NYU

NYU Tandon

Vision Aid

Intro to Machine Learning

Instructor: Sundeep Rangan

Presentation By:
Bereket Deneke (bd2249)
Maheen Eatazaz (me2400)
Tanzia Nur (ttn309)

Team

Bereket Deneke

Major: Computer Science

Year: Junior

Campus: NYUAD

Maheen Eatazaz

Major: Computer Science

Year: Junior

Campus: NYUAD

Tanzia Nur

Major: Computer Engineering

Year: Senior

Campus: Tandon

Project Objective

Our program aims to assist the visually impaired by enabling safe navigation through indoor spaces while identifying and alerting them of potential obstacles and hazards in their path.

- **Problem Context**: Many visually impaired individuals struggle to form a mental image of their surroundings.
- **Objective**: Provide an auditory-friendly description of scenes to assist with navigation and environmental awareness.
- **Relevance**: Addresses the real-world challenge of aiding the visually impaired with spatial understanding.

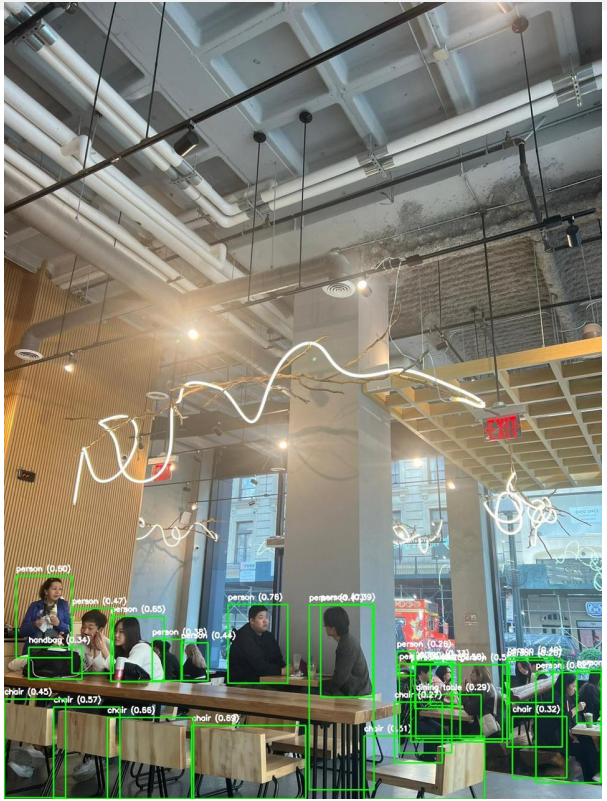
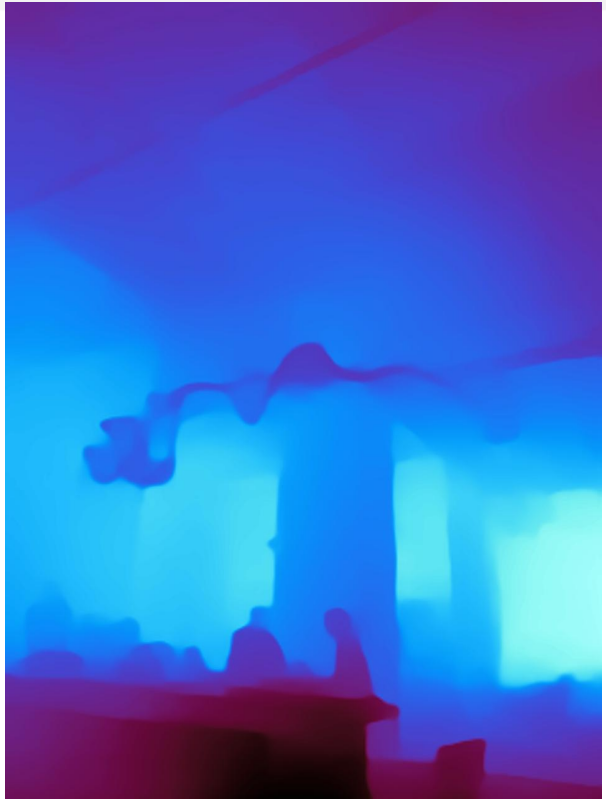
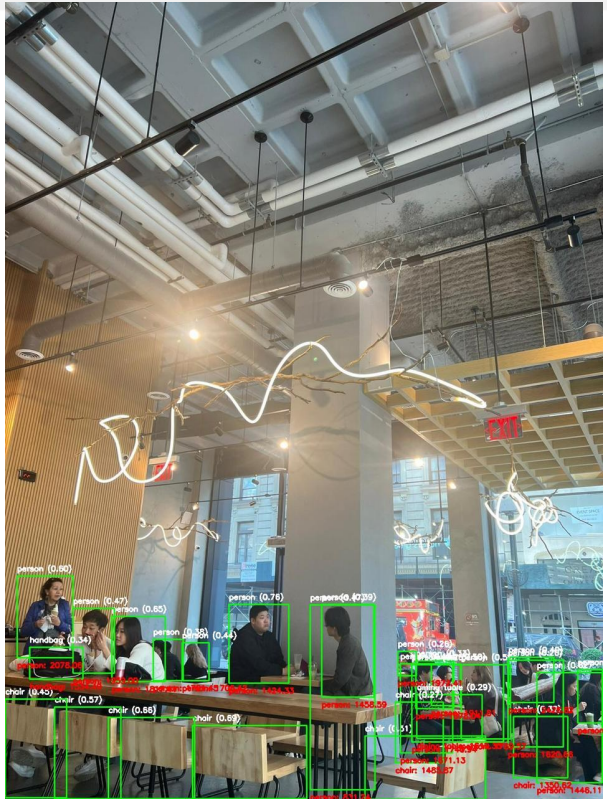
Approach and Key Features

Key Components:

- **YOLOv5** for Object Detection: Identifies objects like people, cars, signs.
- **MiDaS** for Depth Estimation: Estimates distances, giving spatial context.
- **OpenAI GPT** for Language Generation: Converts object positions and depths into natural, directionally-aware language.
- **Text-to-Speech (TTS):** Converts the GPT-generated description into spoken words, offering direct auditory feedback.

Workflow:

1. Input Image/Video
2. Detect Objects (YOLOv5)
3. Estimate Depth (MiDaS)
4. Generate Descriptive Output (GPT)
5. Read Aloud (TTS)

Example of indoor scene test images

NYU

Demo!

Future Improvements

- **Model Fine-Tuning**:
  - Fine-tune YOLOv5 and MiDaS on custom datasets not previously seen.
- **Experiment with Different Models**:
  - Test alternative object detection or depth estimation models.
  - Incorporate more advanced language models or specialized visual question answering (VQA) techniques.
- **Extended Scope**:
  - Real-time deployment on mobile or AR devices.

Challenges & Reflections

Technical Challenges:

- Prompting GPT to use only the provided scene information was challenging. It often attempted to fill in details not given.
- Utilizing Streamlit for visualization was tricky due to how Streamlit handles asynchronous tasks.
- Getting depth map of a particular object detected by YOLO was slight difficult

Reflections:

- Highlight the importance of prompt engineering and careful model tuning.
- Recognize that integrating multiple models (vision + language) requires balancing complexity and clarity.
- Consider iterative refinements in model selection and prompt design for more reliable outputs

Thank You!

Any questions?