SightSense: Enhancing Vision for the Blind, Where Sound Paints the World with Computer Vision and Voice Assistance

MRUNAL RAJIV CHAVAN[1], CHANDNI SANJAY RANA[2], PROF. POONAM DHARPAWAR[3]
[1, 2, 3] Department of Data Science Usha Mittal institute of Technology, Mumbai, India

Abstract— SightSense AI represents a groundbreaking innova- tion aimed at enhancing the independence of visually impaired individuals by providing advanced navigation capabilities. Lever- aging the power of YOLOv4 for object recognition, coupled with Python Text-to-Speech (Pyttsx3) technology, SightSense AI offers real-time auditory feedback to users, enabling them to navigate and understand their surroundings more effectively. Beyond basic object recognition, SightSense AI serves as a comprehensive travel companion, facilitating navigation and exploration of unfamiliar environments. It excels in assisting users in safely navigating busy intersections, reading signs, and identifying currency notes. The application's multifaceted functionalities significantly improve the independence and autonomy of visually impaired individuals, demonstrating the transformative potential of technology in transcending barriers and enhancing quality of life. The integration of Python Text-to-Speech technology ensures seamless communication of environmental information to users, further enhancing their ability to navigate independently. By bridging the gap between sound and vision, SightSense AI revolutionizes the way visually impaired individuals interact with their surroundings, ultimately promoting greater autonomy and inclusion in everyday life.

Index Terms- YOLOv4, Pyttsx3, navigating, currency

I. INTRODUCTION

In a world that relies heavily on visual information, the vi- sually impaired face unique challenges in their daily lives. The simple act of navigating and comprehending the world around them can be a daunting task. Recognizing these challenges and fueled by the desire for inclusivity and empowerment, we introduce ”SightSense.”

The heart of SightSense lies in its fusion of computer vision and voice assistance. By harnessing the formidable power of sophisticated computer vision algorithms, this application can interpret live video feeds from cameras or smartphones, transforming the visual world into meaningful, context-rich information. It accomplishes this by intelligently identifying and describing objects, deciphering text, and providing real- time auditory feedback through the acclaimed Pyttsx3 (Python Text-to-Speech) module.
However, SightSense goes beyond mere object recognition. It aspires to empower and enhance the independence of the visually impaired community. Serving as a reliable and inclusive travel companion, it assists in navigating unfamiliar environments, reading signs, and even identifying currency notes. Whether crossing busy intersections or deciphering critical information, SightSense becomes an invaluable tool in daily life scenarios.

In this project, we delve into the intricate blend of tech- nology and humanity, where SightSense transcends barriers, enabling visually impaired individuals to interact with and understand their world more effectively. It's a testament to the power of innovation and compassion, where technology becomes a means of empowerment and inclusion, enriching the lives of those who rely on it.

The visually impaired community grapples with ongoing and significant challenges in their daily lives, primarily due to the inherent complexities of understanding and navigating a visual world that many people often take for granted. In a society where a considerable amount of information is conveyed through visual means, this community encounters formidable barriers that not only impede their autonomy and safety but also restrict their full participation in various aspects of their environment.
Understanding the gravity of these challenges and motivated by a commitment to inclusivity and

empowerment, we have chosen to embark on the ”SightSense” project. This project addresses a pressing issue: the need to empower visually impaired individuals with an enriched visual experience enabled by cutting-edge technology.

II. PROPOSED SYSTEM ARCHITECTURE

1. Real-Time Video Analysis:

Data Capture: The core of SightSense lies in its ability to process live video streams in real-time. The application actively captures video feeds using the device’s camera. Data Analysis: These live video feeds are then transmitted to the program’s background for analysis. The program’s background is equipped with powerful algorithms that can swiftly and accurately process the incoming video frames. Object De- tection: Within the program’s background, the video frames undergo object detection using metrics provided by the COCO datasets object detection model. This model has been trained to recognize a wide variety of objects, from common everyday items to specific objects within various contexts.

2. Voice Assistance Integration:

Transforming Object Paths: After successful object detec- tion, SightSense incorporates voice modules into the frame- work. These modules are responsible for translating the paths and positions of the detected objects into standard voice notes. User-Friendly Voice Notes: The generated voice notes are thoughtfully designed to be easily understandable and im- mensely helpful for visually impaired individuals who rely on the application. This ensures that the information is conveyed in a clear and concise manner, enabling users to comprehend their surroundings effectively.

3. Assistance for the Visually Impaired: Delivering Essential Information: SightSense’s primary objective is to provide visually impaired users with critical information about their environment. The transformed voice notes serve as the means to achieve this. They are promptly delivered to blind users, enabling them to gain a better understanding of their surround- ings.

4. Alarm-Based Distance Estimation: Distance Calculation: In addition to object detection and assistance, SightSense implements an alarm system with the capability to calculate distances. It evaluates the proximity of the detected objects to the visually impaired person. Voice-Based Feedback: De- pending on the calculated distance, the framework generates voice-based feedback. For instance, if the user is very close to an object, the system will provide immediate and clear vocal alerts. Alternatively, if the user is in a safer area further away, the system communicates this information using relevant units, ensuring security and situational awareness.

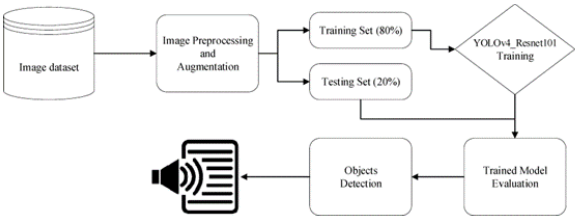

Fig. 1. Block diagram of the proposed work.

The block diagram describes the proposed system for Sight- Sense.

III. METHODOLOGY

The architecture of SightSense is designed to enable real- time object detection and voice assistance for visually im- paired individuals. This system leverages computer vision techniques and voice recognition to provide users with essen- tial information about their surroundings, thereby enhancing their independence and safety.

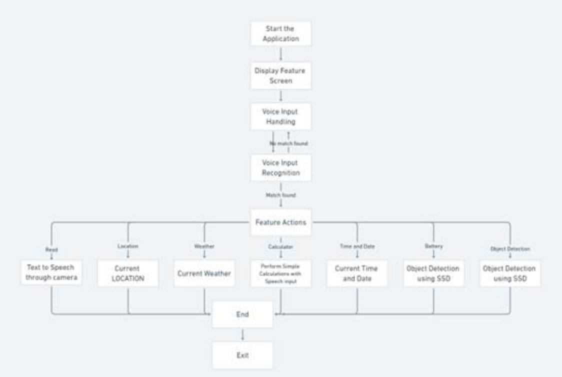

Fig. 2. The architecture of SightSense

1) Data Acquisition:: SightSense starts with the acquisition of data through a camera, typically integrated into a smartphone or an external webcam. This camera captures the live video feed of the user’s environment.
2) COCO-SSD:: COCO-SSD is an object detection model powered by the TensorFlow object detection API. Single-Shot MultiBox Detection (SSD) is an acronym for Single-Shot MultiBox Detection. In the COCO Dataset, this model can detect 90 different classes.
3) Real-time Video Analysis:: The captured video feed is sent to the system’s backend for analysis. This backend is responsible for processing the video frames in real-time.
4) Object Detection with YOLO:: Within the backend, SightSense employs the YOLO (You Only Look Once) object detection algorithm. YOLO is known for its efficiency and speed in detecting objects in real-time. YOLO identifies and localizes objects within the video frames, providing bounding boxes around each detected object.
5) YOLO: Provides a framework that allows detection of objects in near real time speeds. For deployment in a mobile device we are using Tiny YOLO, which is a lightweight YOLO framework for mobile and edge devices. YOLO, a state-of-the-art object detection algorithm, is employed for real-time visual recognition. Its speed and accuracy make it ideal for identifying objects, text, and environmental elements from live video feeds captured by cameras or smartphones. YOLO’s ability to process images quickly ensures that visually impaired users receive prompt and accurate information about their surroundings.
6) Object Classification: : Once objects are detected, the system classifies them into predefined categories (e.g., ”chair,” ”car,” ”person”) using the object’s features and characteristics. Confidence scores are assigned to each classification, indicating how confident the system is in its predictions.
7) ResNet (Residual Network):: ResNet is a deep convolutional neural network architecture designed to address the problem of vanishing gradients in very deep neural networks.

It introduces skip connections, also known as residual connections, which allow the gradient to flow directly through the network, mitigating the vanishing gradient problem.

maps are then utilized by additional convolutional layers to detect objects at multiple scales. SSD utilizes anchor boxes, which are pre-defined bounding boxes of various aspect ratios and scales, to predict the presence of objects and refine their bounding boxes. By combining classification scores and bounding box offsets, SSD effectively identifies objects while accounting for variations in size and aspect ratio. Finally, SSD applies non-maximum suppression to refine the final set of detected objects. Renowned for its efficiency and accuracy, SSD has found extensive use in real-time applications such as autonomous driving, surveillance, and robotics.

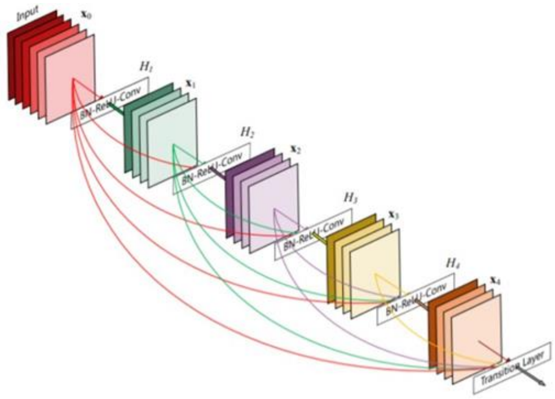

Fig. 3. Resnet Architecture

MobileNet

MobileNet is a family of neural network architectures designed for efficient on-device image classification, object detection, and other computer vision tasks. It was developed by re- searchers at Google, aiming to create lightweight models suitable for deployment on mobile and embedded devices with limited computational resources.

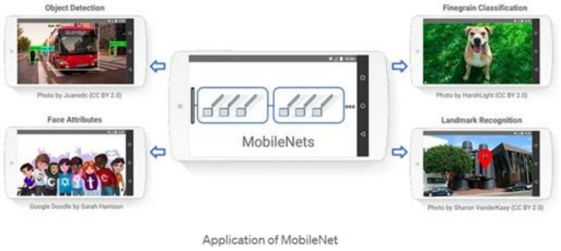
Object Detection
Photo by Juanedc (CC BY 2.0)
Face Attributes
Google Doodle by Sarah Harrison
MobileNets
Finegrain Classification
Photo by HarshLight (CC BY 2.0)
Landmark Recognition
Photo by Sharon VanderKaay (CC BY 2.0)
Application of MobileNet

Fig. 4. MobileNet Architecture

The key innovation in MobileNet architectures is the use of depthwise separable convolutions, which decouple standard convolution into two separate operations: depthwise convolu- tion and pointwise convolution. This separation significantly reduces the computational cost while maintaining reasonable accuracy in comparison to traditional convolutional neural networks (CNNs).

Single Shot MultiBox Detector (SSD): The Single Shot MultiBox Detector (SSD) is a state-of-the-art object detection algorithm designed for efficient and accurate detection of objects within images. SSD employs a base convolutional network, often based on architectures like VGG or ResNet, to extract feature maps from input images. These feature maps are then utilized by additional convolutional layers to detect objects at multiple scales. SSD utilizes anchor boxes, which are pre-defined bounding boxes of various aspect ratios and scales, to predict the presence of objects and refine their bounding boxes. By combining classification scores and bounding box offsets, SSD effectively identifies objects while accounting for variations in size and aspect ratio. Finally, SSD applies non-maximum suppression to refine the final set of detected objects. Renowned for its efficiency and accuracy, SSD has found extensive use in real-time applications such as autonomous driving, surveillance, and robotics.

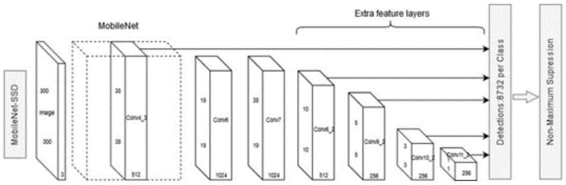

Fig. 5. SSD Architecture

5. Voice Assistance Integration:
After object detection and classification, the system inte- grates voice assistance modules. These modules are responsi- ble for converting the detected objects' paths and classifica- tions into voice notes.

Pyttsx3 (Python Text-to-Speech) : is integrated to provide the voice assistance component. Pyttsx3 converts the iden- tified visual information into descriptive auditory feedback, allowing users to hear clear and natural-sounding voice in- structions.Pyttsx3 is user-friendly, cross-platform, and offers customization options for voice, rate, and volume. Pyttsx3's robustness and multilingual support make it an excellent choice for delivering personalized voice assistance to visually impaired individuals.

6. Voice Feedback to the User: The transformed voice notes are delivered to visually impaired users in real-time. Users receive auditory feedback that describes the positions and movements of detected objects. The voice feedback is clear and user-friendly, designed to enhance understanding and navigation.

7. Alarm-Based Distance Estimation:
In addition to object detection and voice assistance, Sight- Sense implements an alarm system that calculates distance estimations. This system assesses the proximity of detected objects to the user. Depending on the calculated distance, the framework generates voice-based feedback. This feedback informs the user about the distance to nearby objects, enhancing security and situational awareness.

8. User Interface:
SightSense can include a user interface, which may be a mobile app or a web application. The interface allows users to interact with the system, adjust settings, and provide feedback.

III. RESULT AND ANALYSIS

The initial user interface (UI) element encountered upon launching the SightSense mobile application is a landing page. This page serves a dual purpose: information dissemination and user onboarding. A prerecorded voice message instructs the user to perform a leftward swipe gesture, leveraging the concept of kinesthetic learning to introduce the core function- alities of the application. This initial text-based and audio- prompted interaction exemplifies the human-machine interface (HMI) design principles employed within SightSense. It is optimized for intuitive user experience (UX) by incorporating both visual and auditory communication channels. This design decision caters to a diverse user base and fosters accessibility by accommodating users with varying learning preferences.

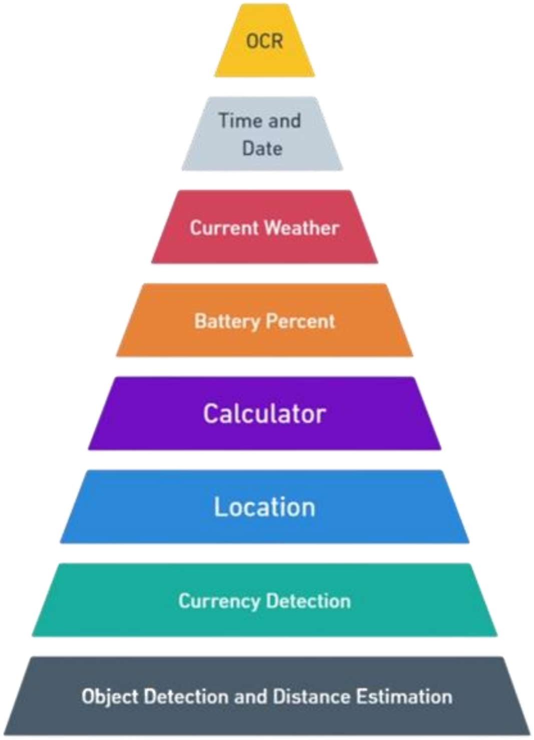

Fig. 6. Features of SightSense

The landing page describes functions as a tutorial to guide the user through the app's functionalities. Swiping left reveals a list that details these functionalities, which can all be controlled through voice commands. Here's a breakdown of the functionalities listed:

1) *Text Reading:* Double tapping the screen activates the microphone and initiates text-to-speech functionality by instructing the application to ”read”. This suggests the appli- cation integrates Optical Character Recognition (OCR) for text capture and conversion into an audio stream.
2) *Location Services:* The user can access their current lo- cation by uttering ”location”. This indicates the app leverages geolocation services to retrieve and display the user's physical coordinates.
3) *Weather Information:* The user can obtain real-time weather data by saying ”weather”. This suggests the appli- cation fetches real-time weather data through an integration with a weather API.
4) *Calculator:* Mathematical calculations can be performed by saying ”calculator”. This implies the application has a built-in calculator module.
5) *Date and Time:* The current date and time can be retrieved by saying ”time and date”. This suggests the applica- tion accesses the device's system clock to display the current date and time.
6) *Battery Status:* The user can inquire about the battery level by saying ”battery”. This functionality likely retrieves data from the device's battery management system.
7) *Application Termination:* The application can be closed by saying ”exit”. This indicates the application integrates with the device's application manager to terminate the SightSense SightSense is a comprehensive assistive technology application meticulously designed to cater to the needs of visually impaired individuals, offering a multifaceted approach to en- hance their environmental awareness and navigation abilities. At its core, SightSense leverages real-time video analysis ca- pabilities, harnessing the device's camera to capture live feeds seamlessly. These feeds are then subjected to rigorous analysis in the program's background, where sophisticated algorithms are deployed for object detection, drawing from the extensive COCO DATASETS for comprehensive recognition spanning from commonplace items to specific contextual objects.

Upon successful object detection, SightSense seamlessly integrates voice modules into its framework, transforming detected object paths and positions into intelligible voice notes. Crafted with utmost consideration for clarity and concise- ness, these voice notes serve as invaluable aids for visually impaired users, delivering essential information about their surroundings in a manner that is easily comprehensible and immensely helpful. This integration ensures users can effectively understand and interpret their environment, fostering a greater sense of independence and confidence in navigation. Furthermore, SightSense enhances navigation capabilities by audibly narrating the positions and movements of detected objects, empowering users to traverse both indoor and outdoor environments with greater ease and confidence.

Furthermore, SightSense incorporates an alarm-based distance estimation system, enhancing security and situational awareness for users. This system calculates the proximity of detected objects and delivers voice-based feedback accordingly, providing immediate vocal alerts or informative updates on safer distances. This feature ensures users are informed and aware of their surroundings, contributing to their overall safety and well-being.

SightSense represents a significant advancement in assistive technology, offering a holistic solution to the challenges faced by visually impaired individuals in navigating their environment. Through its seamless integration of real-time video analysis, voice assistance, enhanced navigation capabilities, and alarm-based distance estimation, SightSense empowers users with critical information and tools to navigate their surroundings with confidence, independence, and security.

IV. CONCLUSION AND FUTURE SCOPE

The SightSense project stands as a pioneering advancement in technology, specifically tailored to enhance the quality of life and foster independence among visually impaired indi- viduals. By seamlessly integrating real-time object detection, intelligent classification, and voice feedback functionalities, SightSense offers invaluable assistance in navigating and comprehending the surrounding environment. This innovative solution effectively bridges the gap between sound and vision, empowering users to interact confidently and safely with their surroundings. The cohesive integration of core components such as YOLO-based object detection and gTTS voice assistance en- sures an intuitive and user-friendly experience. Furthermore, the incorporation of alarm-based distance estimation adds an extra layer of situational awareness, further enhancing user safety.

Utilizing a Raspberry Pi, coupled with a camera module and leveraging machine learning techniques alongside software libraries like OpenCV and TensorFlow, enables efficient object detection of everyday items. The process involves gathering and preparing a dataset, training the machine learning model, and ultimately deploying the Raspberry Pi for real-time object detection and classification.

In conclusion, the SightSense project not only showcases the innovative potential of technology but also underscores its profound impact on improving the lives of visually impaired individuals. With its robust functionalities and user-centric de- sign, SightSense embodies a significant step towards fostering inclusivity and autonomy in navigating the modern world.

In addition to its technological advancements, the Sight- Sense project has the potential to spark broader societal shifts in attitudes towards disability and accessibility. By showcasing the capabilities and contributions of individuals with visual impairments, SightSense challenges stereotypes and promotes a more inclusive society. Through increased awareness and understanding, it paves the way for greater acceptance and support for people with disabilities in all facets of life, from education and employment to social interactions and public infrastructure.

Furthermore, the success of the SightSense project underscores the importance of interdisciplinary collaboration in addressing complex societal challenges. Its development required expertise from diverse fields, including computer science, engineering, psychology, and accessibility studies. As such, the project serves as a testament to the power of collab- oration and collective problem-solving in driving meaningful innovation. By fostering collaboration across disciplines and sectors, SightSense not only improves the lives of visually impaired individuals but also exemplifies a model for tackling pressing global issues through collective action and ingenuity.

REFERENCES

[1] J Ramesh Babu, Chandra Sekharaiah, and G Mahesh Kumar. A navigation tool for visually impaired persons. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 938–940, 2015.

[2] Ms. Manjeet Kaur Kavita. A survey paper for face recognition technolo- gies. In *Int J Sci Res Publ 6(7)*, pages 96–100, 2018.

[3] Nirmal A Kumar, Yazin Haris Thangal, and K Sunitha Beevi. Iot enabled navigation system for blind. In *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC) (47129)*, pages 186–189, 2019.

[4] Trupti Shah and Sangeeta Parshionikar. Efficient portable camera based text to speech converter for blind person. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 353–358, 2019.

[5] Sunit Vaidya, Naisha Shah, Niti Shah, and Radha Shankarmani. Real-time object detection for visually challenged people. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 311–316, 2020.