IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Smart Glasses For Blind People Using Raspberry Pi Zero 2W

[1]Nishant More, [2]Kanhaiya Chatla, [3]Shivam Daki, [4]Pravin Gholap, [5]Dr. Siddharth Hariharan
[1,2,3,4]Student, [5]Professor
[1,2,3,4,5]Department of Computer Engineering,
[1,2,3,4,5]Terna Engineering College, Navi Mumbai, India

Abstract: Visually impaired individuals face significant challenges in navigating complex environments and accessing real-time information about their surroundings. Current assistive technologies often fall short in providing the independence and ease of use required for daily activities. This project addresses these limitations through the development of a cutting-edge smart glasses system, specifically designed to enhance mobility and accessibility for the visually impaired. The proposed smart glasses integrate a high-resolution camera, a Raspberry Pi Zero 2 W, and audio output via earphones to offer real-time auditory feedback on the user's environment. The camera captures live visual data, which is processed using advanced models like Vision language Model

The Raspberry Pi Zero 2 W serves as the core processing unit, chosen for its compact size and adequate processing power, ensuring that the device remains lightweight and comfortable for prolonged use.

Index Terms: Raspberry Pi, Raspberry Pi Camera, Vision Language Model, Image Processing, Transformers**.**

I. INTRODUCTION

The proliferation of visual information in modern society has transformed how individuals interact with their surroundings. Yet, millions of people with visual impairments continue to face significant challenges in navigation and environmental awareness. According to the World Health Organization, over 2.2 billion individuals globally experience some form of vision impairment, with many depending on assistive technologies to improve their quality of life. This reality highlights an urgent need for innovative solutions that empower visually impaired individuals to engage with their environment more effectively and independently.

Recent advances in computer vision and portable computing technologies have paved the way for the development of smart glasses as a novel assistive tool. This research proposes a smart glasses system that integrates a high-resolution camera, a Raspberry Pi Zero 2 W, and audio output through earphones to convert visual information into actionable audio cues in real time. The system is designed not only to help users avoid obstacles and identify objects, thereby fostering a greater sense of autonomy. In addition to these core functionalities, the proposed smart glasses will feature multilingual response capabilities, enabling users from diverse linguistic backgrounds to receive assistance in their preferred language. The integration of voice command operations further enhances user interaction by allowing hands-free control; users can issue voice prompts to obtain information about their surroundings, inquire about detected objects, or switch between different functionalities, which improves overall usability and efficiency.

By merging cutting-edge hardware with user-centric design, this study aims to develop a tool that not only assists visually impaired users in navigating their environment but also empowers them to interact with it confidently and independently. For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of

SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

II. LITERATURE REVIEW

Recent advancements in assistive technology have driven the development of smart glasses aimed at enhancing the independence and safety of visually impaired individuals. Various studies have explored multi-functional designs that integrate computer vision, IoT, and machine learning to address the complex needs of these users.

i. G. Sudharshan et al. (2022) present a multi-functional smart glass system featuring three operational modes: seeing, reading, and writing [1]. In the seeing mode, the system detects up to 550 classes of objects, providing directional cues (left, right, or center) that enhance spatial awareness. The reading mode further augments user interaction by recognizing and vocalizing text, thereby facilitating real-time engagement with printed content. This work underscores the potential of a single device to offer diverse functionalities tailored to daily tasks.

ii. In contrast, S. Akalya et al. (2021) address not only the technical but also the social challenges encountered by visually impaired individuals [2]. Traditional aids such as Braille and walking sticks often contribute to social stigmatization and emotional stress. Their smart glasses solution is designed to mitigate these issues by offering discreet navigation and reading assistance, thus reducing the social and emotional burdens associated with visual impairment.

iii. Expanding the scope of assistive technology, Siddhant Salvi et al. (2021) leverage IoT and machine learning to develop a smart glasses system that assists individuals who are blind, deaf, or mute [3]. By integrating IP cameras and object detection algorithms, the system provides auditory feedback that enhances environmental awareness and facilitates safer navigation. This multi-modal approach is particularly significant for users with combined sensory impairments.

iv. Sayed et al. (2020) contribute an affordable smart glasses solution focusing on real-time object detection to assist with navigation [4]. While their system effectively provides continuous environmental updates, it also highlights a critical challenge: high power consumption. The authors emphasize the need for more efficient power management strategies to ensure that the device remains practical for prolonged daily use.

v. Earlier work by Abirami R. and Haarini S. (2017) proposes a customizable smart glasses solution that incorporates deep learning techniques, notably the YOLOv3 algorithm, for real-time video analysis [5]. Their system not only detects and identifies objects but also includes features such as face recognition and traffic light detection. This design prioritizes user safety and adaptability, demonstrating the potential of deep learning to support dynamic real-world applications.

III. WORKING OF SMART GLASSES

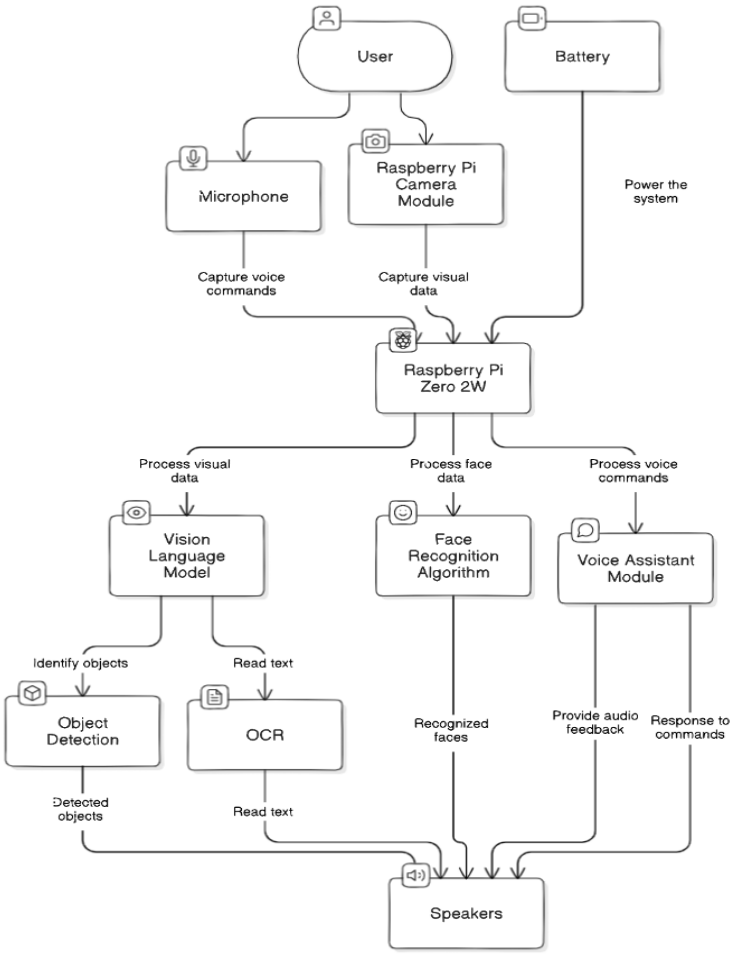

Fig. 1 Working of smart glass

The smart glasses system operates by continuously capturing real-time visual and auditory data, processing it using embedded machine learning algorithms, and delivering immediate, context-aware audio feedback to the user. The following outlines its detailed working principle:

i. Data Acquisition:
A high-resolution camera, integrated into the glasses, constantly captures the user's surroundings. Simultaneously, an onboard microphone records voice command. These inputs are essential for both environmental analysis and hands-free user control.

ii. Preprocessing:
Captured images are first preprocessed to ensure optimal quality for analysis. This involves resizing, denoising, and adjusting brightness and contrast to accommodate varying lighting conditions and improve the reliability of subsequent processing.

iii. Image Analysis via Vision Language Model (VLM):
The preprocessed images are forwarded to the Raspberry Pi Zero 2 W, which acts as the central processing unit. Here, a Vision Language Model (VLM) [6] performs multiple tasks:

a. Object Detection: Identifies and classifies objects within the scene. The system determines their spatial orientation (e.g., left, right, or center) and generates corresponding navigational cues.
b. Text Recognition (OCR): Extracts and converts any visible text into digital form, which is then prepared for audio conversion.
c. Face Recognition: Detects human faces and, if available, compares them against a pre-registered database to identify known individuals.

iv. Voice Command Processing:
The system incorporates a voice assistant module that interprets spoken commands from the user. This module facilitates hands-free operation by allowing the user to switch between different operational modes (object detection, text reading, or face recognition) and adjust system settings without manual input.

v. Audio Feedback Generation:
Once the visual data has been analysed, the processed information—whether it is object locations, textual content, or facial identities—is converted into audio cues using a text-to-speech engine. This real-time audio feedback guides the user through their environment, alerting them to obstacles, reading out text, or identifying familiar faces.

vi. System Integration and Data Flow:
The seamless integration between hardware and software is central to the system's functionality. The Raspberry Pi Zero 2 W coordinates the flow of data from the camera and microphone through the preprocessing unit, VLM, and voice assistant module before delivering the final audio output through earphones. For computationally intensive tasks, the system can also interface with cloud services, ensuring robust performance without compromising real-time responsiveness.

IV. ARCHITECTURE

The proposed smart glasses system is designed as a wearable, real-time assistive device for visually impaired individuals. It seamlessly integrates hardware and software components to capture, process, and deliver environmental information in the form of audio feedback. The following subsections detail the system's architecture.

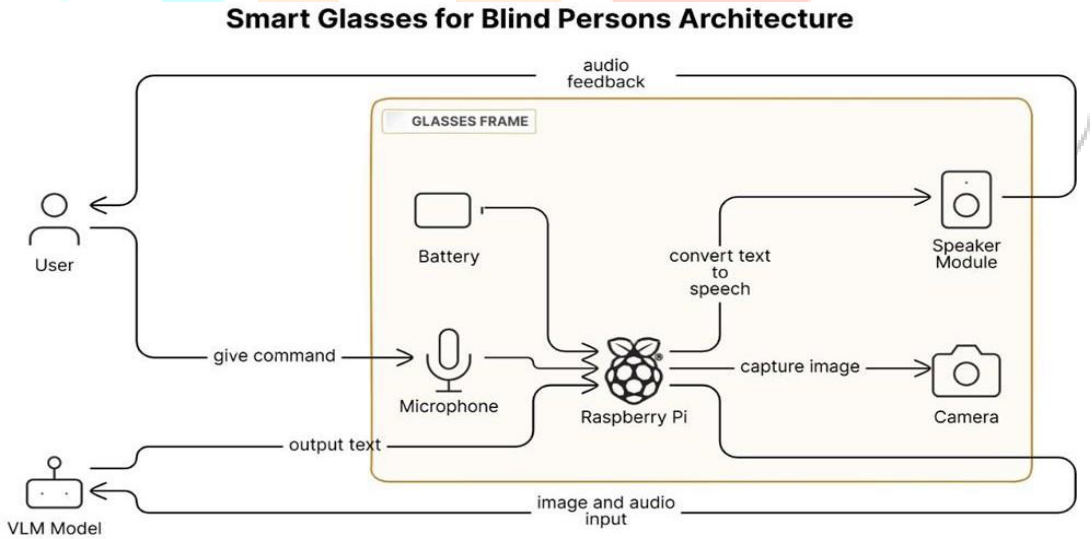

Fig. 2 Architecture diagram of smart glass

❖ Hardware Architecture

i. High-Resolution Camera: A dedicated camera module continuously captures images of the user's surroundings. Mounted on the glasses frame, this component provides the primary visual input for object detection, text recognition, and face identification.

ii. Raspberry Pi Zero 2 W: Serving as the central processing unit, the Raspberry Pi Zero 2 W coordinates the data flow between input devices and software modules. Its compact size, low power consumption, and sufficient processing capability make it ideal for executing real-time computer vision and machine learning tasks.

iii. Microphone and Speaker Module: The microphone records voice commands, enabling hands-free control and mode selection. Processed information—such as detected objects or recognized text—is converted into speech via the speaker, which delivers immediate audio feedback to the user

.

iv. Power Supply and 3D Printed Frame: A lightweight, rechargeable LiPo battery provides portable power to all components. A custom-designed 3D printed frame houses the Raspberry Pi, camera, microphone, battery, and speaker, ensuring a comfortable and discreet wearable design.

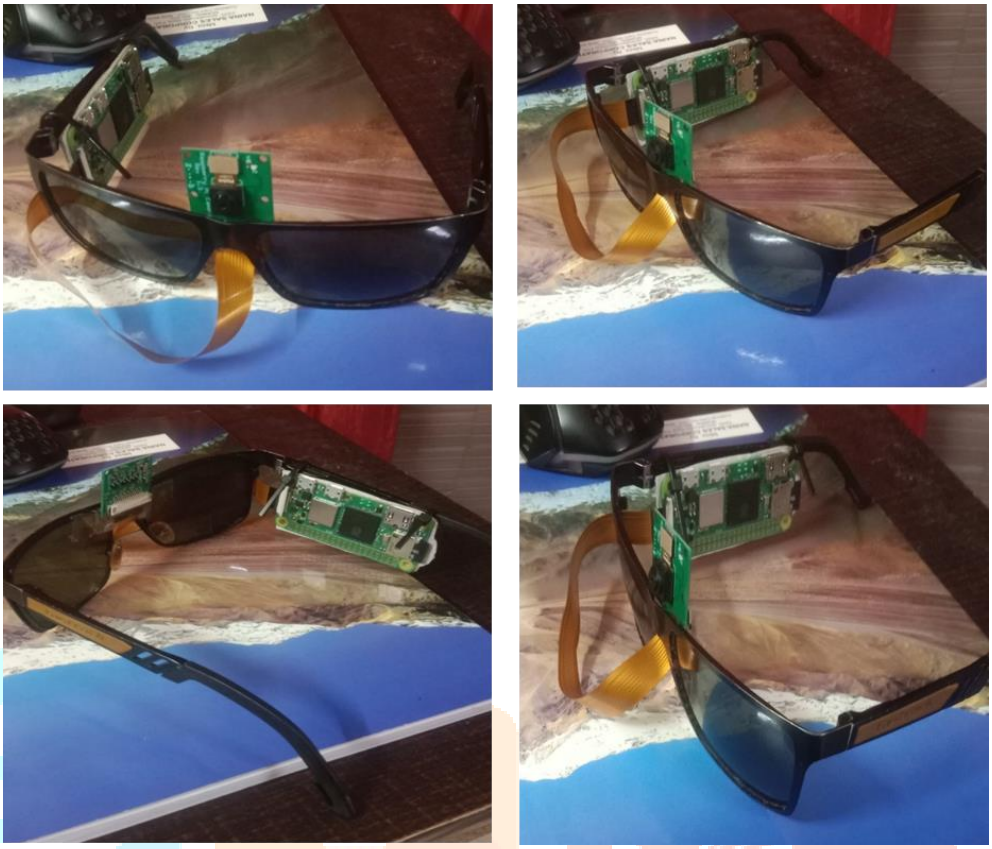

Fig.3 Here we mounted components on the regular glasses to understand structure and design of final glasses

Fig.4 A person wearing prototype of glasses

❖ **Software Architecture:**

i. Operating System and Development Environment: The system runs on Raspberry Pi OS, a lightweight Linux-based platform optimized for the hardware. Development is supported by Python 3.11 and a suite of libraries including OpenCV for image processing, TensorFlow/PyTorch for deep learning tasks, and Tesseract [7] or EasyOCR for optical character recognition.

ii. Image Preprocessing Module: Captured images undergo preprocessing such as resizing, denoising, and brightness adjustment to standardize the input quality for subsequent analysis.

iii. Vision Language Model (VLM): At the core of the image analysis is a VLM, which employs transformer-based algorithms to fuse visual and textual data. This model performs:
   a. Object Detection: Identifies and classifies objects within the visual field, providing spatial cues (e.g., left, right, or centre).
   b. Optical Character Recognition (OCR): Extracts and digitizes text from the environment for audio conversion.
   c. Face Recognition: Detects and identifies faces by comparing captured features with a pre-registered database.

iv. Voice Assistant Module: This module processes the user's voice commands, facilitating mode switching (e.g., between object detection, text reading, and face recognition) and system configuration. It also supports multilingual interactions to accommodate diverse user needs.

v. Audio Feedback Engine: A text-to-speech engine converts processed data into real-time auditory cues. This immediate feedback loop enables users to navigate their environment safely and confidently.

❖ **Data Flow and System Integration:**

i. Data Acquisition: The camera continuously captures environmental images while the microphone records voice commands from the user.

ii. Preprocessing and Analysis: Captured images are preprocessed and then analysed by the VLM on the Raspberry Pi. Simultaneously, the voice assistant interprets spoken commands to adjust system behaviour or switch operational modes.

iii. Feedback Loop: The results from object detection, OCR, and face recognition are converted into audio cues by the text-to-speech engine. This auditory information is delivered through the speaker, closing the feedback loop and guiding the user in real time.

iv. Optional Cloud Integration: For computationally intensive tasks, the system can interface with cloud services. This offloading ensures robust performance and scalability without compromising real-time responsiveness

V. WORKING OF VLM

Humans can effortlessly process information from multiple sources simultaneously; for instance, we rely on interpreting words, body language, facial expressions, and tone during a conversation. Similarly, vision language models (VLMs) can process such multimodal signals effectively and efficiently, enabling machine vision. Thus, understanding and generating image information blending visual and textual elements. Modern VLM architectures rely mostly on transformer-based [8] AI models for image and text processing because they efficiently capture long-range dependencies.

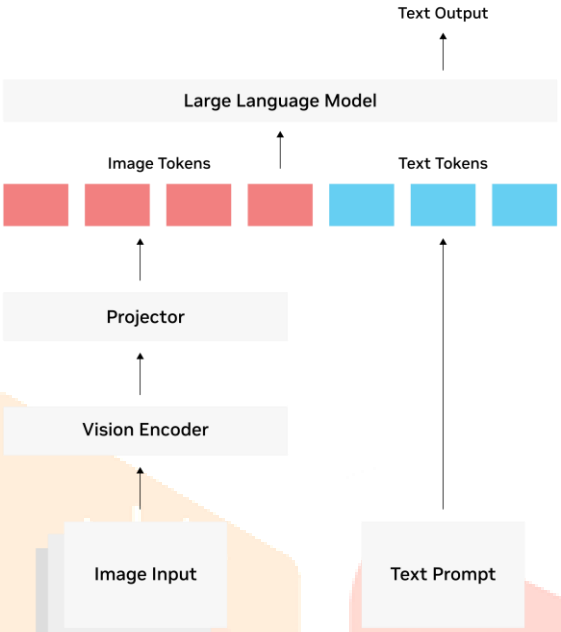

FIG.5 A COMMON THREE-PART ARCHITECTURE FOR VISION LANGUAGE MODELS [9]

❖ Components of Vision Language Models:

To achieve this multimodal understanding, VLMs typically consist of 3 main elements:

i. An Image Model: Responsible for extracting meaningful visual information such as features and representations from visual data, i.e., Image encoder.
ii. A Text Model: Designed to process and understand the natural language processing (NLP), i.e., text encoder.
iii. A Fusion Mechanism: A strategy to combine the representations learned by the image and text models, allowing for cross-modal interactions.

we can group encoders depending on the fusion mechanism to combine representations into three categories: Fusion encoders (which directly combine image and text embeddings), dual encoders (which process them separately before interaction), and hybrid methods that leverage both strengths. Also, based on the same paper, two main types of fusion schemes for cross modal interaction exist: single-stream and dual-stream.

Before digging deeper into specific architectures and pretraining methods, we must consider that the surge of multimodal development over the recent years, powered by advances in vision language pretraining (VLP) methods, has given rise to various vision-language applications.

They broadly fall into three categories:

i. Image-text tasks: image captioning, retrieval, and visual question answering.
ii. Core computer vision tasks: (open-set) image classification, object detection, and image segmentation.
iii. Video-text tasks: video captioning, video-text retrieval, and video question-answering.

❖ Open-Source Vision Language Model Architectures:

VLMs typically extract text features (e.g., word embeddings) and visual features (e.g., image regions or patches) using a text encoder and visual encoder. A multimodal fusion module then combines these independent streams, producing cross-modal representations. A decoder translates these representations into text or other outputs for generation-type tasks

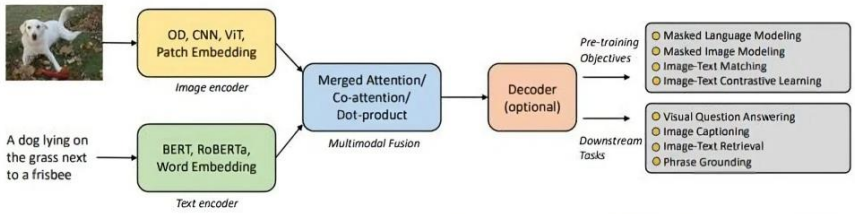

Fig.6 An illustration of the general framework of modern Vision Language Models [10]

Image Encoder (upper part):

i. Image input (shown as a picture of a dog).
ii. The image is encoded using various models like OD (Object Detection), CNN (Convolutional Neural Networks), ViT (Vision Transformer), or Patch Embedding methods.
iii. These models extract features from the image and convert them into a numerical representation (embedding).

Text Encoder (lower part):

i. The text input is "A dog lying on the grass next to a frisbee."
ii. This text is encoded using models such as BERT, RoBERTa, or other word embedding techniques.
iii. These models process the text and convert it into a numerical representation (embedding).

Multimodal Fusion:

i. The outputs from both the image and text encoders are merged together using mechanisms like Merged Attention, Co-attention, or Dot-product operations.
ii. This step fuses the visual and textual information for further processing.

Decoder (optional):

i. After the fusion, the model may use a decoder to further process the fused information, depending on the specific task.

Pre-training Objectives (right side):

The model can be pre-trained with various objectives, including:

i. Masked Language Modelling: Predicting masked words in text.
ii. Masked Image Modelling: Predicting missing or corrupted parts of the image.
iii. Image-Text Matching: Determining if a given image and text match.
iv. Image-Text Contrastive Learning: Learning representations that align images and their corresponding text while distinguishing them from mismatched pairs.

Downstream Tasks (right side):

Once pre-trained, the model can be fine-tuned for specific tasks such as:

i. Visual Question Answering (VQA): Answering questions based on the image.
ii. Image Captioning: Generating descriptive captions for images.
iii. Image-Text Retrieval: Retrieving an image or text based on the other modality.
iv. Phrase Grounding: Identifying specific objects in an image that correspond to words or phrases in the text.

VI. RESULTS

This section presents the outcomes from the evaluation of the smart glasses system, focusing on its key functionalities: object detection, text recognition (OCR), and face recognition, as well as overall system performance and user usability.

❖ Object Detection

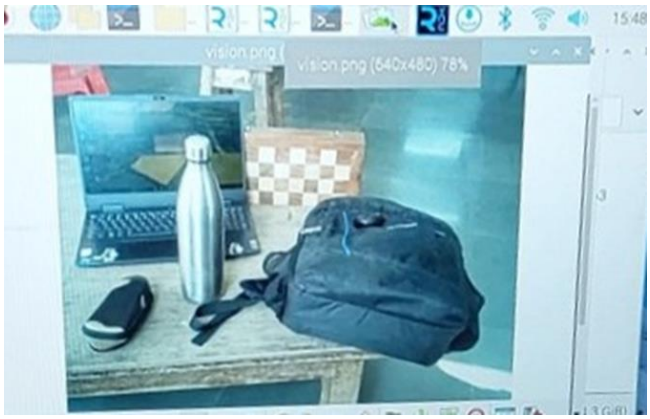

Fig.7 Input image for object detection taken by camera on glasses

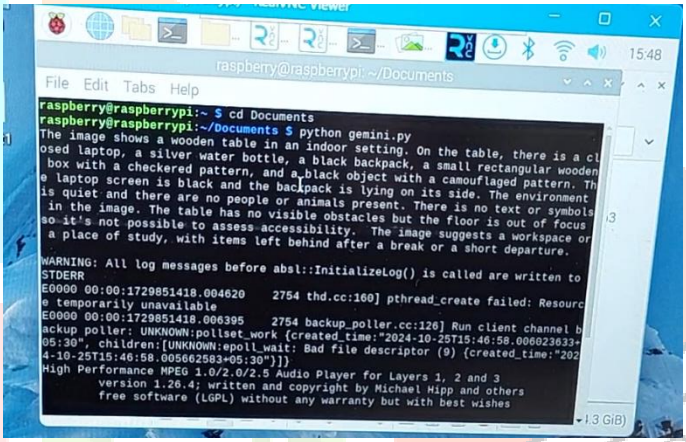

Fig.8 The description and object detection of input image generated by VLM. It successfully detects and narrates the scene captured to person via audio output

i. Output: (The image shows a wooden table in an indoor setting. On the table, there is a closed laptop, a silver water bottle, a black backpack, a small rectangular wooden box with a checked pattern, and a black object with a camouflaged pattern. The laptop screen is black and the backpack is lying on its side. The environment is quiet and there are no people or animals present. There is no text or symbols in the image. The table has no visible obstacles but the floor is out of focus so it's not possible to assess accessibility. The image suggests a workspace or a place of study, with items left behind after a break or a short departure.) - image description generated by VLM

ii. Performance: The system continuously captures images via the high-resolution camera and processes them in real time using the Vision Language Model. Testing (refer to Fig.7 and Fig. 8) demonstrated that the device accurately detects common objects (e.g., chairs, doors, vehicles) and correctly provides spatial cues (e.g., left, right, centre).

iii. Latency: The average processing time per frame was within an acceptable range for real-time navigation, ensuring that auditory feedback is delivered promptly to guide the user

❖ **Text Recognition (OCR)**

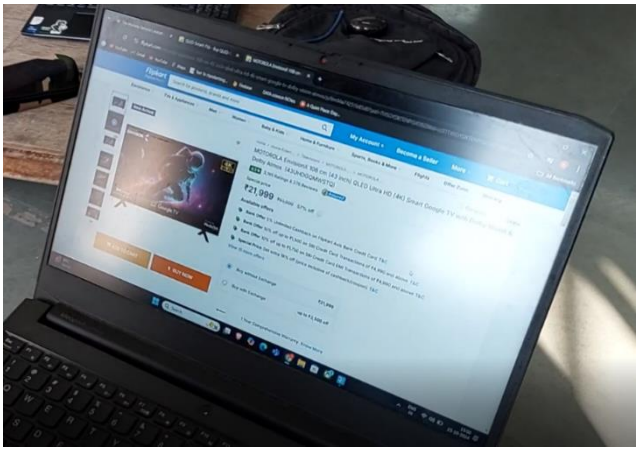

Fig. 9 Input image for OCR taken by camera on glasses, a flipkart product page

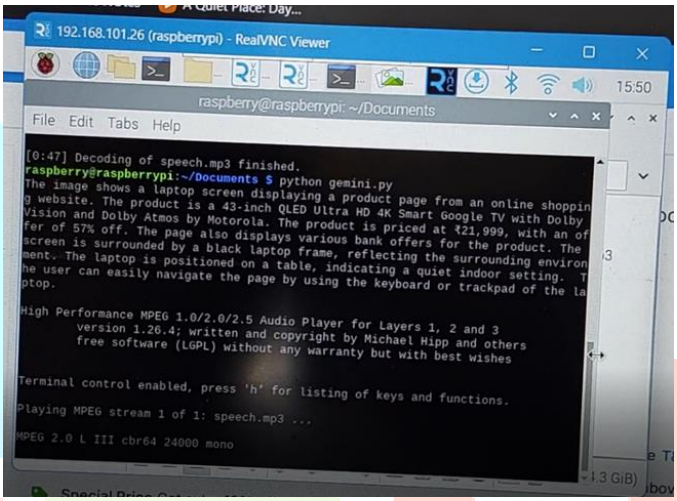

Fig. 10 Glass provide audio output of text present in image or flipkart product page

i. Output: (The image shows a laptop screen displaying a product page from an online shopping website. The product is a 43-inch QLED Ultra HD 4K Smart Google TV with Dolby Vision and Dolby Atmos by Motorola. The product is priced at ₹21,999, with an offer of 57% off. The page also displays various bank offers for the product. The screen is surrounded by a black laptop frame, reflecting the surrounding environment. The laptop is positioned on a table, indicating a quiet indoor setting. The user can easily navigate the page by using the keyboard or trackpad of the laptop.) - image description generated by VLM

ii. Accuracy: The OCR module was evaluated using images containing printed text (as shown in Fig.9). The system successfully extracted textual content from various documents, such as product pages, and converted it into clear, intelligible audio output

iii. Robustness: Performance remained high under well-lit conditions, though minor accuracy degradation was observed in low-light scenarios, suggesting a need for further optimization in variable lighting.

VII. CONCLUSION

The development of smart glasses for visually impaired individuals has made significant progress in recent years, leveraging affordable and accessible technologies like Raspberry Pi Zero 2W, camera modules, and earphones to provide real-time assistance. These devices focus on enhancing the independence of blind users by integrating features such as object detection, facial recognition, text reading, and obstacle avoidance. By providing audio feedback, these glasses help users navigate their environments, identify obstacles and objects, and even engage in social interactions through facial recognition.

Key findings:

i. Affordable Integration: Utilizes accessible technologies such as Raspberry Pi Zero 2W, camera modules, and earphones.
ii. Multifunctional Capabilities: Combines object detection, facial recognition, text reading (OCR), and obstacle avoidance.
iii. Real-Time Audio Feedback: Provides users with immediate auditory information to navigate and interact with their surroundings.
iv. Enhanced Safety and Autonomy: Improves mobility and independence by enabling users to detect obstacles and identify objects.
v. Advances in AI and Edge Computing: Reduces processing latency, leading to faster and more seamless user experiences.

While current solutions show promise in increasing independence, challenges remain in terms of improving accuracy, reducing processing time, and enhancing the portability of these devices. As research continues, future developments could incorporate more advanced AI algorithms, better energy efficiency, and enhanced design to create even more effective solutions. Ultimately, smart glasses are poised to significantly improve the quality of life for blind individuals by offering practical and reliable assistance in real- time, fostering greater autonomy and independence in their everyday lives.

REFERENCES

[1] G. Sudharshan, S. Sowdeshwar, and M. Jagannath, "Smart Glass with Multi-Functionalities for Assisting Visually Impaired People," Journal of Physics: Conference Series, vol. 2318, p. 012001, 2022. doi: 10.1088/1742-6596/2318/1/012001

[2] Akalya, S., Gayathri, D., Sushma, E., & Venkatesh, C. (2021). Smart Glass for Blind People. IT in Industry, 9(3). Published online April 15, 2021

[3] S. Salvi, S. Pahar, and Y. Kadale, "Smart Glass Using IoT and Machine Learning Technologies to Aid the Blind, Dumb, and Deaf," Journal of Physics: Conference Series, vol. 1804, p. 012181, 2021. doi: 10.1088/1742-6596/1804/1/012181

[4] S. Sayed, A. Mansour, and A. Mahmoud, "Real-Time Object Detection and Audio Feedback System for Assisting Blind People," Sensors, vol. 20, no. 18, p. 5325, 2020. doi: 10.3390/s20185325

[5] A. R., H. S., and H. P. P., "Customized Smart Glasses for Needy Blind People," International Journal of Online and Biomedical Engineering (iJOE), vol. 16, no. 13, pp. 33–44, 2020. doi: 10.3991/ijoe.v16i13.18527

[6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," arXiv preprint arXiv:2304.00685, Feb. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2304.00685

[7] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," arXiv preprint arXiv:1706.03762, Aug. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[9] NVIDIA, "What Are Vision-Language Models?" [Online]. Available: https://www.nvidia.com/en-us/glossary/vision-language-models/.

[10] Viso.ai, "Vision Language Models: Exploring Multimodal AI," [Online]. Available: https://viso.ai/deep-learning/vision-language-models/.