Self-Guiding System for Blind People

Dr. J.P. Patra,[a] Abhinaw Jagtap,[b] Ritik Kumar,[c] Shruti Dubey,[d] and Saurabh Bharti[e]

University Teaching Department, Chhattisgarh Swami Vivekanand Technical University
India, Chhattisgarh, Bhilai pin: 491107

[a]Associate Professor, University Teaching Department, jppatra.cse@csvtu.ac.in
[b]Assistant Professor, University Teaching Department, accabhinaw3906@gmail.com
[c]Student, ritik9802official@gmail.com
[d]Student, shrutidubey04669@gmail.com
[e]Student, saurabh104.2011@gmail.com

Abstract. Globally, nearly 43 million people live with blindness, and around 295 million people have visual impairments, facing significant daily challenges in navigating and perceiving their surroundings. To address this, our research presents an intelligent, real-time assistive system that delivers audio-based descriptions of nearby objects and their spatial orientation relative to the user. Utilizing the advanced YOLO v8 model for robust object detection, our system identifies objects within the camera's field of view and communicates their identities and precise locations through Android's Speech Engine. This integration of high-performance detection algorithms with dynamic speech synthesis enhances users' spatial awareness, fostering greater independence and safety. Our approach provides a seamless experience for visually impaired individuals, promoting situational awareness and supporting autonomy in complex environments.

INTRODUCTION

In a world where visual perception often dictates how we interact with our environment, nearly 43 million people living with blindness face unprecedented challenges every day. These individuals navigate a landscape that is largely designed with sighted individuals in mind, encountering obstacles that range from physical barriers to social isolation. The impact of visual impairment extends beyond mere mobility; it permeates every aspect of life, influencing education, employment, personal relationships, and overall well-being. The need for innovative solutions that enhance the quality of life for those with visual impairments is more critical than ever.

Technological advancements have opened new avenues for addressing the challenges faced by individuals who are blind or have low vision. Assistive technologies, which are designed to improve the functional capabilities of individuals with disabilities, are evolving rapidly. Among these innovations are intelligent systems that leverage artificial intelligence and machine learning to provide meaningful assistance. The integration of such technologies into daily life promises to empower individuals with visual impairments, granting them greater independence and improving their situational awareness.

This research is predicated on understanding one's environment is essential for independence. Furthermore, identifying and navigating through obstacles can significantly enhance mobility and safety. Our project seeks to bridge the gap between visual information and auditory comprehension through an intelligent system capable of delivering real-time, audio-based descriptions of the surrounding environment. By utilizing the advanced YOLO (You Only Look Once) v8 object detection model, we aim to create a solution that is not only efficient but also user-friendly. Similar solution exist with a lower version of YOLOv5 [1].

Historically, solutions for blindness, such as guide dogs and long white canes, have been invaluable. They assist and enhance mobility, but they do have limitations. While guide dogs require extensive training and a level of commitment that not all users can provide, traditional canes offer minimal information about the environment outside of direct contact. The introduction of electronic travel aids, such as GPS devices and smartphone applications, has contributed to enhancing outdoor navigation for individuals with low vision or blindness. Yet, many of these systems lack the real-time feedback that can facilitate better decision-making while navigating complex and dynamic environments.

Current technological solutions are increasingly moving towards augmented reality and artificial intelligence. The landscape for assistive technology has the potential to be radically transformed by devices that can not only perceive the environment but also communicate relevant information back to users. The advancements in computer vision, particularly in the realm of rapid object detection and recognition, have paved the way for innovative approaches to accessibility.

The primary objective of this research is to bridge the sensory gap that individuals with visual impairments face in understanding their environment. We propose to develop a system that can offer real-time descriptions of surrounding

objects, their identities, and spatial relationships, aiding navigation and increasing independence. By implementing advanced machine learning algorithms, specifically the YOLO v8 model, we aim for a high-performance object detection system that can be used in everyday scenarios. Through the combination of object detection and speech synthesis technologies, our research aims to create a comprehensive tool that not only identifies objects but also conveys their significance in real-time. By utilizing Android's Speech Engine for auditory feedback, we hope to deliver a seamless experience where users can receive instant and accurate descriptions of their surroundings without needing extensive technological knowledge or training. The significance of this study lies in its potential to improve the daily lives of millions of people living with blindness. By enhancing situational awareness and providing real-time information, the system can foster a greater sense of independence. Individuals will have the ability to navigate their environments with increased confidence, reducing reliance on others and allowing for more spontaneous and fulfilling social interactions. Furthermore, the potential for this technology to be integrated into existing smart devices offers a scalable solution that can be widely adopted.

Moreover, the research contributes to the field of assistive technology by exploring the intersection of artificial intelligence, audio feedback, and user experience design. The findings can serve as a foundation for future developments in this space, potentially leading to more sophisticated and inclusive technologies. To achieve the research objectives, a systematic approach will be taken. The first phase will involve the development of the YOLO v8-based object detection system, focusing on optimizing speed and accuracy. This will be followed by the integration of the speech synthesis engine to convey real-time information effectively. Throughout the development phase, user testing will be conducted to gather feedback from individuals with visual impairments. This user-centric approach ensures that the final product is tailored to meet the needs of its intended users.

Furthermore, we will explore the technical challenges associated with real-time object detection and synthesis, such as processing time, battery consumption, and system responsiveness. By addressing these challenges, we can enhance the practical applicability of the system in real-world settings. In conclusion, the introduction of intelligent systems that cater to the needs of individuals with visual impairments represents a paradigm shift in the field of assistive technology. The ability to provide real-time, audio-based descriptions of one's environment can substantially elevate the quality of life for millions. This research not only aims to develop a cutting-edge solution but also seeks to advocate for inclusion and accessibility in technology design.

The existing technologies are complex to use, many times have solutions which cannot be used alone by visually impaired people, whereas our system provides a fully self guided system. Our endeavor emphasizes that blindness should not equate to invisibility. Through the application of advanced technologies, we can create a future where individuals with visual impairments can confidently navigate their surroundings, embrace independence, and participate fully in society. The journey to achieving this vision is an inspiring challenge, one that requires collaboration between researchers, technology developers, and the community of individuals directly impacted by these advancements. With a commitment to innovation and empathy, we can pave the way towards a more inclusive world.

LITERATURE REVIEW

The work till date has machine learning framework aimed at assisting visually impaired individuals through object detection and recognition technologies. This system leverages SSD architecture and COCO datasets, employing TensorFlow for training and Python programming for implementation. By capturing images through a camera and analyzing them with predefined datasets, the system provides real-time distance estimation and generates audio alerts to guide users in their surroundings.[1]

These works collectively underscore the progression of methodologies and technologies, from CNNs to YOLO models, contributing to real-time, efficient object detection frameworks for various applications. An overview of prior research in object detection and recognition. Key contributions from the literature include: [2]

2019: Object Detection Using Convolutional Neural Networks (CNN) [3]

This study explored vision systems essential for mobile robotics, comparing two models: SSD with MobileNetV1 and Faster-RCNN with InceptionV2, emphasizing the role of CNN in detecting objects for tasks like navigation and surveillance.

2019: Image-Based Real-Time Object Detection and Recognition in Image Processing [4]

Focused on detecting and tracking humans and vehicles, this research reviewed methodologies like feature-based, region-based, outline-based, and model-based detection, highlighting their application in surveillance and image retrieval.

2020: Salient Objects Detecting with Segment Features Using Mean Shift Technology [5]

Proposed a saliency detection method using four steps: regional feature extraction, segment clustering, saliency score computation, and post-processing, aimed at recognizing objects in complex images.

2020: Real-Time Object Detection Using Deep Learning [6]
Presented a deep learning approach incorporating Darknet-53 for feature extraction, with feature map up-sampling and concatenation to enhance object detection in images and videos.

2020: Assistive Object Finding System for Visually Impaired People [7]
Developed a system integrating real-time image processing, GPS, ultrasonic sensors, and YOLO-based deep neural networks to detect objects, separating backgrounds and foregrounds to enhance recognition accuracy for visually impaired individuals.

These works collectively underscore the progression of methodologies and technologies, from CNNs [8] to YOLO models, contributing to real-time, efficient object detection frameworks for various applications.

METHODOLOGY

The goal of our project is to develop a self-guiding Android application for visually impaired users, which identifies and narrates objects in the camera's view. The project involves object detection using YOLOv8 and is implemented on an Android platform with the help of TensorFlow Lite (TFLite).

We begin with understanding YOLOv8 algorithm [9]. YOLOv8 (You Only Look Once version 8) is a state-of-the-art object detection model optimized for high-speed performance and real-time applications. YOLOv8's architecture combines several advanced components that allow it to detect objects effectively. It is developed by Ultralytics, an advanced object detection model designed to balance speed, accuracy, and user-friendliness. Building on the foundational efficiency of its predecessors, YOLOv8 incorporates key innovations, such as anchor-free detection and optimized architecture. The model processes input images in a single network pass, enabling real-time object detection with improved precision and reduced computational overhead compared to multi-stage detection techniques.

The YOLOv8 family consists of multiple variants—n, s, m, l, and xl—each tailored to specific needs by adjusting three key parameters:

- Depth Multiple (d): Determines the number of Bottleneck Blocks used in the C2f block. Increasing depth enhances accuracy at the cost of speed.
- Width Multiple (w): Adjusts the number of channels in convolutional layers. Wider models are more accurate but require greater computational power.
- Max Channels (mc): Limits the number of channels to control model size and avoid overfitting.

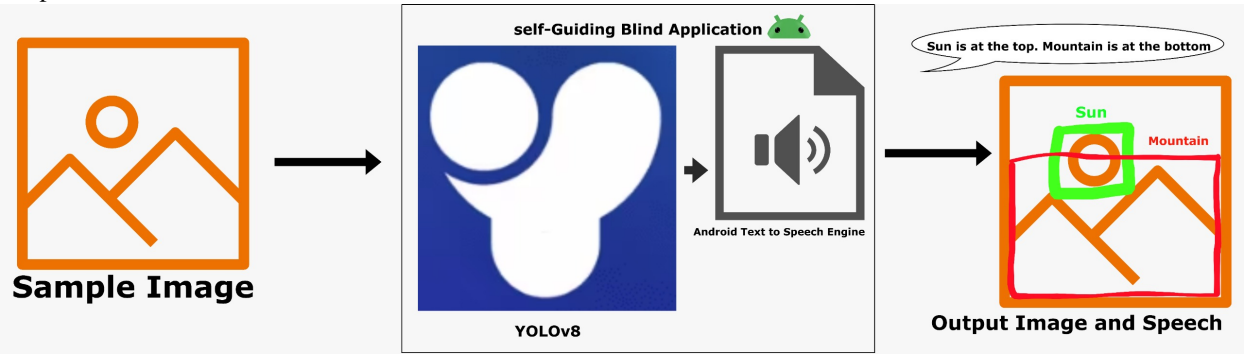
These variations allow YOLOv8 to cater to applications requiring different trade-offs between inference speed and accuracy.

The architecture of YOLOv8 is divided into three sections: Backbone, Neck, and Head, each playing a critical role in processing input images and predicting outputs.

Backbone: Acts as a feature extractor. Begins with a convolutional block that reduces spatial resolution while increasing feature depth. Incorporates C2f and Spatial Pyramid Pooling Fast (SPPF) blocks for feature refinement and multi-scale processing. Example
The C2f block splits feature maps and processes them through Bottleneck Blocks and direct paths, leveraging shortcut connections for efficiency. SPPF generates fixed feature representations without resizing or losing spatial information, enhancing multi-scale detection capabilities. Neck: Combines feature maps from different Backbone layers. Uses upsampling and concatenation to merge fine and coarse details, preparing features for detection. Head: Responsible for the final predictions. Consists of Detect blocks for anchor-free detection, which predict object centers, bounding boxes, and class probabilities directly. Specialized Detect blocks handle objects of varying sizes: small, medium, and large.

As shown in Figure 1, The YOLOv8 model was employed for real-time object detection in the project. The model was exported to the TFLite format for seamless integration into an Android application. TFLite's lightweight and

optimized architecture allowed efficient deployment on mobile devices, enabling on-device inference. The application processes input images or video frames, and YOLOv8 detects objects, outputting bounding boxes and text labels. These results are then leveraged to provide spatial context. An Android text engine synthesizes this information, generating descriptive sentences about the relative positions of detected objects.
Output Flow is as follows:



Object Detection: YOLOv8 identifies objects like "chair," "table," and "laptop," assigning bounding boxes and labels.
Spatial Analysis: Bounding box coordinates are used to determine relative positions, such as "The laptop is on the table" or "The chair is to the left of the table."
Text-to-Speech: The Android text engine converts the spatial descriptions into audio output, enhancing accessibility and user experience.
The Android app, named selfGuidingBlindApp, is developed in Kotlin. The app is structured into several core components that handle object detection, overlaying bounding boxes, and generating audio feedback. The TFLite model, along with its metadata, is loaded using the MetaData class. This class performs following two main taks.

1. Using TensorFlow Lite's MetadataExtractor, the app reads metadata associated with the model. This metadata includes class names and other information required for labeling detected objects. If the metadata extraction fails, the app relies on a predefined set of temporary class names as placeholders.

2. If a label file is provided, the app reads the file to obtain the list of class names for detected objects. This label data is stored in Constants.kt , specifying the paths for both the model and label files.

The detected objects are represented as instances of the BoundingBox data class, which defines the coordinates and attributes for each bounding box. This structure includes Coordinates (x1, y1, x2, y2) for the bounding box corners, Center coordinates (cx, cy), dimensions (w, h), Confidence score (cnf), class index (cls), Class name (clsName) which is the label of the detected object.

To visually represent the detections, the app includes an OverlayView class, responsible for drawing bounding boxes on the screen. The OverlayView uses a Paint object to draw rectangles over detected objects in the live camera feed. The color and thickness of the bounding boxes are customizable. For each bounding box, the app calculates the label's background and position. The object name is drawn near the bounding box, with a solid background for better visibility. Each time new detections are made, the bounding boxes are updated by calling setResults on the OverlayView. This triggers a redraw of the bounding boxes on the preview.

The core object detection logic is as follows.

- The Detector class preprocesses the image to match the model's input requirements, resizing and normalizing it using an ImageProcessor.
- The image is fed to the model, and the interpreter returns an array containing bounding box coordinates, class indices, and confidence scores for each detected object.
- The best box function filters detections based on a confidence threshold. Only bounding boxes with confidence scores above this threshold are considered valid. Non-maximum suppression (NMS) is applied to reduce overlapping boxes, ensuring that only the most relevant bounding box for each object is displayed.

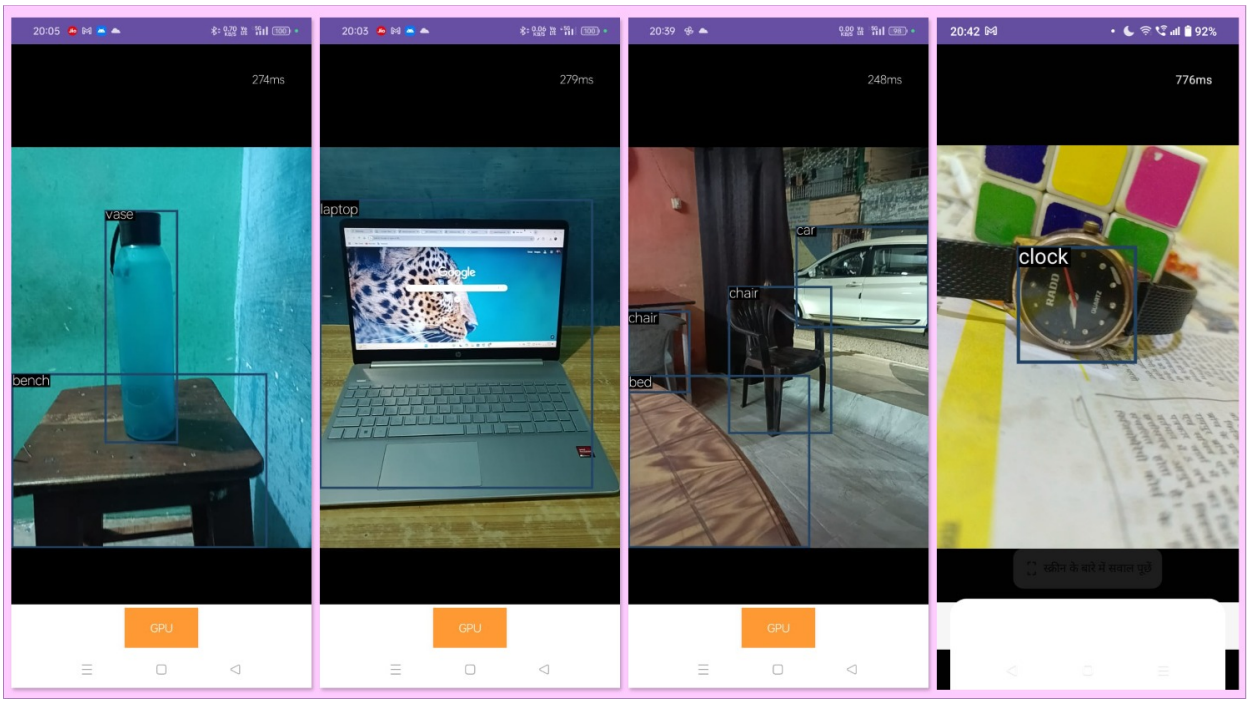
The app uses Android's Text-to-Speech (TTS) engine to provide audio feedback about detected objects. The TTS engine is initialized in MainActivity. It is configured to use US English as the language for narration. When a detection is made, the app determines the relative position (left, center, or right) of each object. A description is generated,

combining the object's class name with its relative position (e.g., "person on the left"). The description is passed to the TTS engine, which speaks the text aloud, informing the user of objects in their surroundings.
We utilized the CameraX library for real-time image capture and analysis. The camera is set up in MainActivity, using CameraX's lifecycle-aware ProcessCameraProvider. The app selects the rear camera by default. Each camera frame is processed through an image analysis pipeline, with the frame being analyzed in the detect function in Detector. kt. The detection results are then passed to onDetect, which updates the overlay and triggers TTS narration if detections are found.

If the device supports GPU acceleration is enabled for the TFLite interpreter, improving inference times and enabling smoother real-time object detection. The camera processing and detection run on separate threads, allowing the UI to remain responsive. The app uses an ExecutorService to manage these background tasks efficiently. The app provides an option to enable or disable GPU acceleration based on the device's capabilities, ensuring compatibility with a wide range of Android devices.

RESULTS AND DISCUSSIONS

The developed application demonstrates exceptional real-time object detection performance using YOLOv8, effectively meeting its goal of enhancing accessibility for visually impaired users, as explained in Figure 2. Compared to prior works, it offers notable advancements in model efficiency and user accessibility. By utilizing the Android Speech Engine for real-time, position-based audio feedback, the system provides seamless guidance, which earlier solutions lacked.



The application's standout features include its cost-effectiveness, ease of access, and fully self-guided functionality, making it a practical solution for a wide range of users. Its deployment on Android devices ensures affordability and widespread availability without requiring additional specialized hardware. The comparison is shown in Table 1.

The dynamic bounding box updates and TTS narration significantly improve spatial awareness. However, occasional interruptions in audio feedback due to overlapping TTS calls during rapid frame processing highlight an area for optimization, especially in environments with high object density or motion.
This self-contained, accessible, and low-cost solution leverages YOLOv8's efficient detection to provide visually impaired users with immediate, actionable feedback, setting a new standard for assistive technologies.

Research Paper	Model Used	Speech System used	Accessibility
OBJECT DETECTION AND RECOGNITION USING TENSORFLOW FOR BLIND PEOPLE	SSD Detection Model	Default voice notes in Python	Less
Third Eye: Object Recognition and Speech Generation for Visually Impaired	YOLOv5	pyttsx3, gTTS	Less
Self-Guiding System for Blind People	YOLOv8	Android Speech Engine	More

TABLE 1. Comparison Table

CONCLUSION

In conclusion, our research addresses a critical issue faced by nearly 43 million people worldwide living with blindness. These individuals navigate environments that are predominantly designed for the sighted, leading to a range of challenges that significantly impact their mobility, safety, and overall quality of life. Traditional assistive devices like guide dogs and long white canes, although valuable, have inherent limitations that do not fully meet the demands of modern mobility and situational awareness. This study presents an innovative solution by integrating advanced technology, specifically the YOLO v8 object detection model, with real-time audio feedback. By providing immediate and relevant information about the surrounding environment, our system not only enhances users' spatial awareness but also empowers them to make informed decisions while navigating complex spaces. This capability is essential in promoting greater independence and confidence for individuals with visual impairments. The potential implications of our research extend beyond mere mobility aid. By fostering independence and improving accessibility, we aim to influence areas such as education and employment opportunities for people with visual impairments, we provide a cheaper and fully guided solution for people with vision defections. This work aligns with global efforts to create inclusive communities and tackle the societal challenges associated with disabilities. Furthermore, as technologies such as artificial intelligence and machine learning continue to evolve, their integration into assistive devices presents an opportunity to redefine how individuals with visual impairments interact with their surroundings. Our research not only emphasizes the necessity of such advancements but also sets a foundation for further exploration in this domain. We believe that presenting our findings at this conference will engage stakeholders in discussions about the future of assistive technologies and raise awareness of the ongoing challenges faced by those with visual impairments. By spotlighting this critical topic, we hope to inspire collaborative efforts towards innovative solutions that promote an inclusive society for all.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all those who supported us throughout the course of this research. We extend our heartfelt thanks to our academic mentors and advisors at the University Teaching Department, Chhattisgarh Swami Vivekanand Technical University, for their invaluable guidance and encouragement. We are also grateful to our peers and colleagues who contributed their time and insights during our discussions and collaborative efforts. Their input helped us refine our ideas and enhance the quality of our work. Additionally, we would like to acknowledge the communities and organizations dedicated to supporting individuals with visual impairments. Their commitment to improving the lives of those with disabilities has inspired us and provided crucial context for our research. Lastly, we thank our families and friends for their unwavering support, patience, and motivation during this project. Without their encouragement, this research would not have been possible.

REFERENCES

1. K. Guravaiah, Y. S. Bhavadeesh, P. Shwejan, A. H. Vardhan, and S. Lavanya, "Third eye: object recognition and speech generation for visually impaired," Procedia Computer Science **218**, 1144–1155 (2023).
2. P. Devika, S. P. Jeswanth, B. Nagamani, T. A. Chowdary, M. Kaveripakam, and N. Chandu, "Object detection and recognition using tensorflow for blind people," International Research Journal of Modernization in Engineering Technology and Science **4**, 1884–1886 (2022).
3. R. L. Galvez, A. A. Bandala, E. P. Dadios, R. R. P. Vicerra, and J. M. Z. Maningo, "Object detection using convolutional neural networks," in *TENCON 2018-2018 IEEE region 10 conference* (IEEE, 2018) pp. 2023–2027.
4. Q. Wu and Y. Zhou, "Real-time object detection based on unmanned aerial vehicle," in *2019 IEEE 8th data driven control and learning systems conference (DDCLS)* (IEEE, 2019) pp. 574–579.

5. A. K. Gupta, A. Seal, M. Prasad, and P. Khanna, "Salient object detection techniques in computer vision—a survey," Entropy **22**, 1174 (2020).
6. W. Tarimo, M. M. Sabra, and S. Hendre, "Real-time deep learning-based object detection framework," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2020) pp. 1829–1836.
7. S. Shaikh, V. Karale, and G. Tawde, "Assistive object recognition system for visually impaired," International Journal of Engineering Research & Technology (IJERT) **9**, 736–740 (2020).
8. S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," (2019), arXiv:1809.03193 [cs.CV].
9. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," (2016), arXiv:1506.02640 [cs.CV].