Personal Assistant for the Visually Impaired

TEAM SP25-42

RUTGERS UNIVERSITY

Sarmad Ahmed, Luis Castellanos

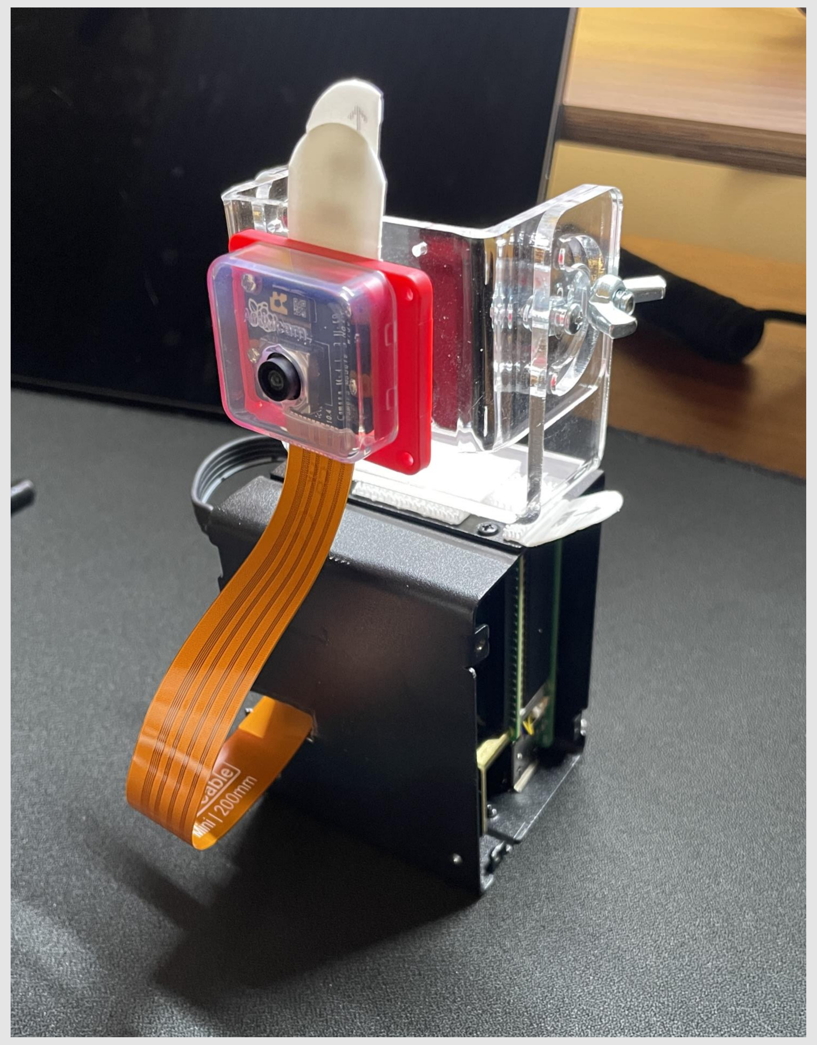
Project Advisor(s): Sasan Haghani

Abstract

Over 6 million Americans suffer from vision lost daily. These people often depend on basic assistive tools for navigating however, such systems do not provide a detailed description of the user's surroundings. To address this, our team has developed a portable, network-free AI assistant. By leveraging modern edge computing and advance software systems we created a system that integrates optical character recognition (OCR) , Text-to-Speech (TTS), Speech-to-Text (STT), and real-time object detection into a cohesive, user-friendly device. Ultimately, the goal of Our project is to enhance the independence and awareness of visually impaired users.

Future Direction

- Limited processing power on Raspberry Pi 5 causes LLaVa Phi model to take approximately 10 – 15 minutes to run.
- Replace Raspberry Pi 5 with NVIDIA Jetson Orin Nano board optimized to run LLMs.
- Enhance YOLOv8 and ResNet-18 performance.
- Replacing Raspberry Pi HQ Camera with a lightweight depth sensing camera.
- Multilingual Support



Challenges

Hardware Limitations

- Optimizing models for deployment on Raspberry Pi 5.
- Creating a balance between accuracy and latency in speech and visual processing to provide the user reliable feedback.

Methods and Implementation

Voice Command Recognition (STT)

- Microphone captures spoken user commands that are converted into text for the system.

Capturing Image

- User's voice command triggers camera to capture image.

Real Time Object Detection

- System analyzes captured image to identify objects in the user's surroundings.
- Detected objects passed on for prompt generation.
- YOLOv8n (Nano version) pretrained on COCO dataset.

Scene Recognition

- System identifies the scene/setting from the captured image.
- ResNet-18 pretrained on Places365
- Predictions > 0.6 confidence passed on for prompt generation.

Prompt Generation

- Generate prompt using the predicted object and scene labels from YOLO and ResNet-18 respectively.

Multimodal Inference

- Generates description of image combining captured image and the generated prompt.
- LLaVa-Phi model composed of:
  - CLIP-based Vision Encoder
  - Phi-2 Language Model
- Runs locally on Raspberry Pi 5 using llama.cpp
- Final output is 25 word sentence describing user's surroundings and potential safety concerns if any.

Audio Feedback (TTS)

- Converts generated surroundings description to audio for the user.
- User can also prompt device to extract text from images using Tesseract OCR.
- Piper TTS

Results

Image Captured by Camera

```
[
[=== YOLOv8 DETECTIONS ===
 {'bowl': 1, 'dining table': 1, 'chair': 3, 'refrigerator': 1, 'microwave': 1}
(assistant-env) sa1790@raspberrypi:~/Capstone/llama.cpp $
```

Objects Identified by YOLO

```
kitchen with 0.7045653462409973 probability
wet_bar with 0.043919432908296585 probability
dining_room with 0.043560612946748734 probability

=== SCENE RECOGNITION (ResNet / Places365) ===
 [('kitchen', 0.7045653462409973), ('wet_bar', 0.043919432908296585), ('dining_room', 0.043560612946748734)]
(assistant-env) sa1790@raspberrypi:~/Capstone/llama.cpp $
```

Predicted Scenes with Confidence

Prompt Generated to Input into LLaVa

Describe this kitchen scene with 1 bowl, 1 dining table, 3 chair, 1 refrigerator, 1 microwave. Briefly mention spatial relationships and safety concerns for visually impaired and blind people. Keep response under 25 words.

Final Output

You are in a kitchen. A dining table sits near the refrigerator. A cupboard is located to the far right. The floor space appears open for safe movement.

REFERENCES

[1] Dmitrii Eliuseev, "YOLO Object Detection on the Raspberry Pi - TDS Archive - Medium," *Medium*, Jul. 11, 2023.
[2] "LLaVA," *llava-vl.github.io*. https://llava-vl.github.io/
[3] A. Rosebrock, "PyTorch image classification with pre-trained networks - PyImageSearch," *PyImageSearch*, Jul. 26, 2021. \(accessed Apr. 18, 2025).