International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Smart Sight: Real Time Scene Captioning for Visually Impaired

Dr Brindha S[1], Ms. Rajeshwari T[2], Kavya Harini M[3], Jui Prerna R S[4], Bhavanthika Shri G S[5], Shrilekhaa M A[6], Srevarsha V P[7]

[1]*Head of the Department, Computer Networking, PSG Polytechnic College, Coimbatore*

[2]*Lecturer, Computer Networking, PSG Polytechnic College, Coimbatore*

[3,4,5,6,7] *Students, Computer Networking, PSG Polytechnic College, Coimbatore*

ABSTRACT:

Visual impairment significantly limits an individual's ability to perceive and interact with the surrounding environment, making independent navigation a daily challenge. To address this issue, this journal presents a vision-based assistive system designed to provide real-time environmental awareness and safety support for visually impaired users. The proposed system integrates a camera-based object detection module with distance sensing and audio feedback to convey meaningful information about nearby obstacles and objects.

A pre-trained YOLO (You Only Look Once) deep learning model is employed for efficient and accurate object detection due to its high detection speed and capability to identify multiple objects in a single frame. To enhance safety during navigation, an ultrasonic sensor continuously measures the distance to nearby obstacles and triggers alerts when objects are detected within a predefined range. Detected information is communicated to the user through audio output, ensuring hands-free operation and minimizing cognitive load.

In addition to environmental awareness, the system incorporates an emergency alert mechanism that allows the user to send their real-time location to a designated caregiver using a GPS and wireless communication module. This feature ensures immediate assistance during critical situations. The proposed system is designed to be portable, cost-effective, and user-friendly, making it suitable for real-world deployment. Experimental observations demonstrate that the system effectively improves situational awareness and enhances the independence and safety of visually impaired individuals.

Keywords— Intelligent Assistive Vision System; Deep Learning–Based Object Detection; Proximity-Aware Obstacle Sensing; Context-Aware Audio Guidance; Emergency Location Broadcasting; Smart Accessibility Device.

1. INTRODUCTION:

Visually impaired individuals often face significant difficulties in understanding their surroundings, identifying obstacles, and navigating safely in unfamiliar environments. Limited access to real-time environmental information increases dependency and raises safety concerns during daily activities. To address these challenges, this project introduces an intelligent vision-based assistance system designed to interpret the user's surroundings and provide meaningful feedback in real time[4]. By combining automated visual analysis with situational awareness features, the system can understand the environment, recognize various items in its field of view, and convey this information clearly through audio output. This enhances accessibility and supports individuals who require additional guidance while navigating everyday spaces.

Beyond visual interpretation, the system integrates safety-driven functionality to support the user during critical situations. It continuously monitors the immediate surroundings for potential obstacles, offers real-time awareness updates, and includes an emergency communication mechanism that can instantly share the user's current location with a trusted contact. This combination of environmental understanding and rapid alert capability makes the system a reliable companion for both daily use and emergency scenarios.

2. CONVENTIONAL ASSISTIVE SYSTEM:

Several assistive solutions have been developed to support visually impaired individuals in navigation and obstacle avoidance [2]. Among these, traditional aids focus primarily on providing basic physical support and limited environmental feedback. However, such solutions offer minimal situational awareness and lack intelligent interpretation of the surroundings.

Fig 2.1 A blind with a cane

As shown in Figure 2.1, the existing assistive system mainly relies on a conventional cane-based approach combined with basic proximity sensing. This system provides only distance-based alerts and does not identify or interpret surrounding objects, resulting in limited environmental understanding. In addition, emergency alert mechanisms in the existing system are poorly integrated and often require manual steps to share location details. The absence of automatic emergency communication reduces system reliability, particularly during critical situations.

3. ARCHITECTURE OF THE PROPOSED ASSISTIVE SYSTEM:

The proposed SMART SIGHT system is designed to provide intelligent environmental awareness by processing real-time visual data and delivering meaningful audio guidance to visually impaired users. The system interprets the surrounding environment using advanced object recognition techniques and converts the detected information into clear spoken feedback, enabling users to perceive their surroundings without visual input.

As shown in Figure 3.1, the system architecture is centered around the Raspberry Pi 4B, which functions as the main processing unit [5]. The camera module captures live video frames of the user's surroundings and transmits them to the processing unit for object detection and scene analysis. This real-time visual interpretation allows the system to identify objects present in the environment.

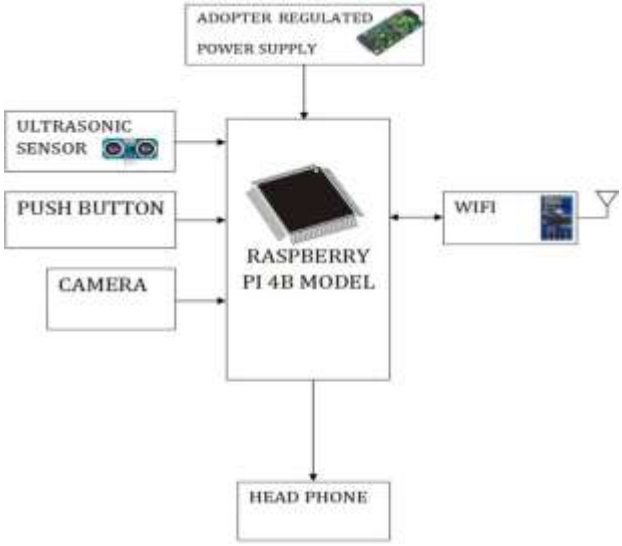

Fig 3.1 Block Diagram of the Smart Sight

To enhance navigation safety, the system continuously monitors the proximity of nearby obstacles using an ultrasonic distance measurement module. The distance information is updated in real time and combined with object detection results, providing layered awareness of both object identity and proximity. This integration supports safer movement in both indoor and outdoor environments.

An emergency alert mechanism is incorporated to improve user safety during critical situations. When the emergency push button is activated, the system retrieves the user's current geographical location through the location retrieval module. The location details are then transmitted to a predefined mobile number using a messaging service via Wi-Fi, ensuring rapid communication and timely assistance.

4. SYSTEM ARCHITECTURE

The system architecture integrates visual sensing, obstacle detection, audio feedback, and emergency communication into a unified framework. A camera module captures real-time video of the surroundings, which is processed by the central unit to perform object detection and scene analysis. Simultaneously, a distance measurement module monitors nearby obstacles and provides proximity data. The integration of object identification and distance information enables accurate hazard assessment and safer navigation [3]. The processed information is delivered as spoken feedback through an audio module, while an emergency trigger enables automatic location sharing with a predefined contact during critical situations.

5. WORKING MECHANISM OF THE SYSTEM

The working principle of the proposed SMART SIGHT system is based on real-time visual processing, obstacle detection, and audio-based guidance. As shown in Figure 5.1, the system begins with input sensing, where a camera module captures live video of the user's surroundings. The captured video is continuously divided into frames and forwarded to the central processing unit for analysis.

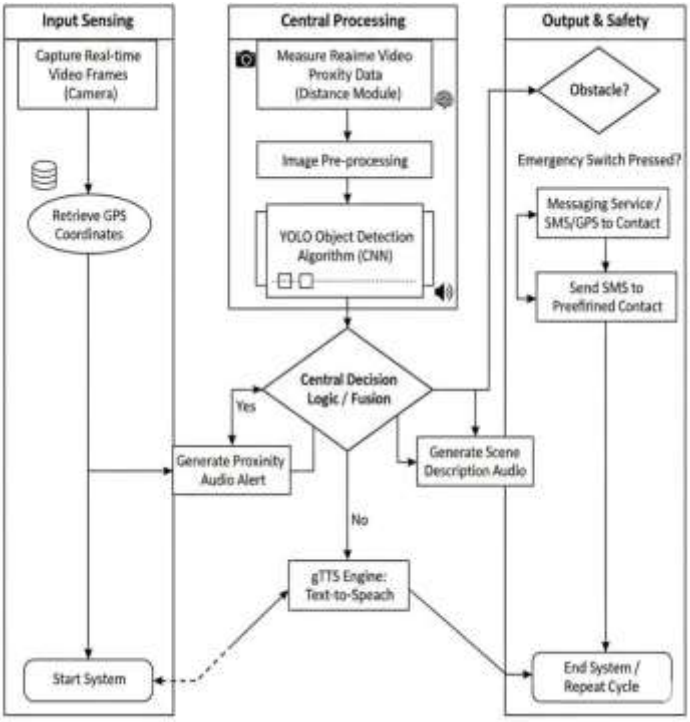

Fig 5.1 Workflow of Smart Sight

The processing unit performs image preprocessing and applies the YOLO (You Only Look Once) object detection algorithm

to identify and classify objects present in the scene. YOLO is selected due to its high processing speed and suitability for real-time applications. Multiple objects can be detected simultaneously,enabling quick understanding of the surrounding environment.

In parallel, the distance measurement module continuously monitors the proximity of nearby obstacles. The distance information is transmitted to the processing unit and fused with object detection results through a central decision logic [6]. This combined analysis helps determine potential hazards and triggers appropriate alerts.

Based on the decision logic, relevant information is converted into spoken output using the Google Text-to-Speech (gTTS) engine. Object descriptions and proximity warnings are delivered to the user through an audio output device, enabling navigation without visual dependence.

For safety enhancement, an emergency trigger switch is integrated into the system. When activated, the system retrieves the user's current geographical location and sends it to a predefined contact through a messaging service. This emergency workflow ensures rapid communication and timely assistance. Overall, the system integrates real-time object detection, distance monitoring, audio feedback, and emergency communication into a unified assistive framework.

6. COMPUTATIONAL TECHNIQUES APPLIED

The proposed SMART SIGHT system integrates multiple algorithms to achieve real-time scene understanding, obstacle awareness, audio feedback, and emergency communication. These algorithms are carefully selected to ensure accuracy, speed, and reliability while maintaining simplicity for real-world assistive use.

6.1. YOLO OBJECT DETECTION ALGORITHM

The system employs the You Only Look Once (YOLO) algorithm for real-time object detection. YOLO is a deep learning–based algorithm that detects objects by analyzing the entire image in a single forward pass through a neural network [4]. This approach significantly reduces processing time compared to traditional region-based detection methods, making YOLO suitable for real-time assistive applications.

YOLO divides the input image into a grid structure, where each grid cell predicts multiple bounding boxes along with

confidence scores and class probabilities. Each bounding box is represented as:

$$(x, y, w, h)$$

Where:

- x, yrepresent the center coordinates of the bounding box
- w, hrepresent the width and height of the bounding box

The confidence score of a detected object is calculated using:

$$\text{Confidence Score} = P(\text{Object}) \times IoU$$

Here, Intersection over Union (IoU) measures how accurately the predicted bounding box overlaps with the actual object:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Objects with confidence scores above a predefined threshold are considered valid detections. This filtering process reduces false detections and improves system reliability. YOLO is chosen because it can detect multiple objects simultaneously with high speed and acceptable accuracy, which is essential for guiding visually impaired users in dynamic environments.

6.2. DEEP LEARNING MODEL (CONVOLUTIONAL NEURAL NETWORK – CNN)

YOLO is built upon a Convolutional Neural Network (CNN) architecture, which automatically learns visual features from image data. CNNs process images through multiple layers, enabling hierarchical feature extraction.

The convolution operation used in CNNs is mathematically expressed as:

$$F(x, y) = \sum I(x + i, y + j) \times K(i, j)$$

Where:

- $I$is the input image
- $K$is the convolution kernel
- $F$is the resulting feature map

Early layers of the CNN detect simple features such as edges and textures, while deeper layers identify complex shapes and object patterns. Pooling layers reduce spatial dimensions and computational complexity, improving processing speed. This deep learning approach allows the system to perform robust object recognition even in varying lighting and background conditions.

6.3. DISTANCE MEASUREMENT ALGORITHM

To support safe navigation, the system uses a sensor-based distance measurement algorithm. The distance measurement module emits a signal toward nearby objects and measures the time taken for the signal to return after reflection.

The distance is calculated using the standard formula:

$$\text{Distance} = \frac{\text{Signal Speed} \times \text{Time Delay}}{2}$$

The division by two accounts for the round-trip travel of the signal. Continuous distance monitoring enables the system to identify nearby obstacles and generate warnings when objects are within a critical range. This algorithm enhances user safety by preventing collisions.

6.4. LOCATION RETRIEVAL AND EMERGENCY MESSAGING ALGORITHM

For emergency support, the system incorporates a location retrieval algorithm that obtains the user's geographical position in terms of latitude and longitude:

(Latitude, Longitude)

When the emergency trigger switch is activated, the system formats the location data into a readable message and transmits it to a predefined contact using a messaging service. This algorithm ensures accurate location sharing and enables quick response during emergency situations.

6.5. TEXT-TO-SPEECH ALGORITHM USING GTTS

The system uses Google Text-to-Speech (gTTS) to convert detected object information and distance alerts into spoken output. Once object detection and distance evaluation are completed, the system generates descriptive text messages such as object names and proximity warnings.

As shown fig 6.5 The gTTS algorithm processes the text through linguistic and phonetic analysis and generates a natural-sounding audio waveform. The conversion process includes:

1. Text normalization
2. Phoneme generation
3. Speech synthesis
4. Audio playback

This voice-based feedback allows the user to receive real-time guidance in a clear and understandable manner, eliminating the need for visual interaction.

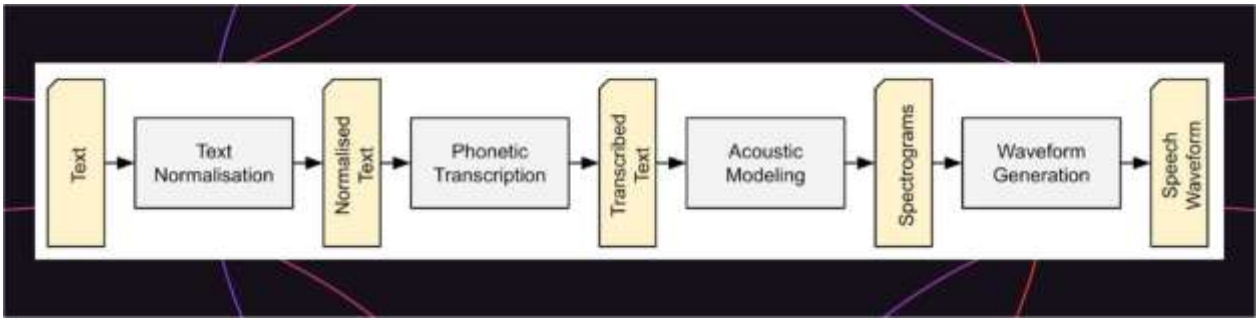

Fig 6.5 Working of GTTS

7. IMPLEMENTATION DETAILS

The implementation of the proposed SMART SIGHT system is carried out through the effective integration of software intelligence and supporting hardware components. The software implementation forms the core of the system and is responsible for visual processing, object detection, speech generation, decision making, and system testing. The hardware implementation provides the necessary interfaces for data acquisition, user interaction, and output delivery. Together, both implementations ensure real-time performance, reliability, and user safety.

The system is designed to operate continuously and respond dynamically to changes in the surrounding environment. Special emphasis is placed on real-time object detection using deep learning techniques, audio-based feedback for accessibility, and emergency handling. The overall implementation ensures that the system remains practical, responsive, and suitable for real-world usage by visually impaired individuals.

7.1 HARDWARE IMPLEMENTATION (CIRCUIT DESCRIPTION)

The hardware implementation of the proposed SMART SIGHT system is centered around the Raspberry Pi Model 4B, which functions as the main processing and control unit. As shown in Figure 7.1 and Figure 7.2, the complete hardware circuit comprises a regulated power supply section, Raspberry Pi board, ultrasonic sensor, camera module, Wi-Fi communication interface, emergency push button, and audio output device. The circuit is designed to ensure stable power delivery, accurate sensing, and reliable communication between all hardware components.

7.1.1 POWER SUPPLY SECTION

As shown in Figure 7.1, the power supply unit is designed to provide a stable 5 V DC output required for the Raspberry Pi and other connected peripherals. A step-down transformer is used to convert the AC mains supply into a lower AC voltage. This reduced voltage is then applied to a bridge rectifier, which converts the AC voltage into pulsating DC.

A smoothing capacitor is connected after the rectifier to reduce ripples and fluctuations in the DC output. The filtered DC voltage is regulated using a 7805 voltage regulator, which provides a constant 5 V output. An LED indicator along with a current-limiting resistor is connected at the output to indicate proper power supply operation. This regulated 5 V supply powers the Raspberry Pi and associated components, ensuring stable and safe system operation.

7.1.2 RASPBERRY PI AS CENTRAL CONTROLLER

The Raspberry Pi Model 4B serves as the central processing unit of the system.

As shown in Figure 7.2, it receives power from the regulated 5 V supply and executes all software-related operations, including object detection, distance processing, audio generation, and emergency handling. The GPIO pins of the Raspberry Pi are used to interface with the ultrasonic sensor and emergency push button, while dedicated ports support camera and communication modules.

The Raspberry Pi is selected due to its sufficient processing capability to handle deep learning inference, availability of multiple GPIO pins, built-in Wi-Fi support, and compatibility with camera modules.

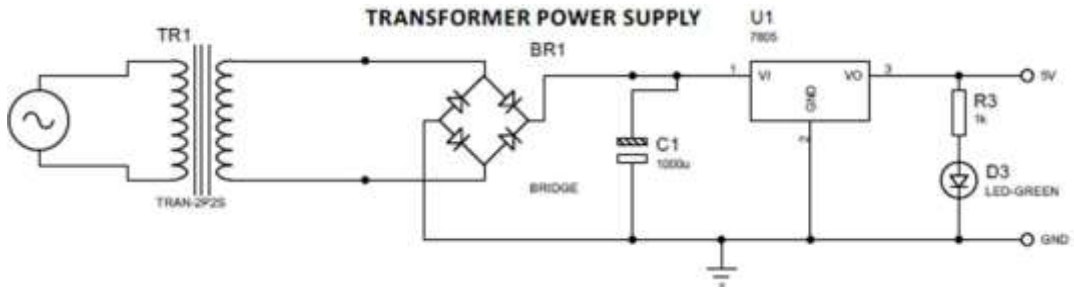

Fig 7.1 Power Supply Connection

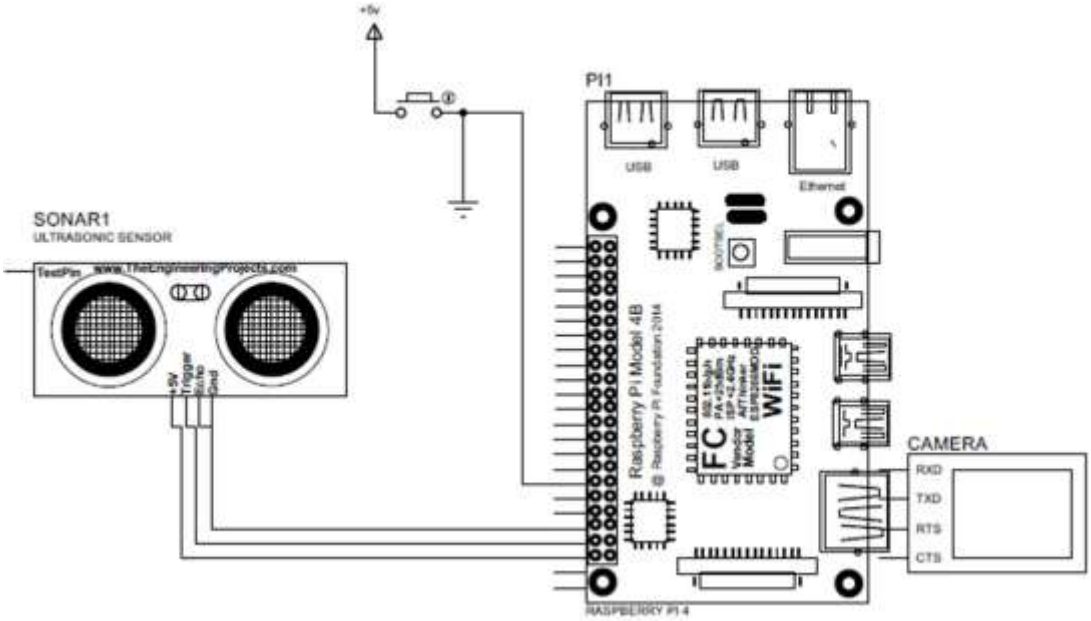

Fig 7.2 Central Controller

7.1.3 ULTRASONIC SENSOR INTERFACE

As illustrated in Figure 7.2, an ultrasonic sensor is interfaced with the Raspberry Pi to measure the distance between the user and nearby obstacles. The sensor consists of Trigger and Echo pins along with power and ground connections. The Trigger pin is connected to a GPIO pin of the Raspberry Pi to generate ultrasonic pulses, while the Echo pin is connected to another GPIO pin to receive the reflected signal.

The Raspberry Pi calculates the time difference between the transmitted and received signals to determine the distance of nearby obstacles. This distance information is continuously monitored and forwarded to the software module for decision making and generation of audio alerts.

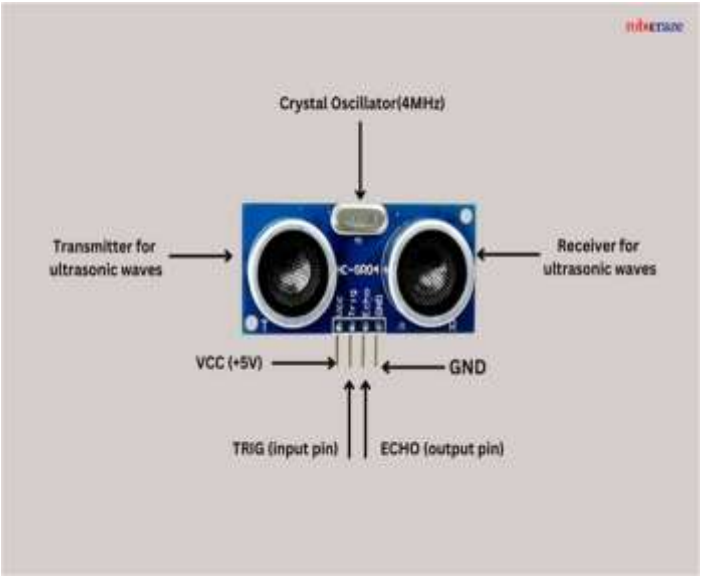

Fig 7.1.3 Ultrasonic Sensor

7.1.4 CAMERA MODULE CONNECTION

The camera module is connected directly to the Raspberry Pi through its dedicated camera interface port, as shown in Figure 7.2. The camera continuously captures live video of the user's surroundings. The video stream is accessed by the software running on the Raspberry Pi for real-time object detection using the YOLO deep learning model.

The dedicated camera interface ensures high-speed data transfer between the camera and the Raspberry Pi, enabling smooth real-time video processing without significant delay.

7.1.5 COMMUNICATION AND CONNECTIVITY

Wi-Fi connectivity is provided through the Raspberry Pi's built-in wireless module, as shown in Figure 7.2 . This connectivity enables the system to access online services required for emergency alert messaging. During emergency situations, the Raspberry Pi uses this network connection to transmit alert messages containing the user's current location to a predefined contact.

7.1.6 OVERALL CIRCUIT OPERATION

All hardware components share a common ground to maintain signal stability and ensure proper operation. The regulated power supply provides consistent voltage, while the Raspberry Pi coordinates data flow between the ultrasonic sensor, camera module, communication interface, and audio output device. As shown in Figures 7.1 and 7.2, the integrated circuit design enables seamless interaction between hardware and software modules, supporting real-time environmental awareness and enhanced safety for visually impaired users.

7.2 SOFTWARE IMPLEMENTATION

The software implementation represents the intelligence layer of the proposed system. It is developed using Python due to its simplicity, flexibility, and extensive support for computer vision and deep learning libraries. The software architecture is modular in nature, allowing independent execution of video processing, object detection, audio generation, and emergency handling modules.

7.2.1 VIDEO CAPTURE AND FRAME PROCESSING

The software begins by initializing the camera module using OpenCV. The camera continuously captures live video from the user's surroundings. This video stream is divided into individual frames to enable real-time processing [1]. Each frame is resized and formatted to match the input

requirements of the deep learning model. This preprocessing step ensures stable performance and consistent detection accuracy.

The frame-by-frame processing approach allows the system to continuously monitor the environment and respond immediately to changes, such as moving objects or new obstacles appearing in the user's path.

7.2.2 YOLO-BASED OBJECT DETECTION

For object detection, a pre-trained YOLO (You Only Look Once) deep learning model is used. YOLO is selected because of its high detection speed and ability to detect multiple objects simultaneously in a single forward pass. This makes it highly suitable for real-time assistive applications where delay must be minimized.

The YOLO model is initialized using its configuration file, trained weight file, and class label file. The model has been trained on standard object datasets, enabling it to recognize commonly encountered objects such as people, vehicles, furniture, and obstacles. Since a pre-trained model is used, the system does not require additional training, reducing complexity and computational cost.

Each processed video frame is passed through the YOLO network, which outputs bounding boxes, class labels, and confidence scores. Objects with confidence scores above a predefined threshold are selected as valid detections. This thresholding mechanism helps eliminate false detections and improves system reliability

Fig 7.2.1 Detected output

A pre-trained YOLO deep learning model is used for real-time object detection due to its high speed and ability to detect multiple objects in a single forward pass. As shown in Fig. 7.2.1, the detected output displays bounding boxes, class labels, and confidence scores, with valid objects selected based on a predefined threshold to improve detection reliability.

7.2.2 INTEGRATION OF DETECTION AND PROXIMITY INFORMATION

Along with visual detection, the system continuously receives distance information from the distance measurement sensor. This data is synchronized with the object detection output. By combining object identity with proximity data, the system is

Able to determine which detected objects pose a potential risk to the user.

For example, objects detected within a predefined safe distance are prioritized for alert generation. This integration ensures that the system does not overwhelm the user with unnecessary information and instead focuses on relevant and nearby obstacles.

7.2.3 TEXT GENERATION AND SPEECH OUTPUT

Once the detection and distance analysis are completed, the system generates meaningful textual descriptions. These descriptions include object names and proximity-related alerts that help the user understand the surrounding environment.

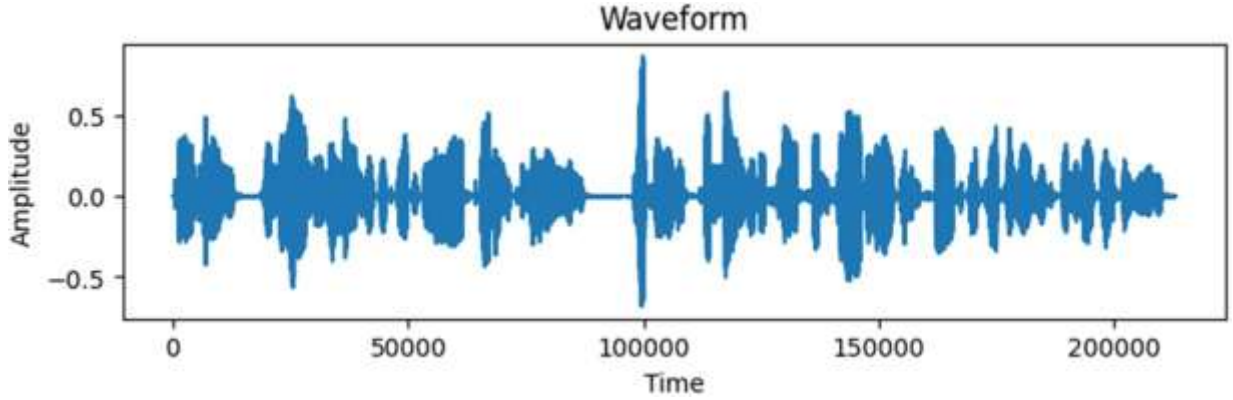

Fig 7.2.2 Speech Waveform of GTTS

The textual information is converted into spoken output using Google Text-to-Speech (gTTS). The gTTS engine transforms the generated text into natural-sounding speech, which is played through the audio output device. The speech output is designed to be clear, concise, and timely, ensuring effective communication with the user.

Audio feedback is generated dynamically and updated continuously as new objects are detected or as the user moves through different environments.

7.2.4 EMERGENCY HANDLING AND LOCATION SHARING

The software continuously monitors the emergency trigger input. When the emergency switch is activated, normal system operation is temporarily overridden. The system retrieves the user's current geographical location using the location module. The obtained coordinates are formatted into a readable alert message and transmitted to a predefined emergency contact through a messaging service.

This emergency routine is designed to operate quickly and reliably, ensuring timely assistance during critical situations.

7.2.5 TESTING AND VALIDATION

Extensive testing is carried out to validate the performance of the software implementation. The system is tested under various indoor and outdoor conditions to evaluate object detection accuracy, response time, audio clarity, and emergency alert reliability. Performance optimizations are applied to reduce delay between detection and audio output. The testing process confirms the system's stability and suitability for real-time deployment.

7. 3 RESULTS AND DISCUSSION

The experimental results demonstrate that the proposed SMART SIGHT system is capable of identifying objects in real time with reliable accuracy. The use of the YOLO-based deep learning model enables the system to detect multiple objects simultaneously from live video input. The detected objects are converted into clear audio descriptions, allowing the user to understand the surrounding environment without relying on visual cues. The real-time processing ensures minimal delay between object detection and audio feedback, which is essential for effective assistance.

The distance measurement module plays a significant role in enhancing user safety. Continuous distance monitoring allows the system to identify nearby obstacles and generate timely alerts when objects are within a critical range. This functionality helps users avoid collisions and move more confidently in both indoor and outdoor environments. The combination of object recognition with distance information provides more meaningful guidance compared to systems that rely on a single sensing mechanism.

The emergency alert feature is observed to function reliably during testing. When the

emergency trigger is activated, the system successfully retrieves the user's current location and transmits it to a predefined contact. This ensures rapid communication during critical situations and adds an important safety layer to the system. The integration of real-time object detection, obstacle awareness, and emergency communication significantly improves overall system effectiveness.

Finally, the results indicate that the proposed system enhances environmental awareness and supports independent navigation for visually impaired users. The system's performance demonstrates its practicality, reliability, and potential for real-world deployment as an assistive solution.

8. FUTURE ENHANCEMENTS

The proposed system can be further enhanced by integrating additional intelligent features. Face recognition can be included to help users identify familiar people in their surroundings. Currency detection can be added to assist users during financial transactions.

To improve reliability, offline emergency alert mechanisms such as SMS-based communication without internet dependency can be implemented. The system can also be extended with AI-based route guidance, enabling users to receive navigation assistance for reaching specific destinations.

In future versions, the hardware can be redesigned into a wearable form factor, such as smart glasses or a compact wearable device, to improve comfort and portability.

9. CONCLUSION

This project addressed the challenge faced by visually impaired individuals in understanding and navigating their surroundings independently. The proposed SMART SIGHT system successfully combines real-time object detection, distance awareness, audio feedback, and emergency communication into a single assistive solution.

Experimental results demonstrate that the system effectively identifies objects, alerts users about nearby obstacles, and provides clear voice guidance. The emergency alert feature further enhances user safety by enabling rapid communication during critical situations.

Overall, the system improves environmental awareness, supports independent mobility, and enhances the quality of life for visually impaired users. With further enhancements, the proposed solution has strong potential for real-world deployment.

REFERENCES

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
2. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.
3. D. Dakopoulos and N. Bourbakis, "Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 1, pp. 25–35, 2010.
4. S. R. Subramaniam et al., "Vision-Based Assistive Systems for the Visually Impaired," *International Journal of Assistive Technologies*, vol. 8, no. 2, pp. 45–52, 2019.
5. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems (NIPS)*, 2012.
6. Raspberry Pi Foundation, "Raspberry Pi 4 Model B Datasheet," 2019.
7. OpenCV Documentation, "Computer Vision Library," Available online.