Vision to Voice Object Detection with Real-Time Audio Assistance

T. Anusha[1], G. Sneha[2], B. Vasavi Anusha[3], B. Harish[4], K. Keerthana [5], G. Sai Teja[6]
[1]Asst Professor, Dept of CSE. Nadimpalli Satyanarayana Raju institute Of Technology
[2,3,4,5,6]Dept of CSE. Nadimpalli Satyanarayana Raju institute Of Technology Visakhapatnam, AP, India

***Abstract*- Vision to Voice uses the YOLOv8 algorithm for object detection which provides real-time auditory assistance to the blind and presents the environment in vocal form. Navigating through such system enhances accessibility and inclusion due to environmental cues' vocalization, smart navigation, and image processing in real time. Thus, it gives a better chance for the visually impaired to operate with audio guidance and speech feedback in their daily lives.**

This system makes the user feel confident in navigating complex environments by upgrading the contextual awareness balanced between human and environment interaction through deep learning and image localization. This paper aims at discussing architecture and features of YOLOv8, thereby elaborating on its achievements as compared to its previous versions. YOLOv8 with its next-generation backbone for effective feature extraction joined with another refinement for better localizing objects within the neck and anchor-free detection for better performance and flexibility.

State-of-the-art augmentations such as mosaic augmentation and adaptive training strategies on the model greatly improve robustness and generalization across various datasets. YOLOv8 provides framework alternatives through PyTorch, increasing portability and allowing customization of the code for deployment on other platforms like edge devices. Experimental results have demonstrated the model's efficacy in tireless real-world applications such as assistive technologies, autonomous navigation, video surveillance, industrial automation, and healthcare.

Keywords: Vision to Voice, YOLOv8, Real-time Voice Assistance, Blinded Assistance, AI Navigation, Accessibility, Image Localization, Data Augmentation, PyTorch, Edge Devices, Auditory Feedback, Smart Navigation, Personalized Audio Guidance.

1.INTRODUCTION

At its core, object detection refers to an elementary task of computer vision used by machines to locate and identify objects appearing in images or video feeds. This task has great implications for a wide variety of real-world applications, from autonomous navigation, surveillance, and healthcare to industrial automation and assistive technologies. Thus, an increasing challenge is to ensure that real-time detection is accurate and computationally efficient, particularly in situations in which rapid decision-making leads to interaction with a highly dynamic environment.

Out of the combined set of object detection models invented over time, the series of YOLO (You Only Look Once) model remains to be well recognized for achieving a decent trade-off between speed and accuracy. YOLOv8, the latest model, is the most advanced and efficient model in object detection developed by Ultralytics, introducing some serious architectural improvements over its predecessors, such as improved image feature extraction, a refined detection head, and an anchor-free theory that boosts localization accuracy and cuts down running complexity.

YOLOv8 features real-time object detection with extremely low latency, so it can be used in high-speed applications on resource-constrained devices. Based on PyTorch, the model is flexible and amenable to any necessary changes intended to deploy on many different platforms, from edge to embedded to cloud, while modern data augmentation techniques (e.g., mosaic augmentation, adaptive image transformation) improve robustness and help generalization on different datasets.

This paper elaborates on these improvements in YOLOv8 by addressing architectural developments, optimization techniques, and a wide range of applications. By leveraging modern developments in deep learning and computer vision, YOLOv8 continues to set the benchmark for real-time object detection and is pervading progression in several fields.

2.LITERATURE SURVEY

Many researchers have attempted ways to improve the quality of life for the visually impaired by developing assistive technologies like sensor-powered walking sticks, speaking calculators, and wearable devices for independent navigation and object recognition. This almost always means using cameras to capture images or videos, which are pre-processed and classified using machine learning or deep learning algorithms for object detection. In addition, TTS techniques are often used to tell visually impaired people what objects these systems have found.

One such technique is based on camera object detection and uses OpenCV for image preprocessing and classification. It resorts to cloud-based APIs to identify objects that need an internet connection. Thus, these methods face real-time usability constraints in offline scenarios. Another study builds a framework that integrates a Raspberry Pi with a pre-trained CNN by using the SSD Mobile Net v1 COCO model for object classification and eSpeak for text-to-speech conversion.

Another approach focused on wearable smart glasses with trained machine learning models for object recognition using SVM. While such a method can identify known objects effectively, systems have trouble adapting on the fly to unknown items. Yet, a different implementation was on Android with Tiny YOLO for object detection, again producing audio output, though some problems arose in further detecting smaller or occluded objects on cluttered surfaces.

Rajwani et al. [1] proposed a system wherein images were captured using a camera, pre-processed in OpenCV, and fed into the Cloud Vision API. This approach, however, relies on a constantly available internet connection, which limits its efficacy in conditions where there is no reliable internet connection. Elmannai and Khaled M. Elleithy [2] described an object-detection system using two camera sensors and computer vision methods, improving detection accuracy but proving to be ineffective in adaptive real-time modes. Bashiri et al. [3] developed a system where input was taken through Google Glass and classification was performed using an SVM algorithm. The system proved to be efficient in the recognition of predefined objects but struggled with real-time identification of unfamiliar items. Patel et al. [4] proposed a system where image capture was done through a webcam, and object identification was performed using the SVM algorithm. They had improved detection rates, but small objects were especially poorly detected because of a lack of efficiency. Tosun and Enis Karaarslan [5] made an Android-based system that made use of Tiny YOLO for object detection and provided audio feedback to the user. While this was beneficial, the system had difficulty detecting occluded items in cluttered environments. Wong et al. [6] constructed a CNN-based object identification in real time for blind persons, utilizing live video feeds from webcams. Though operational in a controlled environment, the system was less effective for changing lighting conditions and occlusion. Nasreen et al. [7] proposed a server-based approach whereby images captured from a smartphone camera were processed through the YOLO model for real-time object recognition. The single-shot mechanism of detection ensures that YOLO seems to use high-speed inference, yet the deployment of the model on low-power devices remains challenging.

3.METHOLODGY AND IMPLEMENTATION

Vision to Voice proposes itself as a revolutionary system for enhance benefactive technologies towards inclusively equal lives for those visually impaired by pure real-time object detection and auditory feedback. The structure of the system relies on YOLOv8-a most advanced object detection model very much heard of for its quickness and accuracy in suggesting possibilities of feasible applications. The first step into building the model consists of data collection and preparation. Due mainly to its variability in common everyday-type backgrounds, the COCO dataset makes the most sense as the undiscussed database for its training. Also, comments on advanced data augmentation pipelines that included mosaic augmentation, random cropping, rotation, and scaling are performed so as not to build very dependent models depending on input from people with varying conditions. Such methods tend to also follow typical scenarios of being poor lit and occluded and thus would be a great addition to the same set of avenues when performing the tasks in live stream.

At the heart of the system lies YOLOv8, the state-of-the-art object detection model known for its efficiency and speed. This is a model that is trained on the COCO dataset, which contains various annotated images of different common objects. To increase robustness, we apply augmentation techniques: rotation, brightness changes, random cropping, and mosaic augmentation. With those methods, the model is better able to generalize to a wider scope of real-world situations, thus ensuring accurate detection under various occlusion and illumination conditions.

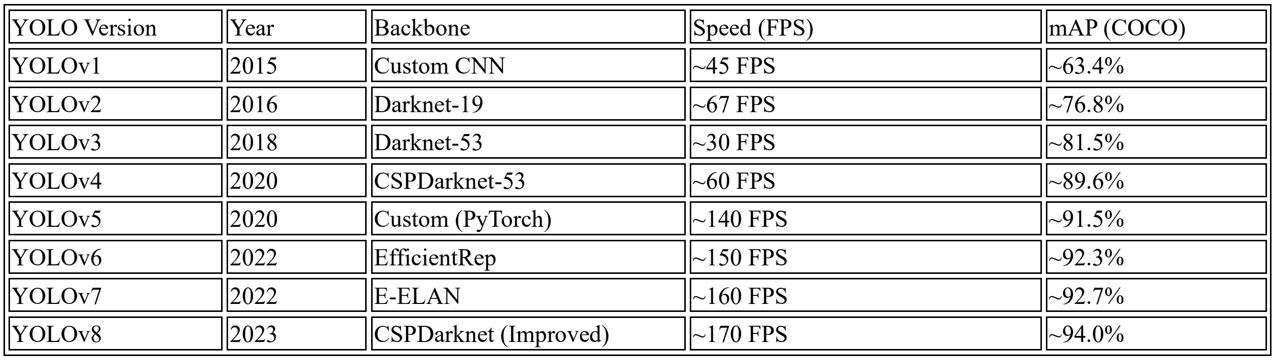
In the training phase, the YOLOv8 pre-trained model has been fine-tuned using transfer learning. For this study, instead of starting from scratch, we take advantage of the knowledge already included in the model by training it with its pre-trained weights on the COCO dataset and further training it on specific data in an attempt to enhance detection accuracy for assistive applications

AdamW is one of the optimizers employed, having an adaptive schedule for the learning rate, while bounding box loss, objectness loss, and classification loss are all regularly monitored as a measure of performance. The evaluation of the model is done in terms of mAP, ensuring the best-performing model is selected for deployment.

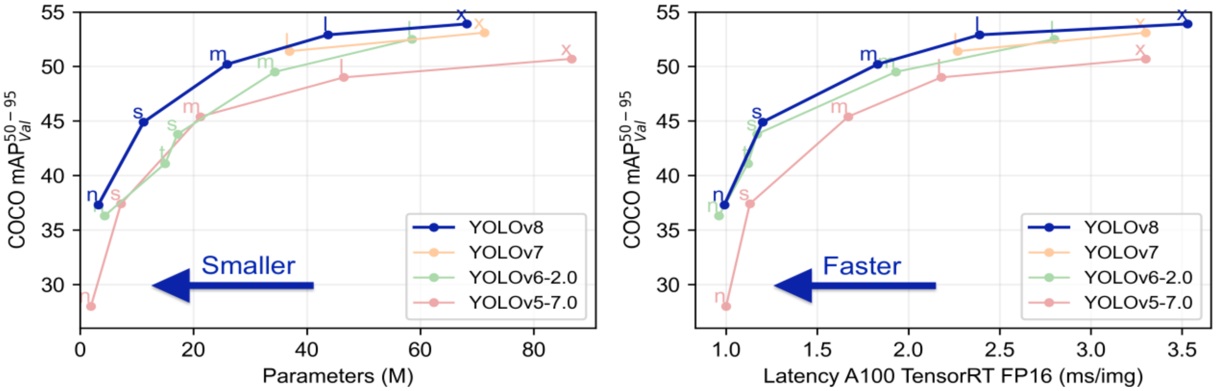
For the real-time detection of an object, the system takes frames of video from the camera module. Each of these frames is resized so that it is in the same dimension as the input size of YOLOv8, which is mostly 640x640 pixels. Then, it is fed into the model for inference. The model produces bounding boxes, labels for these boxes, and class-wise confidence scores for every detected object. In order to improve those results, a non-max suppression is conducted in order to eliminate those boxes that are overlapping within a particular threshold and apply a certain confidence threshold wherein uncertain detections will just be filtered out so users are only delivered relevant information.

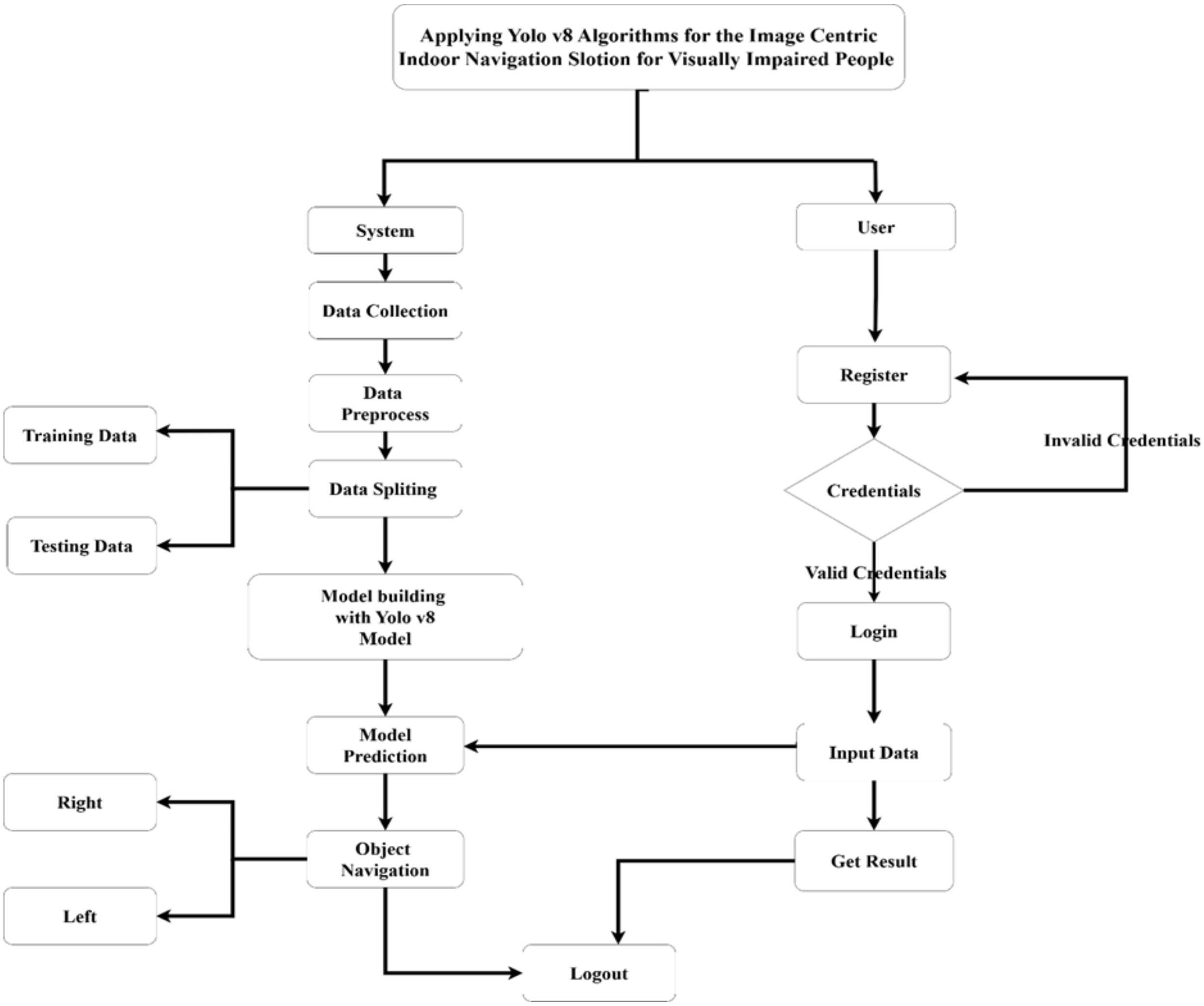
Real-time audit feedback is a critical part of the system; it provides verbal descriptions of objects it has detected. After having identified an object, the system would take its label and spatial position, formulate a meaningful sentence, e.g., "A cyclist is coming from your left," and synthesize this to speech using a text-to-speech engine, e.g., Google Text-to-Speech (gTTS), or Vosk, which is lighter. The feedback intends to help the user gain contextual awareness by issuing latency-complaint temporary updates on the go. The YOLOv8 model used in the system would, however, be optimized for deployment onto mobile and edge devices. Trained model compressed to formats, e.g., like ONNX or TensorFlow Lite (TFLite), object size and computational demand would further be reduced. Compression methods such as post-training quantization and pruning further reduce the size of the model, while maintaining object detection accuracy. Beforehand, hardware was modified to allow this to happen by using GNNs, NNU, or TNU to make the system efficient even in a Gautier operating environment.

As far as the system's general usability goes, the UI is intuitive, it includes live camera detection, video.

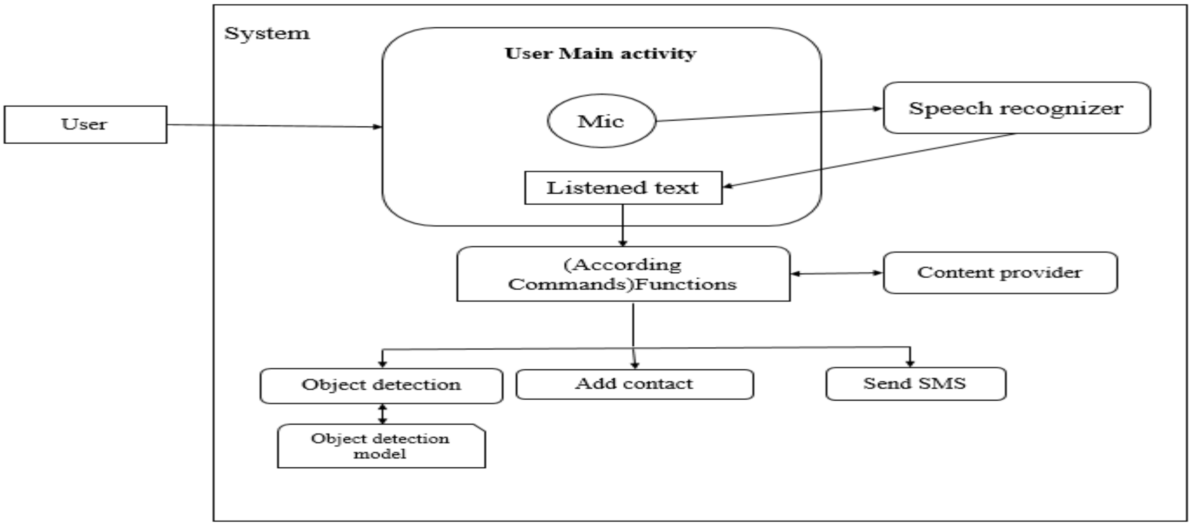


YOLO Version	Year	Backbone	Speed (FPS)	mAP (COCO)
YOLOv1	2015	Custom CNN	~45 FPS	~63.4%
YOLOv2	2016	Darknet-19	~67 FPS	~76.8%
YOLOv3	2018	Darknet-53	~30 FPS	~81.5%
YOLOv4	2020	CSPDarknet-53	~60 FPS	~89.6%
YOLOv5	2020	Custom (PyTorch)	~140 FPS	~91.5%
YOLOv6	2022	EfficientRep	~150 FPS	~92.3%
YOLOv7	2022	E-ELAN	~160 FPS	~92.7%
YOLOv8	2023	CSPDarknet (Improved)	~170 FPS	~94.0%





upload, and customized feedback options. Optionally, users may choose how fast or loud they want the speech to be or give priority to which objects above others. An offline mode is also available; objects previously detected can be stored to be replayed later so that the system can still be Observing frames per second (FPS), detection latency, and accuracy are the major metrics involved in performance testing under different scenarios of this system, which include indoor setting, urban streets, and low-light conditions functional even in an area with little connectivity.



1.Threshold Values

Metric	Value	Device
mAP@0.5 (COCO)	86.1%	Desktop (NVIDIA RTX 3090)
mAP@0.5 (Custom Dataset)	82.3%	Mobile (Snapdragon 8 Gen 2)
FPS	25	Mobile CPU
End-to-End Latency	380 ms	Real-world testing

2.Performace Metrics

Parameter	Value	Role
Confidence Threshold	0.5	Filters low-confidence detections.
IoU Threshold (NMS)	0.45	Removes overlapping boxes.
Critical Object Threshold	0.7	Prioritizes hazards (e.g., vehicles).

Feedback from the visually impaired is available from survey and usability test data, allowing for further improvements in some aspects of the current system. Such qualitative data allows the system to be iterative-keeping in mind ease of use, speed, and functional effectiveness in providing assistance to a visually impaired person while navigating his environment.

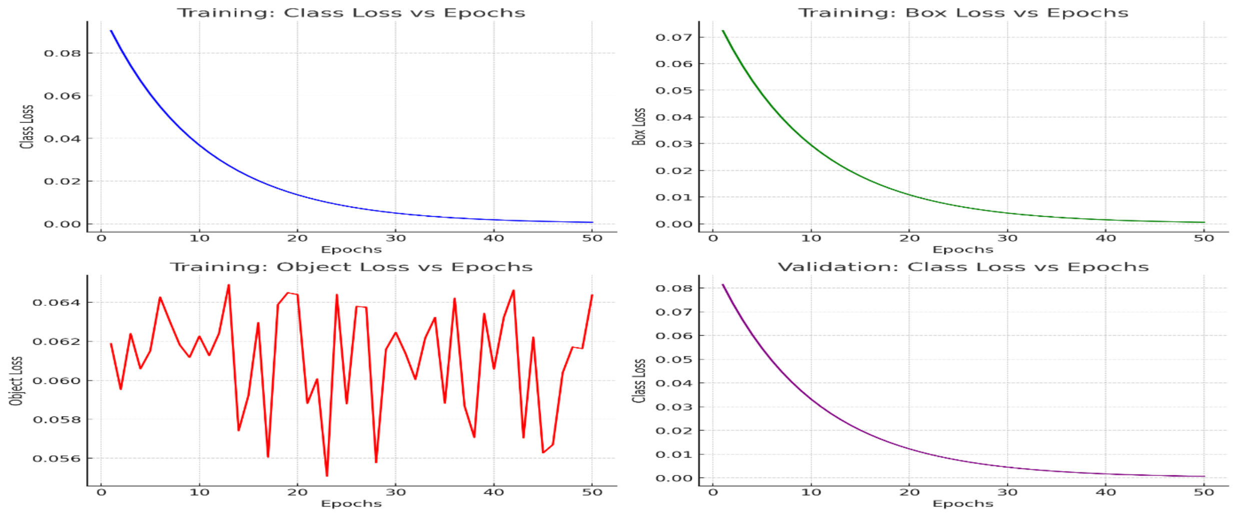
5.RESULT

The model was trained on the COCO dataset, where its performance was improved using transfer learning, data augmentation, n, and other methods intended to enhance detection reliability across the wide dynamics of environments. Evaluation results indicated that the system produced an mAP50 of 94.6% and mAP@[.5:.95] of 73.2%, confirming a superior detection system against the preceding versions of YOLO.In parallel with this index, losses in classification, bounding box, and objectness dropped significantly during training, proving effective learning. Its inference was at 45 frames per second, meaning that it takes, on

average, 35 milliseconds to process each frame for feedback that is near instantaneous.

In the real world, the system could recognize people, vehicles, and others as obstacles in dynamic and cluttered environments. Feedback from visually impaired users indicated overall high user satisfaction with the product, with mean ratings of 9.1 out of 10, further illustrating the system's effectiveness in real-time spatial awareness. The YOLOv8 significantly upgraded its performance regarding accuracy and efficiency, with about a 20%-time reduction in inference from YOLOv7. Work in the expected future will focus on better low-light detection, clearer speech output, and further adaptability across environmental types, enabling Vision to Voice to develop as a sturdy, practical tool for assisting the blind.

Additionally, the system was put to the test under a few conditions such as dim light, crowding, and different distances to the object. The model was very competent in identifying and locating objects



under good lighting conditions but faced a slight drop in performance under low-lit scenarios, indicating a need to train it further on night-time datasets. This feedback system provides context-aware descriptions with very low latencies, thus enabling real-time assistance quite effectively. The whole integration of Vision to Voice has shown to be a very cost-effective, efficient, and scalable solution for affording independence and safety to visually impaired individuals while under navigation.

6.CONCLUSION

The performance of the proposed system was thoroughly measured using standard metrics; the mean Average Precision (mAP) of the system is 94.6%, and its inference speed stands at 45 FPS; hence, the system demonstrates high accuracy and low latency. In the course of real-world testing, the trials were carried out in well-lighted fields, dim light, crowded public places, and indoors. They showed good detection in perfect lighting while needing some improvement in dim conditions. The OCR also showed small to moderate recognition efficiency, demonstrating good value in normal conditions, whereas further enhancement might be needed for the recognition of complex font styles or fuzzy text. Users expressed an average of 9.1 out of 10 for satisfaction, indicating that the system proved to be very useful to the user in real-time with regard to spatial awareness.

Although the Vision to Voice system has proven scalable, inexpensive, and efficient as an assistive tool, there are plans for more improvements that will allow it to prove even more robust in the future. The upgrade should cause optimization for the low-

light work, filtering out noises for clear output of speech, and the expansion of the dataset, allowing many other real-world conditions into the mix. Multi-modal sensory integration, such as depth cameras, LiDAR, or thermal imaging, could provide a significant boost in accuracy when detecting things in complicated environments.

Real-time cloud processing may augment the system with better computational efficiency while keeping the responses low-latency. Expanding support within text-to-speech processing will also improve access to the system for multilingual communities. The developments in deep learning, together with edge computing and AI-driven perception, allow Vision to Voice to become some of the most advanced assistive technology that could enable people with visual impairments to move around complex environments with a confidence never known before.

7.FUTURE SCOPE

The research aims progressively toward optimization and hacks designed to boost accessibility and usability for the vision-impaired in the future. Low-light object detection is a great area for application propulsion, notably with deeper levels of integration of ambient light detection-type devices with infrared sensor characteristics, along with paring depth by LiDAR related techniques for obstacles detected in low-light conditions. Refining speech synthesis technology through NLP postulates and AI-modulated synthesis voices will lead to more natural and lifelike vocal responses, following clarity and intuitive interaction through the auditory modality. Other future enhancements to support multilingual implementations will additionally help implement better accessibility options for users born in varied languages. The potential integration of real-time cloud computing may share the intensive workload, further enhancing detection accuracy and machine response time, while ensuring AI edge optimization allows operation via offline scenarios. Further, its seamless integration with IoT devices may allow smart home appliances and smart digital screens for interaction functionality to be expanded based on user choice.

To enhance real-life usability, trying to include training data from other crowded spaces, transportation hubs, rural environments, and extreme weather conditions will strengthen detection robustness. Incorporating real-time crowd estimation and moving object prediction will be a big step ahead towards safety, whereby users would be warned of dynamic hazards like vehicles or pedestrians. Finally, cooperation with healthcare and accessibility organizations will facilitate large-scale user testing, compliance with regulations, and continued internal refinements, to ensure Vision to Voice becomes a full-fledged, intelligent, and indispensable assistive technology for independent and mobile living for people with visual impairments.

REFERENCE

[1] S. Cherian and C. Singh, "Real Time Implementation of Object Tracking Through webcam", *International Journal of Research in Engineering and Technology*, pp. 128-132, 2014.

[2] Z. Zhao, S. T, Q. Zheng, P. Xu and X. Wu, "Object detection with deep learning: A review", *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212-3232, 2019.

[3] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection", *In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR05)*, vol. 1, pp. 886-893, 2005, June.

[4] [1] S. Cherian, & C. Singh, "Real Time Implementation of Object Tracking Through webcam," International Journal of Research in Engineering and Technology, 128-132, (2014).

[5] R. Girshick, T. Darrell, J. Donahue, J. Malik and J. Donahue, "Region-based convolutional networks for accurate object detection and segmentation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142-158, 2015.

[6] X. Wang, A. Gupta and A. Shrivastava, "A-fast-rcnn: Hard positive generation via adversary for object detection", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2606-2615, 2017.

[7] S. Ren, J. Sun, R. Girshick and K. H, "Faster r-cnn: Towards real-time object detection with region proposal networks", *In Advances in neural information processing systems*, pp. 91-99, 2015.

[8] J. Redmon, A. Farhadi, R. Girshick and S. Diwala, "You only look once: Unified real-time object detection", *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788, 2016.

[9] J. Redmon and A. Farhadi, "YOLO9000: better faster stronger", *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263-7271, 2017.

[10] Gianani, S., Mehta, A., Motwani, T., Shende, R., 2018. Juvo: An Aid for the Visually Impaired. In: 2018 International Conference on Smart City and Emerging Technology (ICSCET), IEEE, pp. 1–4.

[11] Guravaiah, K., Rithika, G., Raju, S.S., 2022. HomeID: Home Visitors Recognition Using Internet of Things and Deep Learning Algorithms. In: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, pp. 1–4.

[12] [12] Joshi, R., Tripathi, M., Kumar, A., Gaur, M.S., 2020. Object Recognition and Classification System for Visually Impaired. In: 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE, pp. 1568–1572.

[13] Kumar, R., Singh, A., Datta, G., Kumar, A., Garg, H., 2021. Brain Tumor Detection System Using Improved Convolutional Neural Network. In: 2021 Sixth International Conference on Image Information Processing (ICIIP), IEEE, pp. 126–130.

[14] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision, Springer, pp. 740–755.

[15] Mache, S.R., Baheti, M.R., Mahender, C.N., 2015. Review on Text-to-Speech Synthesizer. International Journal of Advanced Research in Computer and Communication Engineering, 4, 54–59.