iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34470

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 9 Issue V May 2021- Available at www.ijraset.com

Artificial Vision for Visually Impaired using YOLO Detection

Prof. Ashwini Yerlekar[1], Prachi Gadhe[2], Adesh Adhau[3], Aditya Kubde[4], Aditya Khadatkar[5], Vishal Shende[6]
[1, 2, 3, 4, 5, 6]*Rajiv Gandhi College of Engineering and Research, Department of Computer Science and Engineering, Nagpur, Maharashtra, India*

Abstract: ***One of the biggest challenges faced by complete blind or visually impaired people are their day-to-day physical movements, especially when they encounter some unfamiliar environment. They pose great difficulty in navigating in crowded places or while traveling. For most visually impaired people, having a friend or a family member for as- assistance in navigation is more of a necessity rather than an option. To solve these problems, we have proposed a system that agglomerate technologies like Image processing, Speech Processing, Object Detection, Computer Vision, etc, to help these visually impaired people in navigation with no human interaction and reduce their difficulties to a certain extent. Our system uses the extremely fast YOLOv3(You Only Look Once-version 3) detection algorithm, which uses Darknet neural network framework, for the Object detection and Google's Text-to-Speech API(GTTS) for text-to-speech conversion of the detected object.***

Index Terms: Blind Navigation, Computer Vision, Speech Processing, Object Detection, YOLO Detection Model, Image Processing

I. INTRODUCTION

Out of the 285 million blind and visually impaired people in the world, 62 million of them are in India. Seeing the numbers, it is evident that India has the world's largest number of blind and visually impaired people. Of the 62 million visually impaired people, 54 million have low vision and 8 million of them are blind. Considering the above scenario, currently, India needs as high as 2.5 lakh of eye donation every year to cater to the needs of blind people and make them able to see the world. But the catch is, India can gather only 50,000 eye donations every year which is a mere 20% of the actual donations. Also, a lot of people cannot bear the cost of the medical treatment for eye surgery. That being said, new advancement in Computer Vision and Image processing enables the computer to process the image and detect the objects such as a person, table, chair, etc, with great accuracy. Keeping this in mind, we are presenting this paper to aid the visually impaired and reduce their day-to-day difficulties. The proposed system focuses on assisting blind people or people with low vision to navigate from one place to another without even worrying about anything present in the surrounding. Visually impaired people rely on other sensory information to avoid any hurdles or obstacles that come their way. One of the senses used by them is called "Haptics" which means anything related to the sense of touch. These also produce some noise feedback which their auditory senses use to know the approximate location of the obstacle(object). The first thing that comes to everyone's mind after reading this is, "white cane", which is a device commonly used by visually impaired people in navigation. Though it is a good device and caters to the need for navigation, due to the small and restrictive length of the cane, the blind person can detect obstacles in only proximity.

That being said, a lot of obstacle detection systems have been introduced in the recent decade that can cater to the needs of navigation of visually impaired people to some extent. The two categories of navigation that may play a significant role in detection system are:

1) *Vision Enhancement:* The information, in this case, is being put visually.
2) *Vision Substitution:* This provides information of the surrounding in the form of audio feedback, or tactile feedback, or both.

We propose a system that is based on the "Vision Substitution" category of navigation.

Our system mainly consists of the following components:

a) Webcam
b) Audio device(Earphones)

Our entire system is based upon the above two components where the webcam receives the real-time-video for detection which works as an input device to our system and audio device, or earphones in this context, which gives the real-time-audio feedback of detected objects which is an output produced by the system.

II. LITERATURE SURVEY

A lot of object detection models have emerged in recent decades which use some of the others like Fast-RCNN etc, out of which the YOLOv3 Object detection model proves to be promising in terms of speed and accuracy. The Darknet Neural Network architecture on which YOLOv3 is based is specially designed for object detection which is the main reason why YOLOv3 outperforms the other detection models.

A. YOLOv3 Object Detection

YOLOv3(You Only Look Once) object detection model is implemented using Keras and OpenCV deep learning libraries. It is an object classification system that sorts the images into groups where objects which contain similar characteristics are grouped while the others are neglected or handled explicitly based on requirements. YOLOv3 used a Convolution Neural Network(CNN) for object detection.

CNN's are classifiers based system that can find patterns in the processed input images which are represented as a structured array of data. YOLOv3 separates an image into a grid. Each grid cell predicts some number of boundary boxes (sometimes referred to as anchor boxes) around objects that score highly with the aforementioned predefined classes. Each boundary box has a respective confidence score of how accurate it assumes that prediction should be and detects only one object per bounding box. The boundary boxes are generated by clustering the dimensions of the ground truth boxes from the original to find the most common shapes and sizes. Region-based CNN(R-CNN) and Fast-RCNN work quite similar but the only catch is that YOLOv3 is designed to do classification and bounding box regression at the same time.

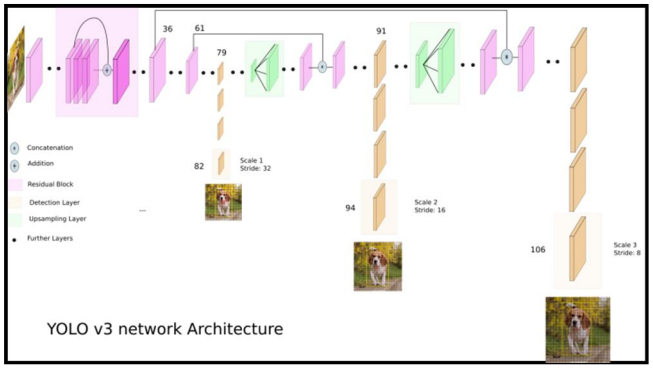


Fig. 1. YOLOv3 Network Architecture

B. Text-to-Speech API

Text-to-Speech(TTS) system converts normal language text into speech. A text-to-speech system is composed of mainly 2 parts a front-end part and a back-end part.

The front-end part has two major tasks to do:

1) *Text-Normalization:* It converts raw text containing symbols like numbers and abbreviations into equivalent of written-out words.
2) *Text-to-Phoneme Conversion:* The output form text normalization is then assigned phonetic transcriptions to each word and divides and marks text into units such as phrases, clauses, and sentences(prosody units).

Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. back-end part also called a synthesizer then converts this linguistic representation into sound., our system uses the state-of-the-art Google Text-to-Speech API(GTTS) as it provides playable audio. Google's Text-to-Speech API converts text input into audio data which in our case is mp3.

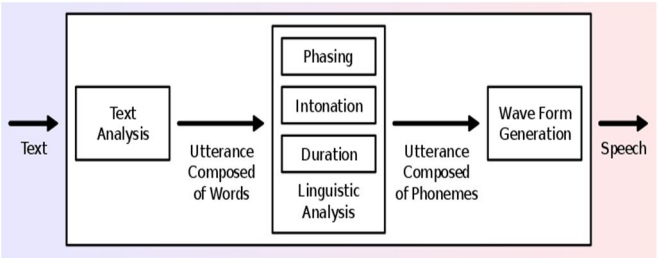


Fig. 2. Typical TTS System

C. Position Estimation

The position estimation system is one of the important modules of the system as it will tell us the position of the detected objects which in turn will be sent to the blind person through a voice feedback mechanism. The system here works based on coordinates, i.e., a Cartesian system where the coordinates of the position of the objects are calculated based on the height and width of the bounding boxes of the detected objects by the YOLO detection model.

The centers of the bounding boxes will be considered as the approximate position of the objects.

III. METHODOLOGY

Our system is mainly divided into three modules that communicate with each other to give the desired output. Following are modules that we introduced into our system:

A. Object Detection Module

The input to this module is through the webcam which provides a real-time video capture through the OpenCV library. The module is then run on each frame. For each input image obtained through the video capture, we are constructing a blob of the image and then perform a forward pass of the YOLOv3 object detector. The dataset which we used to detect the objects is called the *"Common Objects in Context(COCO)"* dataset. The above forward pass gives us the bounding boxes and the probabilities associated with each bounding box. The forward pass gives us a list of layered outputs. For each layer and each detection within a layer, a filtering process is carried out to remove all the weak detection, i.e., the detection with a confidence score(probability) of less than or equal to 50%.

The strong or valid detections are then stored where for each detection their confidence score, bounding boxes, and classes are stored individually. Based on the coordinates of the bounding boxes, i.e., x-coordinate, y-coordinate, height, and width, the centers of these bounding boxes are calculated. All these parameters are then sent to our next module, where they are used to calculate and estimate the position of each object detected in the bounding box.

B. Position Estimation Module

The input to this module is received through the Object Detection module which contains all the valid bounding boxes and confidence scores processed by the detection module. A *non-maxima suppression(NMS)* is performed on these bounding boxes in consideration of the confidence scores associated with each box.

The NMS process is an important part to filter out all the overlapping bounding boxes and also to suppress the weak ones. The suppression is done considering 2 factors:

1) *Score Threshold:* A threshold used to filter boxes by score. This is done to further increase the accuracy of the system if there is such a need. In our system, we have set this threshold as 0.5. This means that any detection if the overlapping boxes have a confidence score of less than 0.5 or if there are still any detection that has a confidence score less than 0.5 will be suppressed by the NMS function.
2) *NMS Threshold:* A threshold used in non-maxima suppression.

The above suppression returns the list of indexes of final bounding boxes of detection which are considered for position estimation. For each object that we detected accurately after suppression, we estimate the position. For each bounding box, the centers we calculated earlier are used. The frame is divided into 9 equal cells in form of a grid where each cell is given a unique name such as top-left, mid-center, top-center, bottom-center, bottom-left, etc. These estimated positions along with their classes are stored as text and sent to a further module, which is, text-to-speech.

C. Text-to-Speech Converter Module

The input to this module is the text obtained from the previous "position estimation module". The text is then passed to the text-to-speech API, which in our case, we are using Google's Text-to-Speech API(GTTS). The reason behind using this API is the availability of vast languages that it supports, plus the accuracy that the API provides.

The main working of GTTS is that it takes the input in raw format or as Speech Synthesis Markup Language(SSML) and outputs speech in the form of an audio file such as MP3 or LINEAR16. In our system, we are using the MP3 audio format to store the text. After the speech conversion is completed, the audio file is played in regular intervals. The intervals are nothing but the frame counts that we are keeping track of, the moment we start the video capture. In our case, we have set the intervals, i.e., frame counts to 30. This is done to minimize the frequency of speech output and to increase the efficiency of the system.

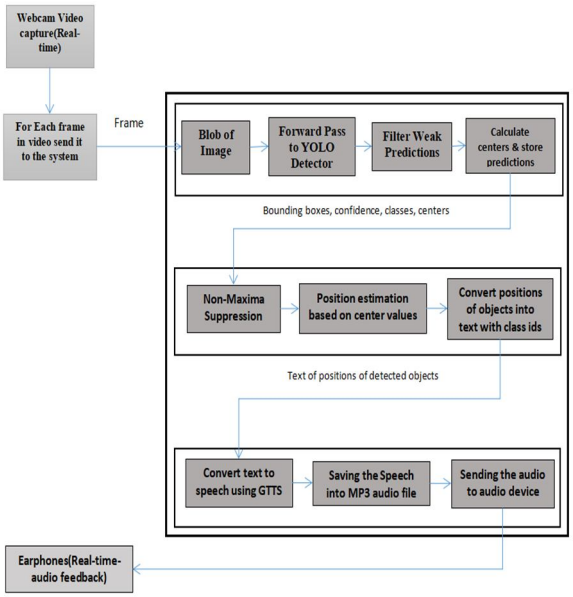


Fig. 3. System Architecture

IV. RESULTS

The system we build successfully recognizes the objects and also is relayed to the user using the audio devices in the form of speech to help blind people navigate easily to some extent. Multiple objects are detected which belong to different classes, thanks to the YOLOv3 detector and its pre-trained weights, which gives great accuracy in detecting objects. The system detects smoothly with 20-30 fps speed.

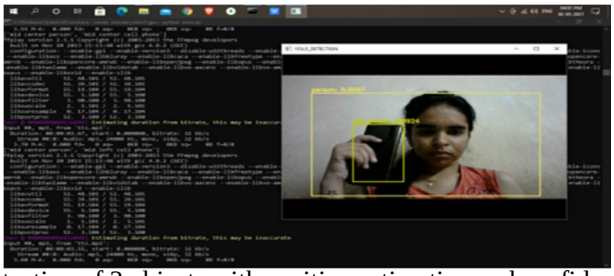


Fig 4.1. Detection of 2 objects with position estimation and confidence score

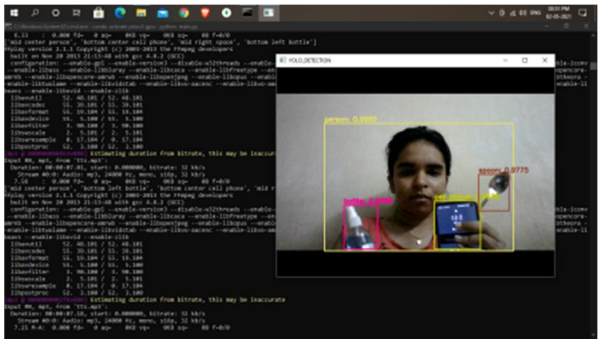


Fig. 4.2. Detection of 4 objects with position estimation and confidence score

Figure 4.1 and figure 4.2 illustrate the detection and recognition of multiple objects with accuracy in the range between 70% - 90%. A variety of different objects are detected all belonging to different classes. The detection is displayed after every 30 frame counts which doesn't take much time to do the computations. Also the overlapping sound feedback problem, i.e., an issue with continuous voice feedback without any delay has been rectified.

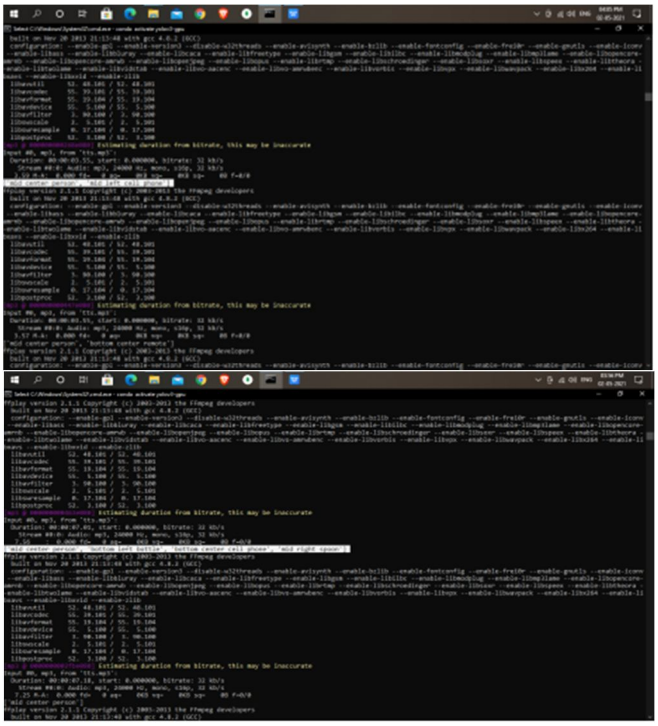


Fig 4.3. CLI output of the detected objects along with position

Figure 4.3 contains two screenshots of the output, where the first screenshot corresponds to the CLI output of figure 4.1 and the second screenshot corresponds to the CLI output of figure 4.2. The CLI output shows the detected object with their respective positions on the screen. For example, the person and the spoon are found in the "mid-center" of the screen, whereas, the bottle and the cellphone are found in the "bottom center" of the screen. The system can detect objects which are in any range between 2m and 7m. Ideally, for the navigation of blind people, the required range is between 20cm and 10m, as any object in this proximity is considered an obstacle. Though the Yolov3 detector can do accurate detection between 2-7 meters, it can be brought close to the ideal range by training the model on the images where the objects are scaled to close to the above ideal distances.

V. CONCLUSION

In this project, we researched and investigated the challenges faced by blind and visually impaired people and proposed a Computer Vision and Image Processing based system to reduce these challenges to some extent. This project provides a real-time solution for the blind visualization and navigation system. The accuracy and speed of the YOLOv3 detection model because of the underlying Darknet Neural Network Architecture proved to be a vital part of this system. That being said, there are a lot of improvements that can be done to this system. We can use a high-resolution camera which will in turn help in improving the obstacle detection in terms of accuracy of the position of obstacles. But doing this may increase processing time. This can be also be rectified by increasing the processing performance, which will, in turn, result in the processing of even higher frame resolution resulting in having a more stable object detection

VI. ACKNOWLEDGMENT

We would like to thank our guide Prof. Ashwini Yerlekar, for the constant support and guidance that she gave us which resulted in the successful execution of the project. We would also like to thank Prof. Neha Titarmare, for allowing us to do this project and also for her valuable suggestions.

REFERENCES

[1] Dunai, Larisa, et al. "Real-time assistance prototype—a new navigation aid for blind people." IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society. IEEE, 2010.
[2] Jabnoun, Hanen, Faouzi Benzarti, and Hamid Amiri. "Object detection and identification for blind people in video scene." 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA). IEEE, 2015.
[3] Kotalwar, Ganesh, et al. "Smart Navigation Application for Visually Challenged People in Indoor Premises." International conference on Computer Networks, Big data and IoT. Springer, Cham, 2018.
[4] Proposed System on Object Detection for Visually Impaired People International Journal of Information Technology (IJIT) – Volume 4 Issue 1, Mar-Apr 2018 Department of Computer Science and Engineering V.E.S.I.T Maharashtra - India

[5] Raihan, Md, and Hossain Mohammad Seym. "An Effective Navigation System Combining both Object Detection and Obstacle Detection Based on Depth Information for the Visually Impaired." Diss. Department of Computer Science and Engineering, Islamic University of Technology, Board Bazar, Gazipur, Bangladesh, 2018.
[6] Lee, Sanghyeon, and Moonsik Kang. "Object Detection System for the Blind with Voice Command and Guidance." IEIE Transactions on Smart Processing & Computing 8, no. 5 (2019): 373-379.
[7] Shewale, A., Mahakalkar, M., Pawar, V., Bharad, Y., & Chiwhane, S. (2019). Electronic Eye for Blind people with Object Detection and Audio Output Using Image Processing.
[8] Rana, S. M., Umme Mafruha Khanam, and Koushik Sarker Antu. "Moving Object Detection for Blind People." (2019).
[9] Kotalwar, Ganesh, et al. "Smart Navigation Application for Visually Challenged People in Indoor Premises." International conference on Computer Networks, Big data and IoT. Springer, Cham, 2018.
[10] Rahman, Ferdousi, et al. "An assistive model for visually impaired people using YOLO and MTCNN." Proceedings of the 3rd International Conference on Cryptography, Security, and Privacy. 2019.
[11] object Detection System for the Blind with Voice Command and Guidance IEEE Transactions on Smart Processing and Computing, vol. 8, no. 5, October 2019 Department of Electronic Engineering, Gangneung–Wonju National University Gangneung, Gangwon 25457, Korea.
[12] Darknet "Darknet.se - About darknet". 2010-08-12. Archived from the original on 2010-08-12. Retrieved 2019-11-05
[13] Shewale, A., Mahakalkar, M., Pawar, V., Bharad, Y., & Chiwhane, S. (2019). Electronic Eye for Blind people with Object Detection and Audio Output Using Image Processing.
[14] Real-Time Object Detection for Blind People Assistant Professor, Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode. (India)
[15] Vaidya, Sunit, Naisha Shah, Niti Shah, and Radha Shankarmani. "Real-Time Object Detection for Visually Challenged People." In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 311-316. IEEE, 2020.
[16] Joshi, Rakesh Chandra, Saumya Yadav, Malay Kishore Dutta, and Carlos M. Travieso-Gonzalez. "Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People." Entropy 22, no. 9 (2020): 941
[17] Abraham, Leo, Nikita Sara Mathew, Liza George, and Shebin Sam Sajan. "VISION-Wearable Speech Based Feedback System for the Visually Impaired using Computer Vision." In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), pp. 972-976. IEEE, 2020.
[18] Mahendru M, Dubey SK. Real-Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3. In2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2021 Jan 28 (pp. 734-740). IEEE.
[19] Mahendru, Mansi, and Sanjay Kumar Dubey. "Real-Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021
[20] Shah, Bela, et al. "Survey on Object Detection, Distance Estimation and Navigation Systems for Blind People." Machine Learning for Predictive Analysis. Springer, Singapore, 2021. 463-472.
[21] Annapoorani, A., Nerosha Senthil Kumar, and V. Vidhya. "Blind-Sight: Object Detection with Voice Feedback." (2021).
[22] Zaman, F. H. K., Abdullah, S. A. C., Razak, N. A., Johari, J., Pasya, I., & Kassim, K. A. A. (2021, February). Visual-Based Motorcycle Detection using You Only Look Once (YOLO) Deep Network. In IOP Conference Series: Materials Science and Engineering (Vol. 1051, No. 1, p. 012004). IOP Publishing.
[23] Shetty, A. K., Saha, I., Sanghvi, R. M., Save, S. A., & Patel, Y. J. (2021, April). A Review: Object Detection Models. In 2021 6th International Conference for Convergence in Technology (I2CT)
[24] Rachburee, N., & Punlumjeak, W. (2021). An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people. International Journal of Electrical & Computer Engineering (2088-8708), 11(4).
[25] Shah, Bela, Smeet Shah, Purvesh Shah, and Aneri Shah. "Survey on Object Detection, Distance Estimation and Navigation Systems for Blind People." In Machine Learning for Predictive Analysis, pp. 463-472. Springer, Singapore, 2021.

AUTHORS

First Author — Prof. Ashwini Yerlekar, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, ashwini.yerlekar@gmail.com

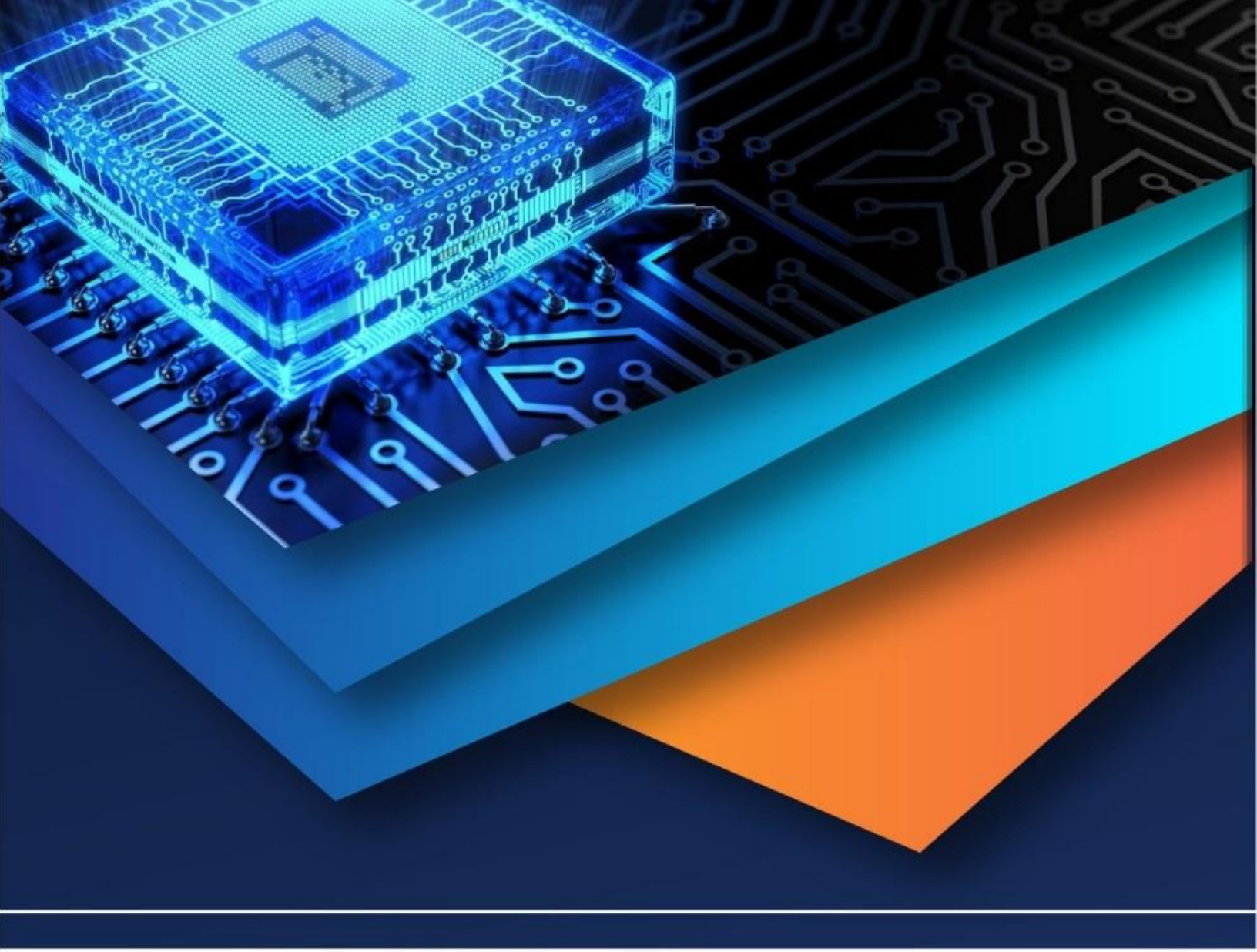
Second Author — Prachi Gadhe, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, prachigadhe21@gmail.com

Third Author — Adesh Adhau, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, adeshadhau3110@gmail.com

Fourth Author — Aditya Kubde, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, kubdeadi@gmail.com

Fifth Author — Aditya Khadatkar, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, apk26012000@gmail.com

Sixth Author — Vishal Shende, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Wanadongri-441110, vbshende122333@gmail.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY