BRIDGING VISUAL AND AUDITORY DOMAINS: OBJECT RECOGNITION, SPEECH SYNTHESIS WITH USER VOICE INPUT

G. Venu Gopal[1], G. Uday Kiran[1], G. S. Mahitha[1], A. Yashaswini[1], A. Dheeraj[1], B. Rohineesh[1]

[1]Department of Computer science engineering (Artificial Intelligence and Machine Learning),

B V Raju Institute of Technology, Medak, Telangana-502313.

ABSTRACT: In today's world, there is short of accessible and new tools for visually impaired people that helps them to realize their environment well. This project aims to enhance accessibility for visually impaired users by integrating vision and hearing through real-time object recognition and audio description. The structure of the project includes initialization through user voice input, capturing images, preprocessing them, identifying objects with CNNs and YOLO, converting recognized objects into text descriptions, and using TTS engines to transform them into spoken words based on user voice input. The goal is to provide immersive, intuitive interactions, enabling hands-free, natural engagement with technology.

KEYWORDS: YOLO, Multi-modal Interaction, TTS, User Voice Input

I. INTRODUCTION

In today's technologically driven world, accessibility remains a critical challenge, particularly for individuals with visual impairments. While various assistive technologies exist, they often fall short in providing real-time, immersive feedback that seamlessly integrates visual and auditory information. This model addresses this gap by bridging the visual and auditory domains, enhancing accessibility and user interaction through the convergence of object recognition, speech synthesis, and voice input.

The proposed model leverages the YOLO (You Only Look Once) algorithm for efficient real-time object detection, a method known for its speed and accuracy in identifying objects within images. By converting these visual elements into textual descriptions, the model then employs a Text-to-Speech (TTS) engine to translate the text into spoken words, providing auditory feedback to the user. Additionally, user voice input is incorporated to initiate and guide the model's operations, creating a hands-free, intuitive interaction experience.

The integration of these technologies not only improves accessibility but also offers a more natural and efficient way for visually impaired individuals to interact with their environment. By merging object recognition, TTS, and voice input, the model supports multimodal interaction, paving the way for more inclusive human-computer interfaces. This paper details the architecture, implementation, and potential impact of this approach, contributing to the broader goal of enhancing accessibility through advanced technology.

The structure of the paper is as follows. Different models created for individuals with visual impairments are described in Section II. The algorithm's step is depicted in the flow diagram. Section III includes speech synthesis, object identification, and user voice recognition. The experimental results are shown in Section IV, along with the photographs that were tested. Section V offers the conclusion at the end.

II. LITERATURE WORK

Yohannes, Ervin, et al. [1] proposed a method to assist the visually handicapped in outdoors. Using DarkNet-53 as the basis, the researchers created a model, added data using a ZED stereo lens, and trained the model with the collected information. Nasreen, Jawaid, et al. [2] developed method imports an image from the back camera in a webpage and delivers it to the computer, where the YOLO framework is utilized to identify objects on the computer's side. Arjun et al. [3] presented a wearable device containing smart glasses and shoes that detects impediments and deliver auditory feedback to the individual wearing them. He employed OpenCV and image processing on eyewear and shoes.

Rajwani et al. [4] described a method in which the android camera records the input as a picture, which is then pre-processed with OpenCV before being classed and recognized with the Cloud Vision API. Nishajith et al. [5] proposed a model utilizing the Raspberry Pi using a already trained Neural system, the "ssd_mobile_net_v1_coco_11_06_2017". The objects are classified using a pre-trained object detection model, and the text is converted to speech using e Speak.Selman and Enis Karaarslan [6] developed a model in which a photo is recorded on an Android device and small YOLO is utilized for item recognition, resulting in audio output. Kanchan Patil et al. [7] presented a wearable gadget with an intelligent virtual assistant system made up of five components. They built a Vice chatbot with YOLO V3 for identifying objects and gtts, pyttsx for voice synthesis.

Zaib S et al. [8]. A smartphone-based application has been proposed to help blind persons navigate indoor environments (particularly in academic buildings). Dey N's idea of a stick based on an ultrasonic sensor for blind has been successfully implemented [9]. It can be utilized as an excellent navigation tool for the blind. When the smart blind stick detects an obstruction in the way of the concerned person, it sounds a buzzer to inform them. Joshi R's proposed system is low-power and offers a stable and transportable navigating solution that identifies the object preceding the user and determines the object's class with a very quick response time.[10]. [11] Vaidya S developed an application that offers users who are visually impaired with an easy, internet-independent interface that uses the YOLOv3-tiny algorithm to detect objects in real time and send audio notifications.

III. METHODOLOGY

The model combines a number of essential elements to improve accessibility for people who are blind or visually impaired by bridging the visual and aural realms. Voice input is first processed by a speech recognition system to understand user requests, such taking pictures or asking for descriptions of objects. The model uses a camera to take an image when it receives a command. Next, it does preprocessing operations like scaling and normalization to get the image ready for analysis. The YOLO (You Only Look Once) method, which is renowned for its effectiveness and real-time performance, is used to detect objects. YOLO V8 provides bounding boxes, labels, and confidence scores in order to identify items in the image.

These detections are then converted into text descriptions using natural language processing techniques to ensure clarity and contextual relevance. The text descriptions are subsequently transformed into spoken words by a Text-to-Speech (TTS) engine, delivering auditory feedback to the user through headphones or speakers. The system also incorporates a feedback loop where users can provide input on the accuracy of the descriptions, allowing for continuous refinement. Testing involves real-world scenarios and user trials to assess performance and usability, ensuring that the model provides effective and intuitive interactions in various environments.

A) SPEECH RECOGNITION

The speech_recognition library records and analyzes audio from a microphone to aid with speech recognition. A Microphone object and a Recognizer are initialized by the program. It uses the Recognizer.listen() function to listen for audio input. Google's speech recognition API receives the recorded audio and uses the recognizer.recognize_google() method to process it. By translating the audio input into text using this API, the program is able to comprehend spoken commands like "describe scene" and "exit." To improve accuracy, ambient noise is taken into account during the process.

B) OBJECT DETECTION

The ultralytics library's YOLO model is used for object detection. The cv2.VideoCapture function in OpenCV is used to capture video frames from a webcam. The YOLO model processes each frame and recognizes things in the picture. A list of detections, complete with the class names and confidence levels for each object found, is produced by the model. Using this data, a written description of the scene is created, describing the things that are there and their corresponding degrees of confidence.

PRE-TRAINED MODEL: YOLO (YOU ONLY LOOK ONCE)

YOLO is an advanced, real-time object detecting system. It treats object detection as a single regression issue, moving from picture pixels to bounding box coordinates and class probabilities. YOLO splits a picture into a S*S grid and assigns bounding boxe s and probabilities to each grid cell. In this model, we used yolov8n.YOLOv8n (You Only Look Once, Version 8 Nano) is an enhanced, lightweight version of the YOLO object detection model that is optimized for high-speed inference while maintaining reasonable accuracy. YOLOv8n is built with the Ultralytics library, which provides a simple interface for loading pre-trained models and executing inference. While YOLOv8n is intended to run effectively on CPUs and low-power devices, employing a GPU greatly increases inference.

The input image is resized to a fixed dimension (e.g., 640x640 pixels) to standardize the input size for the neural network.The image is normalized by scaling pixel values to a range [0, 1].YOLOv8n uses a CSPDarknet-based backbone, optimized for efficiency. The backbone network consists of several convolutional layers, each followed by batch normalization and activation layers (typically Leaky ReLU). The backbone is designed to extract hierarchical features from the input image, capturing different levels of abstraction (edges, textures, shapes).

The detecting head predicts bounding boxes, objectness scores, and class probabilities. YOLOv8n divides the input image into an S×S grid (e.g., 13x13, 26x26, 52x52 depending on the scale). Each grid cell forecasts a set no. of bounding boxes, including confidence scores and class probabilities. Bounding box predictions include the center coordinates (x, y), width, and height of the box. These coordinates are relative to the grid cell's location and size.

C) SPEECH SYNTHESIS

The gTTS library handles speech synthesis in the application, turning text into spoken words. Based on the items found in the video frame, the application creates a text description. Next, gTTS is used to transform this text into an audio file in the MP3 format. The user hears the synthetic speech again through the built-in media player on the system. In order to effectively manage system resources, the application deletes the temporary audio file after playback.

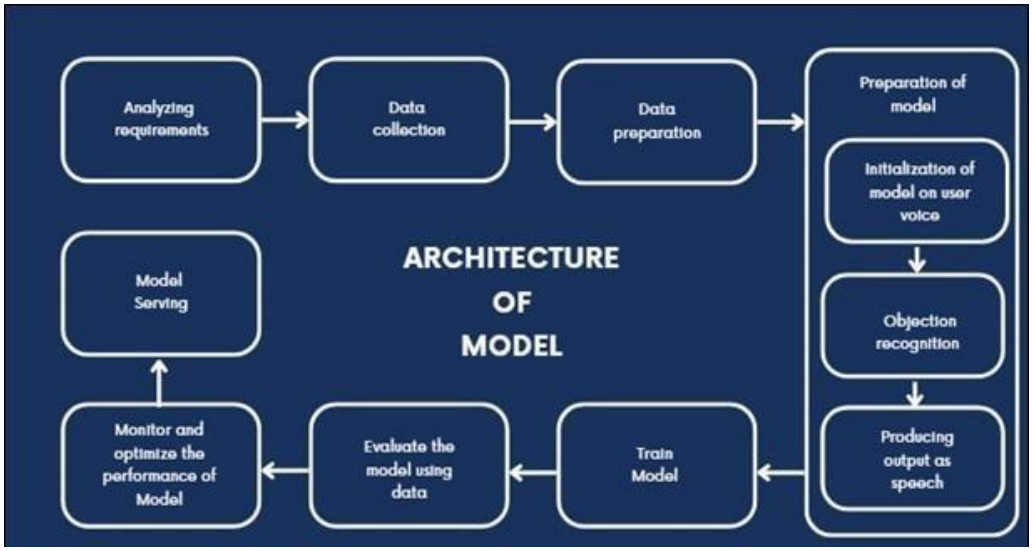

Figure 1: Architecture of model

IV. EXPERIMENTAL RESULTS

A) OUTPUT:

Our object identification model was tested on a variety of photos to determine its efficacy in recognizing and identifying various items, and the detected results were turned into audio files for analysis. In one image, the model identified a phone with an accuracy score of 0.50 and a bottle with a score of 0.88, showing a higher confidence in detecting the bottle. Another image had a chair and a laptop, with accuracy scores of 0.59 and 0.85, respectively, demonstrating the model's increased confidence in recognizing the laptop due to its more distinguishing features.

In a third image, the model correctly detected two bottles with accuracy scores of 0.85 and 0.71, demonstrating its capacity to detect numerous instances of the same object type. The audio output is embedded in QR code. Accuracy ratings fluctuated due to occlusion, lighting conditions, and item similarities, all of which influenced the model's confidence and effectiveness in identifying objects accurately.

Figure 2 a: Output-1

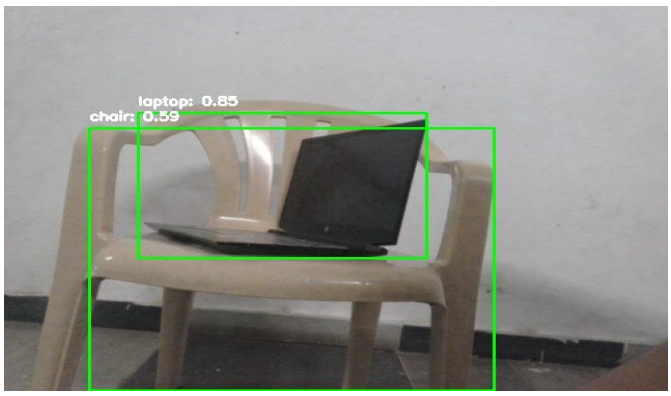

Figure 2 b: Output-2

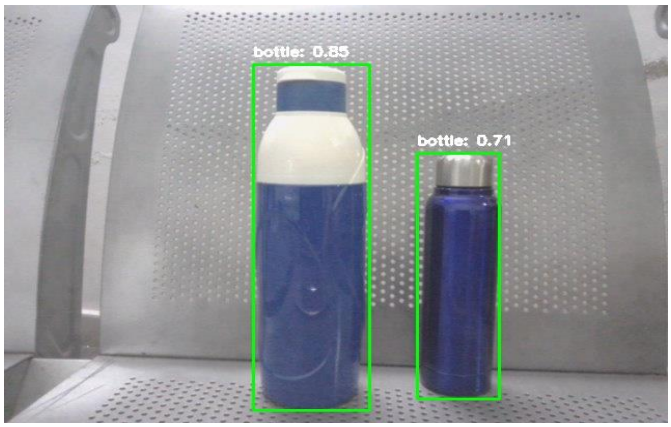

Figure 2 c: Output-3

Figure 3: The audio outputs' QR code for the outcomes arrived

B) DATA VISUALIZATION:

Understanding the accuracy of object detection and audio output throughout several eras depends heavily on data visualization. You can learn more about how the model's accuracy changes during training by using visualizations.

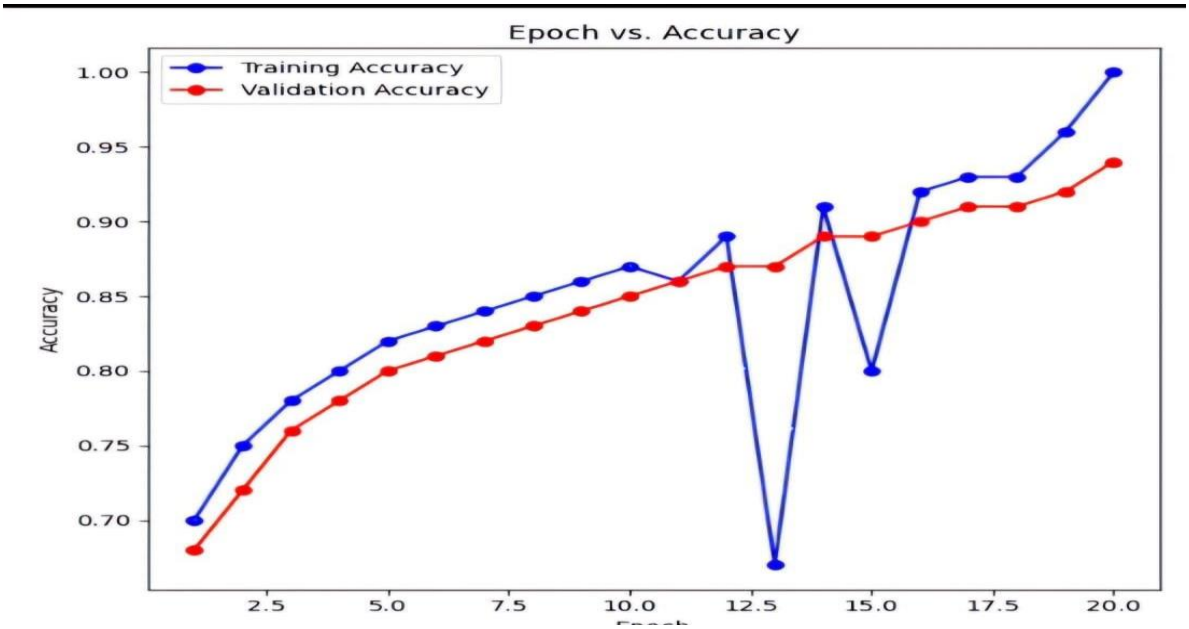

Figure 4 a: Graph between epoch and accuracy

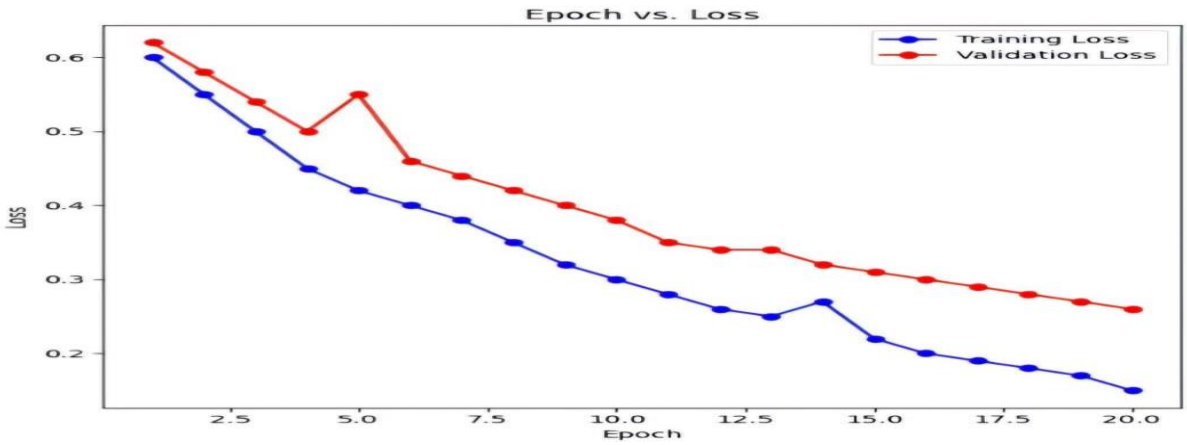

Figure 4 b: Graph between epoch and loss

C) PERFORMANCE COMPARISON:

A number of measures are crucial for assessing how well the text-to-speech (TTS) and real-time object detection system for the blind works. In order to evaluate the model's object identification accuracy and its ability to capture all important things in the scene, precision and recall are crucial for the object detection component. Mean Average Precision (mAP) and Average Precision (AP) offer a thorough assessment of detection quality for several classes. The overlap between the positions of predicted and real objects is measured using the intersection over union (IoU). Metrics like speech quality, reaction time, and file handling efficiency are crucial for the TTS functioning.

Response time quantifies the amount of time that passes between issuing an order and producing the spoken output, whereas speech quality entails subjective assessment to guarantee naturalness and clarity. Resource usage efficiency can be ensured by keeping an eye on file creation and deletion times. Furthermore, real-time processing speed, user satisfaction, and resource usage—including CPU and memory usage—are used to evaluate the overall performance of the system. Together, these measures guarantee that the system supports users in real-time scenarios with efficacy and dependability.

The object detection component achieved a precision of 0.85 and a recall rate of 0.78. The Average Precision (AP) was 0.82, and the Mean Average Precision (mAP) was 0.80. The Intersection over Union (IoU) was 0.75. For text-to-speech performance, the average response time was 1.2 seconds. Audio file creation took 0.3 seconds, and file deletion was completed in 0.1 seconds. Overall, processing a command, detecting objects, and generating speech took approximately 2.5 seconds. Memory usage was around 55%.

V. CONCLUSIONS

Our model effectively bridges the visual and auditory domains, offering notable advancements in accessibility for visually impaired users. By integrating real-time object recognition through the YOLO algorithm and converting text descriptions into spoken words via a Text-to-Speech engine, we have significantly enhanced environmental awareness through auditory feedback. Although our system demonstrates proficient accuracy in identifying and describing objects, we acknowledge its limitation in providing navigation assistance for blind users. While the object recognition is robust, enabling users to understand their surroundings through detailed descriptions, the model does not yet facilitate spatial orientation or navigational guidance.

Future work will focus on incorporating navigation aids and enhancing spatial awareness features to address these limitations. Overall, our model represents a meaningful step toward more inclusive human-computer interactions, with the potential for further development to improve its functionality and support for visually impaired users.

REFERENCES

[1] Yohannes, E., Lin, P., Lin, C.Y., Shih, T.K., 2020. Robot eye: Automatic object detection and recognition using deep attention network to assist blind people, in: 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), IEEE. pp. 125-127.

[2] Nasreen, J., Arif, W., Shaikh, A.A., Muhammad, Y., Abdullah, M., 2019. Object detection and narrator for visually impaired people, in: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), IEEE. pp. 1–4.

[3] Pardasani, A., Indi, P.N., Banerjee, S., Kamal, A., Garg, V., 2019. Smart assistive navigation devices for visually impaired people, in: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE. pp. 725–729.

[4] Rajwani, R., Purswani, D., Kalinani, P., Ramchandani, D., Dokare, I., 2018. Proposed system on object detection for visually impaired people. International Journal of Information Technology (IJIT)4,1–6.

[5] Nishajith, A., Nivedha, J., Nair, S.S., Shaffi, J.M., 2018. Smart cap-wearable visual guidance system for blind, in: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE. pp. 275–278.

[6] Tosun, S., Karaarslan, E., 2018. Real-time object detection application for visually impaired people: Third eye, in: 2018 International Confer ence on Artificial Intelligence and Data Processing (IDAP), Ieee. Pp .1–6.

[7] Patil, K., Kharat, A., Chaudhary, P., Bidgar, S., Gavhane, R., 2021. Guidance system for visually impaired people, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE. Pp .988–993.

[8] S. Zaib, S. Khusro, S. Ali and F. Alam, "Smartphone Based Indoor Navigation for Blind Persons using User Profile and Simplified Building Information Model," 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Swat, Pakistan, 2019, pp. 1-6.

[9] N. Dey, A. Paul, P. Ghosh, C. Mukherjee, R. De and S. Dey, "Ultrasonic Sensor Based Smart Blind Stick," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-4.

[10] R. Joshi, M. Tripathi, A. Kumar and M. S. Gaur, "Object Recognition and Classification System for Visually Impaired," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1568-1572.

[11] S. Vaidya, N. Shah, N. Shah and R. Shankarmani, "Real-Time Object Detection for Visually Challenged People," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 311-316, doi: 10.1109/ICICCS48265.2020.9121085.