iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50182

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 11 Issue IV Apr 2023- Available at www.ijraset.com

Object Detection with Voice Guidance to Assist Visually Impaired Using Yolov7

Dr. P. Boobalan[1], Bhuvanikha. S[2], Sivapriya. M[3], Sivakumar. R[4]
[1]*Associate Professor, Department of Information Technology, Puducherry Technological University, Pududcherry, India*
[2, 3, 4]*B.Tech Student, Department of Information Technology, Puducherry Technological University, Puducherry, India*

Abstract: ***Object recognition technology has revolutionized various domains such as autonomous vehicles, industrial facilities, and many more. However, the visually impaired individuals, who are most in need of this technology, have not been able to utilize it effectively. Therefore, this paper presents a deep learning-based object detection system that is specifically designed for the blind community. The system incorporates the YOLOV7 (You Only Look Once) algorithm for object recognition and uses text-to-speech (TTS) technology to provide a voice-guidance technique that conveys information about the objects around them. The main objective of the proposed system is to empower visually challenged individuals to independently identify objects in a particular space without relying on external assistance. The system's efficiency and performance are thoroughly scrutinized through experiments to validate its accuracy and effectiveness. The proposed system employs technical terms such as deep learning, object recognition, YOLOV7 algorithm, and also image classification techniques for feature identification and extraction of the video frames and categorize them into the respective classes. It uses COCO dataset, which consist of around 123,287 hand-labeled images classified into 80 categories, has been utilized to train the proposed system.***
In conclusion, the suggested system for object detection uses cutting-edge deep learning techniques along with voice guidance methods to assist individuals who are visually impaired in identifying objects independently. The performance of the system is evaluated through experiments, demonstrating its effectiveness and potential for real-world applications.
Index Terms: Object Detection, YoloV7, COCO dataset, TTS, Convolutional layers, Voice Feedback

I. INTRODUCTION

The rapid development of Information Technology (IT) has spurred numerous research endeavors aimed at resolving day-to-day inconveniences and providing conveniences for individuals. However, visually impaired individuals still experience numerous obstacles, with finding object information and navigating indoor spaces representing the most significant challenges. Object detection technology for visually impaired individuals is an important area of research aimed at improving their quality of life and promoting their independence. Object detection is crucial for visually impaired individuals as it helps them navigate their surroundings independently and safely. This can include the use of cameras and computer vision algorithms to detect objects and provide information about them, such as their name. Voice guidance technology is also incorporated to provide auditory feedback to the user, allowing them to navigate and interact with their environment more safely and independently.

This paper proposes a deep learning object recognition technique for obtaining accurate object information and determining object locations. The You Only Look Once (YOLO) architecture, a state-of-the-art object recognition deep learning model, is used to detect objects through a camera. This dissertation presents a system that analyses precise object information and location, utilizing deep learning techniques. To enhance the usability of the system for visually impaired individuals, voice guidance is provided through the gTTS (Google Text-to-Speech) Python library. This allows the system to synthesize spoken announcements of object detections in real-time, providing the user with immediate and relevant information about their surroundings. This voice guidance enables the visually impaired to navigate their environment with greater independence and safety, making the proposed system an effective tool for assisting them in daily life.

II. RELATED WORKS

The paper explores the recent advancements in object detection techniques utilizing region-based convolutional neural networks and region proposal methods. Although the initial cost of these methods was high, it has been lowered by sharing convolutions across proposals. Fast R-CNN, the latest method, accomplishes near real-time rates using very deep networks, but proposals remain the computational bottleneck in state-of-the-art detection systems. The paper introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, enabling nearly cost-free region proposals. RPNs are trained end-to-end to generate high-quality region proposals are used by Fast R-CNN for detection.

However, the paper has some constraints, such as the high computational expense that requires a powerful GPU to achieve real-time performance, and it is solely evaluated on the PASCAL VOC dataset, limiting its generalization to other datasets. Additionally, the proposed method relies on the selective search algorithm for generating region proposals, which may not be the most efficient or accurate approach for all applications. Lastly, the paper does not provide a comprehensive analysis of the learned features or the inner workings of the proposed method.

This paper have presented a system that leverages object recognition algorithms to detect multiple objects in intricate images and offer audio feedback to aid visually impaired individuals in navigating unknown surroundings. However, the study has limitations as the system's performance has only been evaluated using a webcam in different scenarios, leaving its efficacy in real-world environments uncertain. Furthermore, the paper has not addressed the potential obstacles or constraints of employing audio feedback as a means of assisting visually impaired individual.

The authors of the paper investigates the use of convolutional neural network models designed for detection in RGB images in the task of automatic person detection in thermal images. The authors retrained different state-of-the-art object detectors such as Faster R-CNN, SSD, Cascade R-CNN, and YOLOv3 on a small dataset of thermal images extracted from videos that simulate illegal movements around the border and in protected areas. They also experimented with various training dataset settings to determine the minimum number of images required to achieve good detection results on test datasets. The authors achieved excellent detection accuracy in all test scenarios, despite using a modest set of thermal images for training. Additionally, the paper introduces their original thermal dataset, containing surveillance videos captured in diverse weather and shooting conditions, for experimentation purposes. However, the paper has some limitations, such as the relatively small size of the thermal dataset used for experimentation, which may affect the generalizability of the results. The experiments were conducted using simulated videos, which may not fully capture the complexity and variability of real-world scenarios. Moreover, the paper solely focuses on person detection in thermal images and does not explore other potential applications of thermal imaging technology. Lastly, the paper lacks a detailed analysis of the computational requirements and efficiency of the various object detection models tested.

The focus of this paper is on the process of visual object tracking in videos, which involves detecting and linking moving objects over time. Various algorithms are available for video tracking, each with its own unique set of challenges, such as changes in orientation and rapid movement. The paper evaluates the tracking-by-detection approach using YOLO for detection and SORT algorithm for tracking, as well as developing a custom image dataset for specific classes. The authors also highlight the importance of detecting vehicles and pedestrians in videos for traffic analysis. It should be noted that the efficacy of the tracking-by-detection approach using YOLO and SORT algorithm may vary based on several factors, including the quality of the video input, the complexity of the objects being tracked, and the accuracy of the custom image dataset utilized for training. Furthermore, there may be other algorithms or approaches that could potentially yield more favorable results for visual object tracking.

III. PROPOSED WORK

YOLOv7 is an object detection model that uses the You Only Look Once (YOLO) architecture for real-time object detection. It has several modules that work together to detect objects and provide voice feedback.

A. Input Module

The input module takes the video frame as input to be processed by the YOLOv7 model. The obtained video frames may be of different sizes and aspect ratios, so they need to be resized to a uniform size that matches the input size of the YOLOv7 model. This resizing step ensures that the object detection model can process the frames efficiently. The pixel values of the resized frames are normalized to have a mean of zero and a standard deviation of one. This normalization step helps the model to learn better and converge faster during training. The blob function reorders the image channels to match the input format expected by the YOLOv7 model. The blob function groups the preprocessed frames into batches to speed up the training and inference process. The batch size is typically defined in the configuration file of the YOLOv7 model .After the blob function has prepared the input frames, they can be fed into the YOLOv7 neural network for object detection. The network outputs bounding boxes and class probabilities for all detected objects in each frame of the video.

B. Object Detection Module

The object detection module is responsible for detecting objects within the input image or video frame. Yolov7 uses a deep neural network architecture to detect objects with high accuracy and speed. The input image is preprocessed to resize it to a fixed size and normalize the pixel values.

The preprocessed image is then passed through a convolutional neural network (CNN) to extract features from the image. The input to the convolutional layers is a pre-processed image or a batch of images that have undergone resizing, normalization, and channel reordering. The convolutional layers use a set of filters (also called kernels) that are learned during the training process. The filters are applied to the input image to produce a set of feature maps. The feature maps are passed through a non-linear activation function which introduces non-linearity into the model and enables it to learn more complex representations. Pooling layers are used to reduce the spatial dimensions of the feature maps while preserving the important features. The most common type of pooling used in YOLOv7 is max-pooling, which takes the maximum value in a small region of the feature map. The convolutional layers are typically followed by down sampling layers, which reduce the spatial resolution of the feature maps. Down sampling is used to increase the receptive field of the filters and reduce the computational cost of the model. YOLOv7 uses skip connections to connect the output of one set of convolutional layers to the input of another set of convolutional layers. Skip connections help to preserve the low-level details and spatial information in the feature maps and enable the model to detect small objects. The feature maps from different layers are fused together to produce a set of feature maps that capture both low-level and high-level features of the input image. The fused feature maps are then fed into the detection head of the YOLOv7 model to produce the final object detection output. The extracted features are then used to predict bounding boxes and class probabilities for each object in the image. It is capable of detecting multiple objects within an image or video frame, and is trained on a large dataset of images with labeled objects.

C. Non-Maximum Suppression (NMS)

Non-Maximum suppression (NMS) is a post-processing technique used in YOLOv7 for object detection. It is used to eliminate redundant detections and retain only the most confident ones. It compares the predicted bounding boxes of the detected objects and remove the ones that have a high overlap or intersection. The algorithm selects the detection with the highest confidence score and removes all other detections that have a high overlap with it. NMS involves two main steps, thresholding and suppression. The thresholding step filters out detections with low confidence scores. The suppression step removes redundant detections that have a high overlap with each other, retaining only the most confident ones.By performing NMS, YOLOv7 can reduce the number of false positives and improve the accuracy of object detection.

D. Output Module

Voice guidance technologies are synthesized, allowing visually impaired individuals to identify object locations through voice guidance. This approach involves synthesizing the position and name of the object, utilizing Google text-to-speech (GTTS) technology. The output module combines the object detection and Text-To-Speech module to provide voice feedback to the user. It generates a voice message that announces the name of the detected object and its location within the image or video frame.

Overall, YOLOv7 with GTTS voice feedback is a powerful system that can accurately and quickly detect objects within images or video frames, and provide real-time voice feedback to visually impaired individuals.

IV. DESIGN DIAGRAM

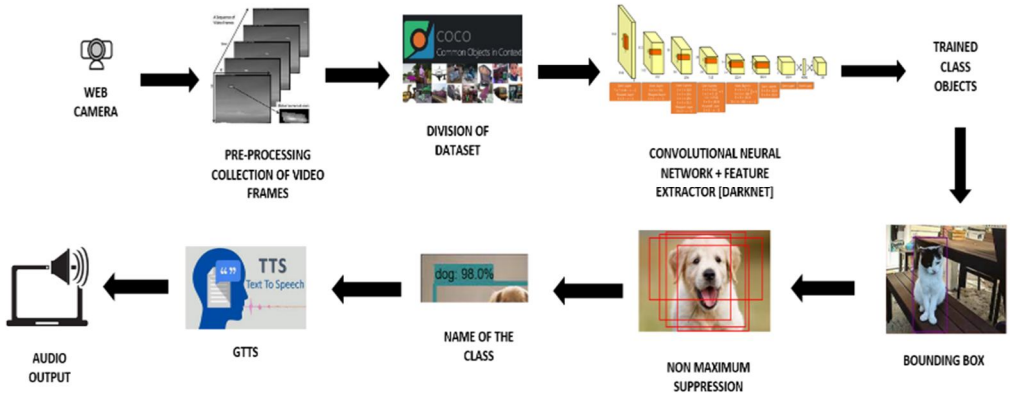


Fig.1 Detailed Design Diagram of the proposes system

V. EXPERIMENTAL RESULTS

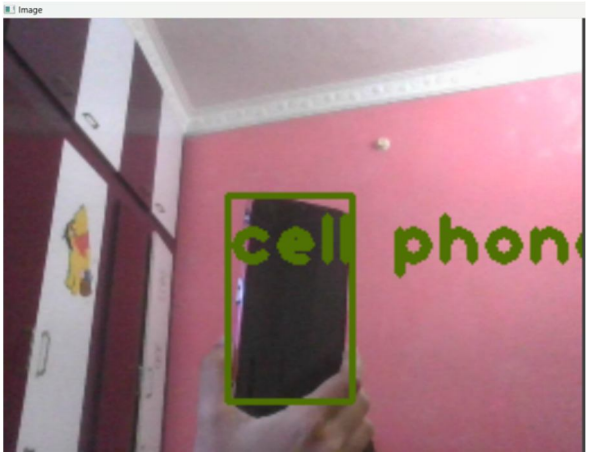


Fig.2 Object detection with audio output

Fig.3 Object detection with audio output

Fig.4 Object detection with audio output

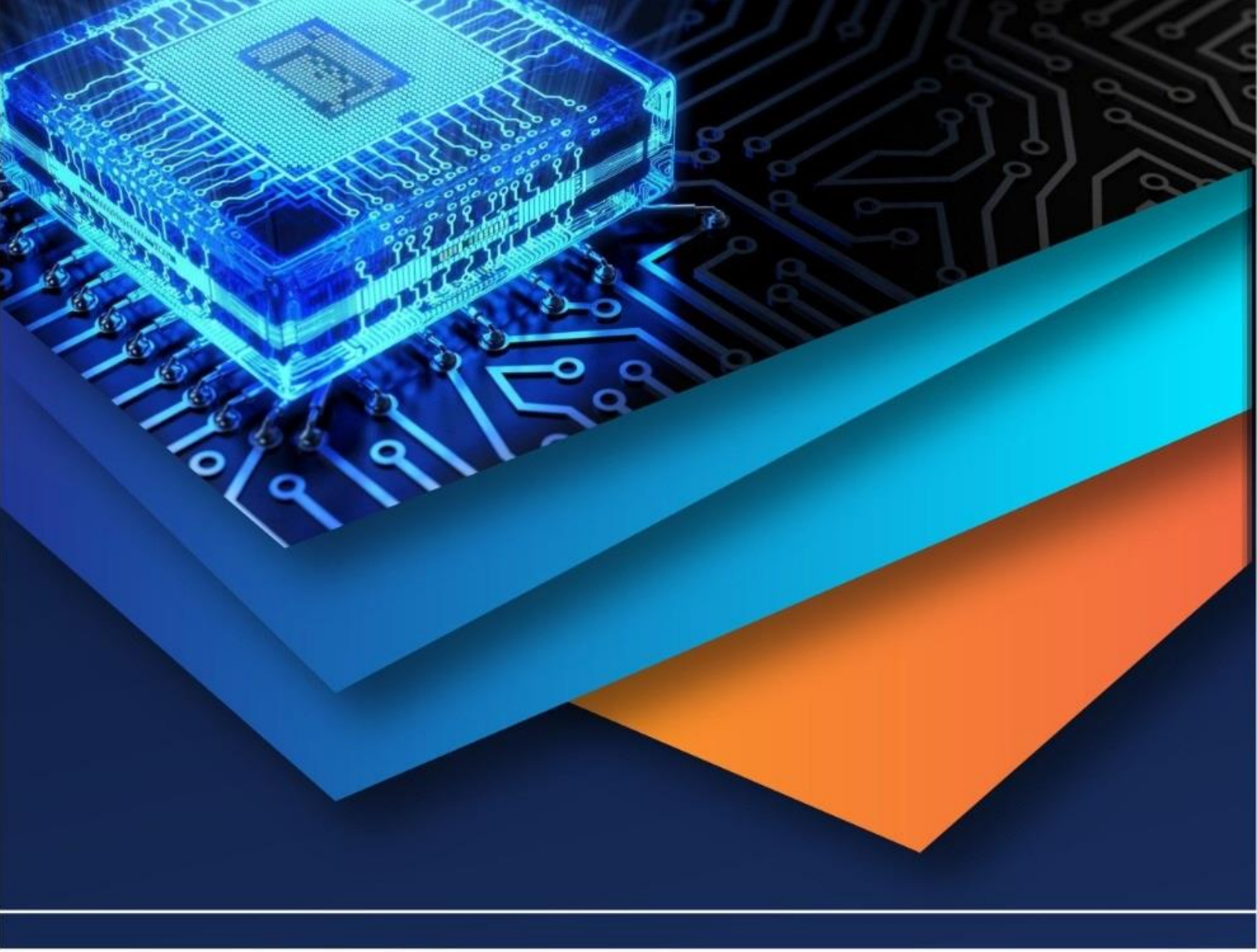
VI. CONCLUSION

The use of YOLOv7 for object detection with voice guidance using GTTS is a promising approach to assist visually impaired individuals in navigating their surroundings. The system's ability to detect and identify objects in real-time, coupled with the voice guidance provided by GTTS, can greatly improve the independence and safety of visually impaired individuals.

This technology has the potential to revolutionize the way visually impaired individuals interact with their environment and improve their quality of life. Further research and development in this area can lead to even more sophisticated and accurate systems that can be used in various settings, such as public transportation, shopping centers, and educational institutions. Overall, this technology has the potential to empower visually impaired individuals and help them lead more fulfilling lives.

REFERENCES

[1] Pirom, Wasin. "Object Detection and Position using CLIP with Thai Voice Command for Thai Visually Impaired." In 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), pp. 391-394. IEEE, 2022
[2] Yang, Wang, B. O. Ding, and Li Su Tong. "TS-YOLO: An efficient YOLO Network for Multi-scale Object Detection." In 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC), vol. 6, pp. 656-660. IEEE, 2022.
[3] Masud, Usman, Tareq Saeed, Hunida M. Malaikah, Fezan Ul Islam, and Ghulam Abbas. "Smart assistive system for visually impaired people obstruction avoidance through object detection and classification." IEEE Access 10 (2022): 13428-13441.
[4] Krishna, N. Murali, Ramidi Yashwanth Reddy, Mallu Sai Chandra Reddy, Kasibhatla Phani Madhav, and Gaikwad Sudham. "Object Detection and Tracking Using Yolo." In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1-7. IEEE, 2021.
[5] Ahmadi, Milad, Zichun Xu, Xinli Wang, Lei Wang, Mingjun Shao, and Youliang Yu. "Fast Multi Object Detection and Counting by YOLO V3." In 2021 China Automation Congress (CAC), pp. 7401-7404. IEEE, 2021.
[6] Fan, Jiayi, JangHyeon Lee, InSu Jung, and YongKeun Lee. "Improvement of object detection based on faster R-CNN and YOLO." In 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), pp. 1-4. IEEE, 2021
[7] Bhole, Swapnil, and Aniket Dhok. "Deep learning based object detection and recognition framework for the visually-impaired." In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 725-728. IEEE, 2020.
[8] Mahmud, Saifuddin, Redwanul Haque Sourave, Milon Islam, Xiangxu Lin, and Jong-Hoon Kim. "A vision based voice controlled indoor assistant robot for visually impaired people." In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-6. IEEE, 2020.
[9] Devi, A., M. Julie Therese, and R. Sankar Ganesh. "Smart navigation guidance system for visually challenged people." In 2020 International Conference on Smart Electronics and Communication (ICOSEC), pp. 615-619. IEEE, 2020.
[10] Cao, Danyang, Zhixin Chen, and Lei Gao. "An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks." Human-centric Computing and Information Sciences 10, no. 1 (2020): 1-22.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY