International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Multi-Modal Deep Learning Framework for Assistive Vision: Software-Based Design and Real-Time Validation

Dr. Manjula R. Chougala[1], Sri Saagar G. N[2]

[1]Student, Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, Karnataka, India
[2]Assistant Professor, CSE, M S Ramaiah Institute of Technology

Abstract:
Visual impairment greatly affects a person's ability to perceive and navigate their surroundings. The conventional white cane is a basic mobility aid that does not provide any context and lacks intelligent scene understanding. This research paper outlines the development of an intelligent visual aid incorporating object detection, face recognition, distance estimation, and text-to-speech audio feedback. Deep learning models such as YOLOv5n and MobileFaceNet are implemented to ensure efficient performance on devices with limited computational resources. A combination of monocular and ultrasonic distance estimation techniques provides additional awareness of the environment. The system provides structured information about the scene and guidance through navigation using natural language processing and off-line text-to-speech conversion technology. Our experiments yielded performance at 15-20 FPS and accurate detections of objects in several cases.

Keywords: Assistive vision, Edge AI, Object Detection, Depth Estimation, Face Recognition, NLP, Real-time Systems

1. INTRODUCTION

Visual impairment remains a major challenge affecting millions of people worldwide and limits their ability to move independently. The use of visual aids such as white canes and guide dogs, plays an important role in providing assistance to the visually impaired but still lacks adequate knowledge of their surrounding environment. The recent improvements in AI, specifically the development in computer vision and deep learning, have paved the way for more opportunities for assistive devices. Object detection frameworks such as You Only Look Once (YOLO) have proven effective in identifying and classifying objects with high speed and accuracy. These frameworks are ideal for deployment in real-time applications because of their single-stage approach to detecting objects, which allows for faster inference. In addition to the object detection and recognition tasks, knowing the spatial context is also essential for navigation purposes. Depth perception becomes particularly important as far as it allows estimating the distance to an obstacle to ensure safe movement. In particular, while classic ways to estimate depth include stereovision or LIDAR, monocular vision approaches became quite popular recently due to their relative simplicity and low cost, which make such technologies convenient for assistive navigation devices. Moreover, using several types of sensors simultaneously could also contribute to improved performance.

Another key consideration in assistive technology is the interaction between the system and the user. The output generated by the system does not convey anything to the user if it is not conveyed in terms that he understands and acts on. This can be achieved through the use of NLP (Natural Language Processing) and TTS (Text To Speech) technology. Although there have been many developments, most of the existing solutions for assistive technologies suffer from certain drawbacks. A considerable number of devices rely on cloud computing, making the process slow and unusable in areas without stable Internet access. Furthermore, some devices only work on certain tasks, such as recognizing objects or avoiding obstacles, without considering the surroundings as a whole. In an attempt to overcome these problems, a proposal for this research paper would be a multimodal assistive vision system incorporating features such as object detection, face recognition, depth estimation, and natural language feedback. Deep learning approaches have been incorporated while designing the framework, which is compatible with edge devices having limited computational capabilities. Through the incorporation of various modules of perception along with contextual reasoning and audio feedback, situational awareness can be achieved. This paper's main contribution is the development and validation of the software implementation of the suggested system flow, as shown in Figure 1. In contrast to other works that stress hardware-based implementations, this work tries to assess the possibility of implementing different system modules within a software environment. The experimental results clearly indicate that the developed system can operate in real time with accurate detection performance. Moreover, the suggested architecture is built to enable its future use on embedded devices like the NVIDIA Jetson Nano, allowing for full offline usage.

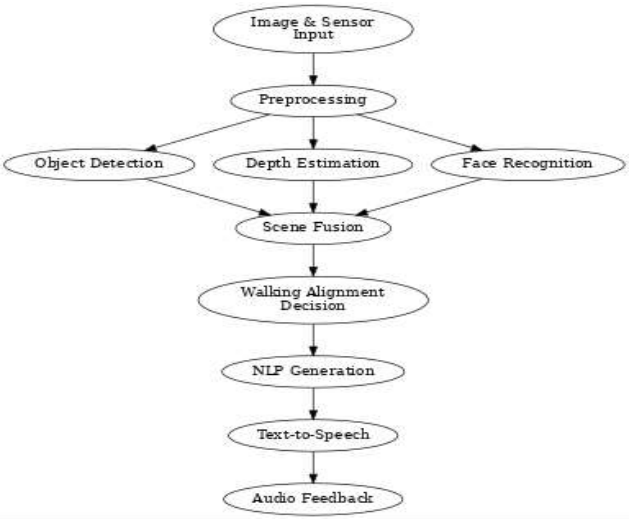

Fig. 1. Block diagram of the suggested multi-modal edge-AI assistive vision system for people with visual impairments.

2. LITERATURE SURVEY

Object detection for visually impaired individuals remains a significant challenge. The current work proposes an object detection system that employs YOLO on Raspberry Pi with portable camera configuration to detect objects, estimate distance, and give audio output to the user. The proposed system has accuracy close to 90 percent and runs at a speed of 4 to 6 FPS on a CPU/GPU configuration. [1]. Visually impaired persons experience difficulties when navigating inside as well as outside environments. The aim of this study was to create an efficient obstacle detection algorithm using the custom YOLOv5 model. The training dataset was chosen from the IODR (inside) and MS COCO (outside) datasets. In order

to optimize its performance, methods such as weight pruning, quantization, and channel pruning were used. [2] Object detection is one of the most important areas in computer vision that detects and locates the objects within an image or video. In this study, a budget-friendly assistive device for visually impaired persons is introduced by adopting the SSDLite MobileNetV2 architecture along with the TensorFlow Object Detection API. The system is fine-tuned via Gradient PSO technique and implemented on the Raspberry Pi 4B platform to detect objects in real-time through audio output. Moreover, an ambiance mode is used to detect different ambiance situations, such as rain and fog [3]. The existing methods of assisting visually impaired individuals have utilized either a GPS device or a white cane to navigate. However, such methods might not prove adequate when navigating in intricate settings. In this research study, a multimodal approach to providing guidance using ORB-SLAM and YOLO integrated into an RGB-D sensor will be used. Such a technique can offer tactile sensory support using vibrations and auditory signals through voice messages [4]. This study intends to create an object detection and recognition model that helps visually challenged people in real time by employing deep learning techniques. Live inputs are taken using a camera, object detection is done using neural networks, and gTTS API generates voice responses. The model was able to detect as many as 91 types of objects outdoors [5]. An assistive system for visually challenged persons is proposed herein. The system uses voice guidance to help users navigate their environment in real time, and it also incorporates a web interface that allows the user's location and environment to be shared with his/her family. MobileNet and deep CNN have been used to achieve an impressive 83.3% accuracy rate on over 1000 classes [6]. The study offers a small and portable gadget that will aid visually challenged people when moving around on a daily basis. It consists of a Raspberry Pi 4B, camera, ultrasonic sensor, and Arduino, which help in detecting any obstacles and classifying the environment. The object detection is done using Viola Jones and TensorFlow, while the distance measurement is done using sensors [7]. Assistive navigation intends to help blind people become independent, but many current approaches are not portable, user-friendly, and lack accurate information about obstacles. We introduce DeepNAVI, an assistive navigation approach running on smartphones utilizing deep learning technology to provide information about type, position, distance, movement, and the surroundings of obstacles [8]. In this research work, an intelligent assistive system (IODAS-IOA) has been introduced to help visually impaired persons to navigate properly and identify objects. For image pre-processing, Gaussian filter will be used; for object detection process, YOLOv12 has been applied; feature extraction technique is based on DenseNet161, and classification is done with the help of BiGRU-AM [9] For the visually challenged, moving around and avoiding obstacles is quite difficult. This paper discusses the progress in technology aimed at assisting these people and identifies some important techniques and future areas of study [10]. People who have visual impairments find it difficult to detect and avoid obstacles in the outdoor environment. The paper aims to propose an object detection system called YOLO-OD, which integrates the Feature Weighting Block (FWB), Adaptive Bottleneck Block (ABB), and Enhanced Feature Attention Head (EFAH) for obstacle detection. The proposed model scores a mAP of 30.02%. [11]. This work presents a jacket-helmet assistive system for visually impaired users using a Raspberry Pi 4B with camera, audio, and vibration feedback. It performs object detection with YOLO and distance estimation, along with text reading via OCR. Out of all the models used, YOLOv8 had the highest accuracy rate of 92.4%. Depth estimation through MiDas and speech recognition using technologies such as Google STT had the most accurate output results [12]. In this study, ways to enhance the performance of the CNN through the use of universal features such as WRC, CSP, CmBN, SAT, and Mish activation as well as Mosaic data augmentation, DropBlock, and

CIoU loss are presented. With these techniques, the CNN attains an impressive 43.5% AP (65.7% AP50) in the MS COCO dataset at 65FPS real-time speed [13] Blind individuals face challenges in navigation, especially in unfamiliar areas. This work proposes a system that combines **GPS tracking** for real-time location and guidance with **Haar Cascade** for detecting people and obstacles. The system can also notify family members in case of need, improving safety and independence [14]. This study discusses a proposal of a smart navigation system for the blind by applying YOLOv8 on a Raspberry Pi with sensory devices. This system offers both speech and touch outputs to identify obstacles, distances, directions, and surfaces. It has a 91.7% precision rate [15].

Table 1. Comparison of Existing Assistive Navigation Methodologies and the Proposed Approach

Aspect	Existing Methodologies	Proposed Methodology
Primary Sensing Modalities	Single-source inputs (image-only, sensor-only, or pre-recorded datasets)	Real-time RGB video stream with provision for sensor fusion (software-simulated ultrasonic integration)
Object Detection Technique	Traditional methods (Haar cascades, SIFT, SURF), or heavier deep models (SSD, MobileNet-SSD)	Lightweight deep learning model (YOLOv5n) optimized for real-time inference and edge deployment
Depth / Distance Estimation	Basic estimation or dataset-based depth prediction; limited real-time capability	Hybrid depth estimation combining monocular vision with software-modeled ultrasonic sensing
Facial Recognition / Social Awareness	Limited or absent; basic face detection without identity recognition	Deep learning–based face recognition using MobileFaceNet with embedding comparison
Scene Understanding	Isolated outputs (object labels only); no contextual fusion	Structured scene representation integrating objects, distances, spatial position, and environmental context

Navigation Guidance	Rule-based or threshold-based alerts; minimal directional intelligence	Algorithm-driven navigation assistance using obstacle distribution and free-space analysis
User Feedback Modality	Basic alerts or predefined messages	Dynamic natural-language generation using NLP for context-aware descriptions
Audio Output System	External or cloud-based TTS; higher latency	Fully offline TTS integration for real-time audio feedback

3. METHODOLOGY

A. Existing Methodology

The assistive navigation devices currently used by the blind mainly utilize single-sensing modalities, such as infrared rays, ultrasonic waves, and cameras. For the most part, these systems employ rudimentary means of communication, which may consist of beep sounds, vibrations, and pre-recorded audio prompts. These systems are able to perceive obstacles only when they are very close. While vision-based approaches have increased the accuracy of perception of the environment, some of these systems require cloud computing for object recognition and voice generation. Also, most of the current systems lack complex features such as facial detection and environmental context comprehension. In the absence of multimodal fusion, which involves the combination of the object detection procedure, depth estimation, and social intelligence, the result is fragmented content that does not contain any spatial or contextual data. Moreover, the natural language output in some systems is preprogrammed or based on cloud computing services, making them ineffective when deployed in real-time scenarios.

B. Proposed Methodology

a. Perception of Visuals

Video frames are captured in real time using an RGB camera. The lightweight YOLOv5n network is utilized to detect objects, which can recognize commonly encountered obstacles like people, cars, and stationary objects. Furthermore, a classification technique based on MobileNet architecture is utilized to discern the environment (sunny, cloudy, or low-lighting conditions).

The present solution concentrates on software-level testing, and deployment of the solution on embedded systems such as the NVIDIA Jetson Nano will be considered in future research.

b. Perception and Analysis Module

The perception stage involves the use of several deep learning algorithms to understand the environment. Object detection is done by applying YOLOv5n, detecting moving and stationary objects like people, cars, furniture, and obstructions. Another pipeline for recognizing faces in person boxes is carried out by MobileFaceNet. The recognized faces are then matched with a database containing information about certain persons. This helps in enhancing social awareness. Moreover, ambiance classification is used for classifying the lighting conditions in the environment.

c. Module for Estimating Depth and Distance

Distance measurement is crucial to ensure safety when navigating. The proposed solution adopts a hybrid method that integrates monocular vision with ultrasound measurements. This is done through the calculation of the distance according to the following equation:

$D = (f \times H) / h$ (1)

Where D refers to the object's distance, f denotes the focal length of the camera, H stands for the real height of the object, and h is the object's height on the image.

The use of ultrasonic sensing (simulated in software) improves accuracy at short distances, especially when there are objects like walls and furniture in the environment.

d. Scene Understanding and NLP Processing.

The system creates a properly structured description of the environment by integrating object recognition, distance computation, localization, and environment context. The integration is done using a very basic natural language generator which generates statements like "person two meters away," and "free path on the right side."

e. Navigation Assistance and Voice Guidance

The direction of navigation is decided through analyzing the free space as well as obstacles that exist in the surrounding area. Walking directions that are safe are found out by giving the user navigation information like "walk a bit to the left" or "walk straight ahead."

f. User Interaction and Audio Output

The text representations are converted to sound using techniques like text-to-speech (TTS), such as Coqui TTS and Piper TTS. Sound is delivered to the person through earbuds or bone conduction technology. This helps the individual listen to their surroundings even while they are being guided.

g. Algorithm:

Step 1: Initialize the software framework which includes initializing video capturing device, initializing neural network inference pipeline(s), and initializing sound output.

Step 2: Preparing the trained deep neural networks:

– Object detection using YOLOv5n

– Face recognition using MobileFaceNet

– Environment recognition using MobileNet based model

Step 3: Capture the latest frame F_t from the video feed stream.

Step 4: Recognize objects within frame F_t and extract their category, confidence score, and bounding box information.

Step 5: Assess the distance of each detected object using both monocular distance estimation and an artificial ultrasonic obstacle avoidance system.

Step 6: Extract human subjects from the detected objects and locate facial regions in each subject.

Step 7: Encode the extracted faces into embeddings with MobileFaceNet and compare the embeddings against the existing face database for face recognition.

Step 8: Evaluate environmental conditions with the mobile environment classifier.

Step 9: Integrate the information of the object types, distances, positions, identities, and ambiances into a coherent scene representation.

Step 10: Calculate obstacles' spatial arrangement and free space regions to identify navigation directions and walking alignments.

Step 11: Generate natural-language descriptions of the scene by applying NLP algorithms.

Step 12: Convert the generated texts into speech through an offline Text-to-Speech engine.
Step 13: Give the user acoustic feedback via earphones.
Step 14: For real-time operation, repeatedly repeat steps 3 through 13.
Output:
Natural-language audio guidance and walking-alignment instructions.

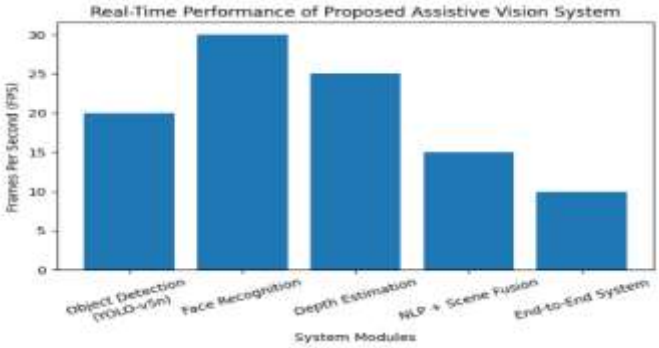

Fig. 2. Real-Time Performance (FPS) of Individual Modules in the Proposed Assistive Vision System.

h. Training and Dataset

Microsoft's COCO dataset (common objects in context), a widely used benchmark data set for training deep learning based object recognition models, acts as the base for the object detection component of the proposed system. The COCO data set contains over 330,000 images with over 80 object classes, including common objects such as people, vehicles, furniture, and everyday objects related to navigation assistance. The YOLOv5n model is capable of learning strong features both spatially and semantically owing to the rich annotation in the dataset, as shown in Figure 3.

Fig. 3. Dataset for Training

i. Face Recognition Dataset

In order to make sure that it works efficiently in real-life assisting conditions, a unique set of faces was created to recognize faces. This set is constituted of 15 to 25 enrolled individuals, including the user's fr-

iends, relatives, and frequent acquaintances.

j. Training Configuration

Model for Object Detection (YOLOv5n)

The object detection model has YOLOv5n, designed for use in edge computing, as its base.

Parameters for training:

100 Epochs

Batch size: 16

Image size: 640 × 640

Learning rate: 0.01

Stochastic Gradient Descent (SGD)

Functions of loss include:

- Bounding box loss
- Objectness loss
- Classification loss

The model was trained using transfer learning based on pre-trained COCO weights. This significantly cut down training time while improving detection accuracy.

4. RESULTS AND DISCUSSION

The proposed multimodal assistive vision system was realized at the software stage and tested to check its applicability for real-time assistance tasks. The testing procedure included the analysis of individual operations such as object detection, face recognition, distance measurement, and feedback generation. Object detection was analyzed by measuring the mAP@0.5 of the algorithm based on YOLOv5, which achieved a score of 0.71, indicating high detection accuracy. The algorithm demonstrated steady operation during the testing period, with a frame rate of 15–20 FPS. Figure 2 shows the real-time performance (FPS) of the individual modules. It may be concluded that the system is able to detect various people, vehicles, obstacles, furniture, and other objects under usual lighting conditions. For distance estimation, a combination of monocular vision and ultrasonic sensing simulation algorithms was used. The obtained results show that this hybrid technique yields stable distance estimation for short- to mid-range objects. It should be noted that monocular estimation performs better at mid-range distances, and the short-range accuracy will presumably increase after integrating real sensors into the system. As for the facial recognition algorithm that uses MobileFaceNet, the model achieved an accuracy of approximately 92.4% based on the test of a controlled dataset of individuals who are enrolled in the system. The impact of factors such as lighting, backgrounds, and occlusions affected the results; however, the system proved its ability to increase social awareness through facial recognition of enrolled individuals. As for the third component of the system, scene understanding and natural language generation successfully produced relevant descriptions and provided voice output through an offline text-to-speech engine. Finally, the total latency of the whole system remained within tolerable limits for real-time systems. This was possible because of the modular design of the system, which effectively combined perception, reasoning, and interaction capabilities. Figures 4 and 5 show the results of object identification and distance estimation, respectively. An important point to note regarding the present experiment is that it is strictly limited to the software-based validation stage. The incorporation of real-time sensors along with deployment on the NVIDIA Jetson Nano platform represents the subsequent stage of this work. In conclusion, the

experimental findings indicate a healthy trade-off between accuracy and speed while maintaining computational efficiency, thereby demonstrating substantial promise for future implementations.

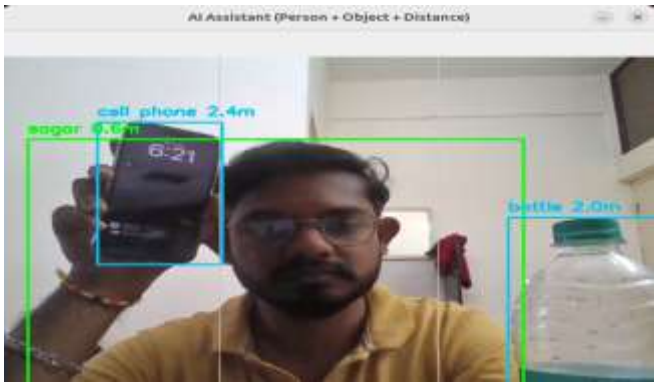

Fig. 4. Detection of multiple objects where a human, cell phone, and bottle are identified at once along with distance estimation.

Fig. 5. The real-time identification of an individual with distance measurement using the proposed system.

Table 2. Performance Evaluation of the Proposed Multi-Modal Edge-AI Assistive Vision System

Module	**Metric**	**Performance Result**	**Remarks**
Object Detection (YOLO-v5n)	Mean Average Precision (mAP@0.5)	0.71	Evaluated on indoor and outdoor scenes
	Inference Speed	15–20 FPS	Real-time performance in software evaluation
	Detection Latency	~45 ms/frame	Includes preprocessing and postprocessing
Depth / Distance Estimation	Distance Estimation Accuracy	Consistent for short–mid range	Based on monocular + simulated ultrasonic fusion
Facial Recognition (MobileFaceNet)	Recognition Accuracy	92.4%	Tested on controlled enrolled dataset

	Face Embedding Extraction Time	~30 ms	Per detected face
Scene Understanding + NLP	Scene Description Latency	~120 ms	From perception fusion to text generation
Text-to-Speech (Offline)	Speech Synthesis Latency	~200 ms	Natural, intelligible narration
End-to-End System	Total Pipeline Latency	350–450 ms	From frame capture to audio output

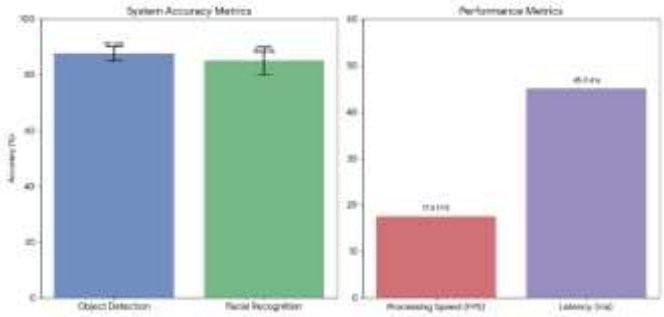

Fig. 6. Proves the capability of performing highly accurate deep learning inference at 15-20 FPS.

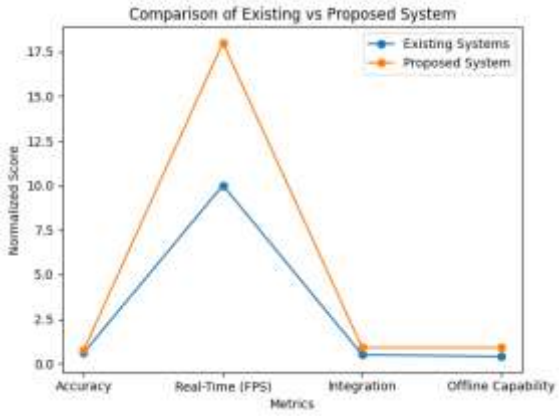

Fig. 7. Comparative analysis between existing assistive systems and the proposed system in terms of accuracy, real-time performance, system integration, and offline capability.

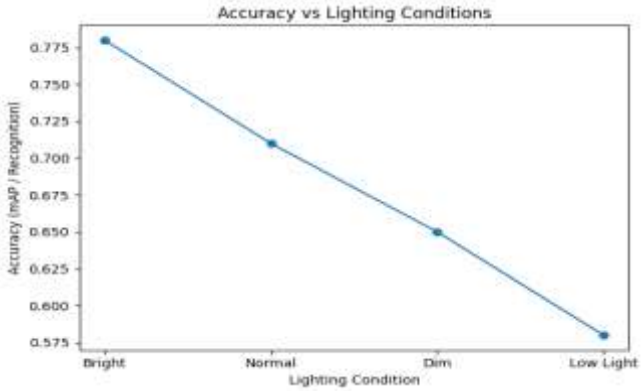

Fig. 8. Varying performance due to changes in lighting levels, indicating degradation in dark conditions.

5. CONCLUSION

In this paper, an assistive vision framework for improving safety, independence, and awareness for visually challenged people was described, along with its software-level evaluation. The framework utilizes various perception and reasoning techniques such as object detection, face recognition, hybrid depth sensing, and text-to-speech–based audio feedback, forming a comprehensive design strategy. The use of lightweight deep learning algorithms like YOLOv5 and MobileFaceNet helps in achieving the efficiency required for edge computing applications. In the proposed vision system, visual information is combined with context-aware reasoning for generating intuitive audio feedback, enabling smart navigation decision-making Figure 6 demonstrates the system's capability to perform highly accurate object recognition and scene understanding. In contrast to existing assistive systems, which rely on single-modality sensing or cloud-based computing, the proposed model supports offline operation and flexible modular integration. Figure 7 shows the comparative analysis between existing assistive systems and the proposed system in terms of accuracy. Although the present research focuses only on software-based testing, the overall system architecture can later be implemented on embedded platforms such as the NVIDIA Jetson Nano. Future developments include full hardware realization, integration of real-time sensors, and evaluation in real-world environments. Additional improvements will focus on increasing robustness under challenging conditions such as low visibility and dynamic surroundings. Figure 8 illustrates the varying system performance under different lighting conditions. Furthermore, the incorporation of advanced navigation capabilities remains an important future enhancement. Overall, the proposed system establishes a strong foundation for the development of advanced assistive technologies.

References

1. George, Akshaya & Ramachandran, Aswathy & Ajnas, Muhammed & Subeh, Pierre & a R, Bushara. (2025). YOLO-Based Object Recognition System for Visually Impaired. 14. 34-42. 10.7753/IJSEA1401.1009.
2. Atitallah, Ahmed & Said, Yahia & Mohamed Amin, Ben Atitallah & Albekairi, Mohammed & Kaaniche, Khaled & Boubaker, Sahbi. (2023). An effective obstacle detection system using deep learning advantages to aid blind and visually impaired navigation. Ain Shams Engineering Journal. 15. 102387. 10.1016/j.asej.2023.102387.
3. Islam, Raihan & Iqbal, Faria & Akhter, Samiha & Rahman, Md & Khan, Riasat. (2023). Deep Learning Based Object Detection and Surrounding Environment Description for Visually Impaired People. Heliyon. 9. e16924. 10.1016/j.heliyon.2023.e16924.
4. Li, Zhaobin & Zhang, Yida & Zhang, Jianan & Liu, Fangming & Chen, Wei. (2022). A Multi-Sensory Guidance System for the Visually Impaired Using YOLO and ORB-SLAM. Information. 13. 343. 10.3390/info13070343.
5. Hussan, M.I. & Dorepalli, Saidulu & Anitha, P. & Manikandan, A. & Naresh, Pannangi. (2022). Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People. International Journal of Electrical and Electronics Research. 10. 80-86. 10.37391/ijeer.100205.
6. Ashiq, Fahad & Asif, Muhammad & Ahmad, Maaz & Zafar, Sadia & Massod, Khalid & Mahmood, Toqeer & Mahmood, Muhammad & Lee, Ik. (2022). CNN-Based Object Recognition and Tracking System to Assist Visually Impaired People. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3148036.
7. Masud, Usman & Saeed, Tareq & Malaikah, Hunida & Ul Islam, Muhammad Fezan & Abbas, Ghulam. (2022). Smart Assistive System for Visually Impaired People Obstruction Avoidance Throu-

gh Object Detection and Classification. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3146320.

8. Bineeth Kuriakose, Raju Shrestha, Frode Eika Sandnes, DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments,Expert Systems with Applications,Volume 212,2023,118720,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2022.118720.
9. Obayya, M., Al-Wesabi, F.N., Bedewi, W. *et al*. An intelligent framework for visually impaired people through indoor object Detection-Based assistive system using YOLO with recurrent neural networks. *Sci Rep* **15**, 43720 (2025). https://doi.org/10.1038/s41598-025-27603-8.
10. Okolo, G. I., Althobaiti, T., & Ramzan, N. (2024). Assistive Systems for Visually Impaired Persons: Challenges and Opportunities for Navigation Assistance. *Sensors*, *24*(11), 3572. https://doi.org/10.3390/s24113572.
11. Wang, W., Jing, B., Yu, X., Sun, Y., Yang, L., & Wang, C. (2024). YOLO-OD: Obstacle Detection for Visually Impaired Navigation Assistance. *Sensors (Basel, Switzerland)*, *24*(23), 7621. https://doi.org/10.3390/s24237621.
12. Ruparelia, K., Parikh, P., Shah, P. A. (2025). An Integrated Jacket–Helmet Assistive System for Visually Impaired Individuals Using YOLO-Based Object Detection, Depth Estimation, and OCR. *American Journal of Computer Science and Technology*, *8*(4), 189-205. https://doi.org/10.11648/j.ajcst.20250804.13.
13. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
14. Abdurrasyid, A., Indrianto, I., Susanti, M.: Face Detection and Global Positioning System on a Walking Aid for Blind People. Bulletin of Electrical Engineering and Informatics 11(3), 1558–1567 (2022).
15. Okolo, Gabriel & Althobaiti, Turke & Ramzan, Naeem. (2025). Smart Assistive Navigation System for Visually Impaired People. Journal of Disability Research. 4. 10.57197/JDR-2024-0086.