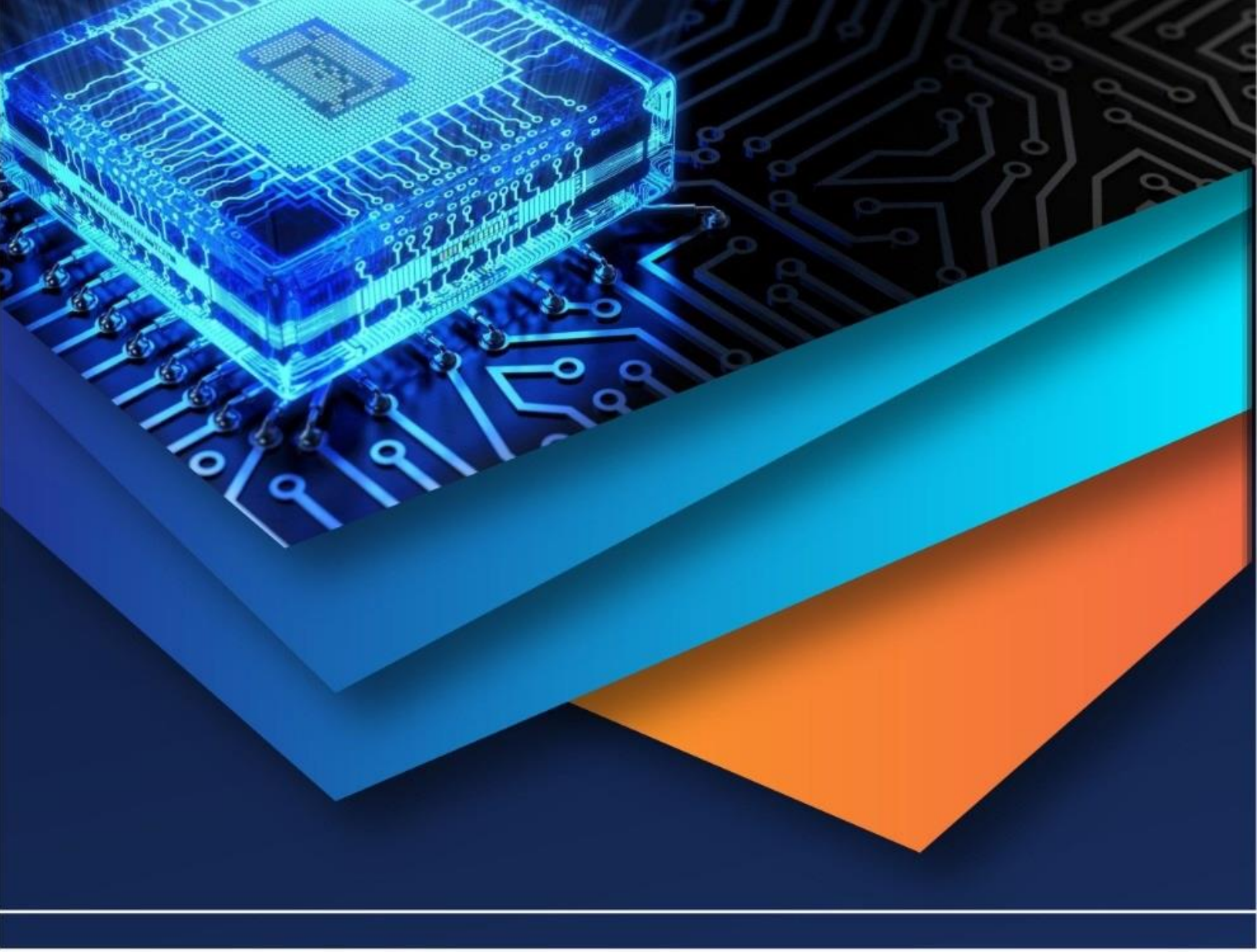
V. ACKNOWLEDGEMENT

I gratefully acknowledge the support, guidance, and encouragement of my Dissertation Guide Assistant Professor Ms. Foram Shah for this novel work.

REFERENCES

[1] Matteo Brisinello, Ratko Grbi, Dejan Stefanovi and Robert Pekai- Kova, Optical Character Recognition on images with colorful background, 2018, IEEE 8th International Conference on Consumer Electronics – Berlin (ICCE-Berlin).
[2] M.S. Akopyan, O.V. Belyaeva, T.P. Plechov and D.Y. Turdakov, Text recognition on images from social media, 2019, Ivannikov Memorial Workshop (IVMEM).
[3] Neha Agrawal, Arashdeep Kaur, An Algorithmic Approach for Text Recognition from Printed/Typed Text Images, 2018, 8th International Conference on Cloud Computing, Data Science & Engineering.
[4] K. Karthick, K.B. Ravindrakumar, R. Francis, S. Ilankannan, Steps Involved in Text Recognition and Recent Research in OCR; A Study, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.
[5] Anupriya Shrivastava, Amudha J., Deepa Gupta, Kshitij Sharma, Deep Learning Model for Text Recognition in Images, 10th ICCCNT 2019 July 6-8, 2019, IIT – Kanpur, Kanpur, India.
[6] Pratik Madhukar Manwatkar, Dr. Kavita R. Singh, A Technical Review on Text Recognition from Images, IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)