NaviGatr:

Visual Assistance for Walking Persons

Eliav Hamburger, Naitik Gupta, Micah Wright

ENEE408N

April 7, 2025

Contributions

Eliav Hamburger:
I spearheaded the depth estimation module of the project. I researched various monocular depth estimation algorithms, tested a few of them, and then made a selection based on benchmarking data. I create a .yml environment file to make the testing environment reproducible. I edited the demonstration code to return metric depth in the image preview. I created the project plan flowchart as well as all the slides related to depth estimation in the milestone report.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination.

Micah Wright:
I was responsible for object detection slides on the milestone review presentation. I evaluated possible object detection models based on the constraints outlined on the milestone review presentation. I successfully deployed the chosen model based on supporting documentation and the implementation of an IP camera. I have drafted documentation related to the object detection component for the replication of results and informational support for a NaviGatr product user.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment.

Naitik Gupta:
I am responsible for the emotion detection component of our project, which processes the face detected by the object detection model to predict emotion.. I also helped set up the GitHub repository and manage our team's workflow and internal deadlines. In addition to contributing to the presentation and documentation, I have been setting up the hardware system, including the Raspberry Pi and Coral TPU. and designed, 3D printed the prototype, and I am handling the integration of the hardware, firmware, and deployment onto our embedded system.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment.

Executive Summary

Our project, NaviGatr, is driven by the goal of making navigation for the visually impaired a richer and more detailed experience. In researching state-of-the-art technology for the vision impaired, we have found most aids, both manual and virtual, to be lacking in some capacity. NaviGatr is built as a lightweight, extensible platform to address many of the pitfalls of current aids for the visually impaired. From a bird's eye view, NaviGatr seeks to identify common objects in a user's line of sight, assess their distance from the user, in the presence of nearby people, assess their emotional state, and provide the user with an audio output describing their surroundings.

NaviGatr is borne out of the convergence of three different machine learning algorithms: monocular metric depth estimation, object detection, and emotion recognition. These three models serve as the engine behind the NaviGatr platform, which runs on a Raspberry Pi, an edge computing device, with a Google Coral Tensor Processing Unit, an external battery, and a basic Pi camera.

Our primary constraint under consideration while designing NaviGatr is speed. For our platform to provide useful information to users, it must be able to identify and measure obstacles in real-time. To this end, we tailored our model selection to fast models and included a TPU in our design to maintain quick computation even on an edge device. Another potential roadblock we expect to face is formulating a useful output for the user given the plethora of information provided by the models. One thought we had was to provide the data to a large language model to distill all the objects and distances into a useful soundbite for the user.

Our prototype is still under construction, with our first 3D print of the headband device holder being completed today. Additionally, while each model has been successfully deployed, we have yet to integrate them into a cohesive pipeline for navigation.

When we complete our first prototype, we hope to design an obstacle course to test the accuracy and efficacy of our platform. Tests can range from basic metrics such as simple accuracy scores for objects detected and distance measured to more interactive tests like navigating a person to a specific object or obstacle avoidance.

Introduction

According to the World Health Organization in 2023, about 2.2 billion people have vision impairment worldwide.[1] In the US, the Centers for Disease Control and Prevention (CDC) estimates that 7 million Americans experience vision impairment or complete blindness.[2] Given the severity of the impact of vision impairment and the large portion of society it impacts, it's clear that aids for the visually impaired are critical navigational tasks and quality of life. Our project aims to enrich the navigational and perceptive experience of those with visual impairments.

There are many already existing assistance technologies for those with visual impairment. An article from 2022 reviewed available assistive tools for navigation. This includes camera-based approaches where motion can be detected or pre-provided still images of the environment to allow for mapping.[3] Unfortunately, these technologies fall short of providing a real-time assessment of a user's surroundings. Another technology is audible walking canes which can inform the user of an object's size.[3] This approach has proximity limitations and prevents the usage of a person's hands.

We propose a revolutionary approach for pedestrian navigation with the purpose of assisting visually impaired people, NaviGatr. This approach leverages depth estimation and object detection models on frames pulled from a camera attached to a Raspberry Pi. NaviGatr further enriches the user experience by inferring the emotions of nearby people using yet another machine learning model. NaviGatr parses the outputs of each of these models to craft an auditory output describing and pinpointing nearby obstacles and objects to the user.

Our prototype of NaviGatr will be mounted on a 3D-printed headband. The headband will be equipped with a Raspberry Pi 4, a basic Pi camera, a Google Coral tensor processing unit, and a battery pack. The Raspberry Pi will pull frames from the Pi Camera and proceed to run each model on the frame. For object detection, we're deploying the NanoDet model; for depth estimation, we're deploying the Depth Pro model; for emotion detection, we're deploying an EfficientNet model. Using the model outputs, it will identify where objects are and the emotions of nearby people and craft this knowledge into an auditory output.

Goals and Design Overview

NaviGatr's design is managed from the official Github repository located at https://github.com/naitikg2305/NaviGatr/.

Given our motivation to assist the visually impaired, we have developed a set of goals focused on safe and informative navigation. Furthermore, we have formulated 3 main objective statements that act as technical milestones. These objectives are distinct from the goals but are motivated by them.

Goals:

- Assist visually impaired individuals with real-time environmental awareness
- Detect and classify surrounding objects using object detection models
- Estimate object proximity using depth-sensing algorithms
- Recognize human facial emotions for social context when someone is nearby
- Provide intuitive voice-based feedback to inform the user
- Implement the system on embedded hardware for portability while maintaining low latency

Objectives:

A. Accurately locate objects in 3D space that are relevant to footpaths
B. Alert person of meaningful objects with positioning/property information
C. Direct navigation to circumvent or interact with objects
D. Help recognize and interact with people around them better

Our objectives are realized through specifications with our goals as the guiding principles. These specifications are broken down per objective (obj.) and are provided in Table 1.

Objective	Spec 1	Spec 2	Spec 3	Spec 4
A	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall detect the chair's distance within a deviation of 0.762 meters.	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall detect object classification of 'chair' with a minimum confidence of 98%	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall process depth and object detection using an identical frame and shall be able to provide object depth given a request with a detected	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall provide a boolean True/False on object presence given a request for service to the object detection component with pixel location as a parameter

			bounding box' object center pixel location. The depth response should be within a deviation of 0.762 meters.	
B	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall alert the user of an object.	With a static user (anchor pt.), with a single chair in FOV and 10 meters from anchor pt., the system shall not alert the user of an object.	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall indicate the object type to the user.	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt., the system shall alert object distance to the user.
C	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt. and center, the system shall determine rotation about the anchor's vertical axis to recenter the anchor pt. with no object at the new center.	With a static user (anchor pt.), with a single chair in FOV and 1.524 meters from anchor pt. and center, the system shall alert of the required rotation from a static position to avoid the object.	N/A	N/A
D	With a static user (anchor pt.), with a single person in FOV and 1.524 meters from anchor pt. and center, the system shall alert of the user of the presence of a human.	With a static user (anchor pt.), with a single person in FOV and 1.524 meters from anchor pt. and center, the system shall indicate what emotion the person displays and alert it the user.	N/A	N/A

FOV- Field of view

Table 1 - Objective and Relevant Metrics

For obj.A-spec1, 1.524 meters represents 2 walking strides away from the user. The threshold 0.762 meters is given to express that the system should not be off by more than 1 walking stride.

In total, there are 12 specifications needing validation.

To meet these design specifications, we have developed a system that deploys machine learning models on a Raspberry Pi4 edge device. This report delves into the details of this system in the *Alternative Design and Design Choices*, *Technical Analysis for System and Subsystems*, and *Design Validation for System and Subsystems* sections. An overview block diagram is presented in Figure 1.

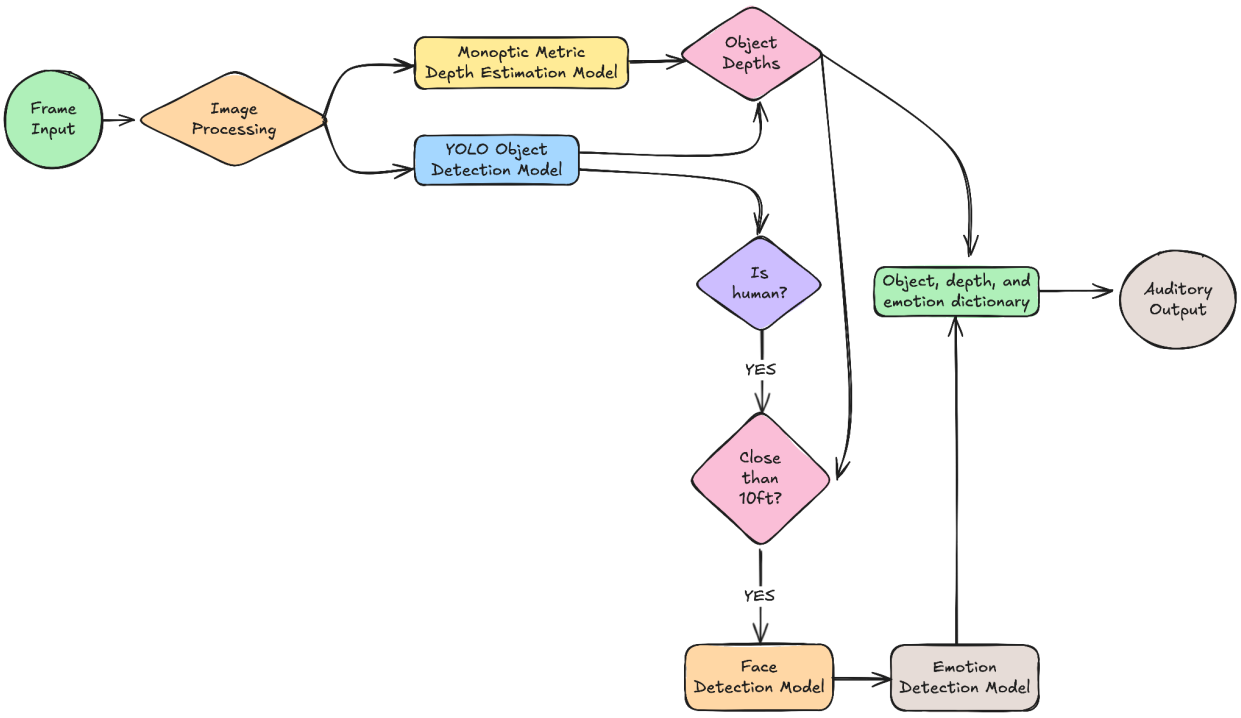

Figure 1 - Project Flowchart

Constraints

Unclassified Objects:

Our system plan is currently limited to detecting objects in the COCO dataset on which NanoDet was trained.[15] Due to this limitation, objects which are not classified in the limited dataset will go undetected. This could result in users encountering dangerous objects without forewarning. For example, there are a large number of animals that are not found in the COCO dataset[5]; if users were to encounter these animals, they might not be alerted to the animal's presence, which could lead to disastrous results. While using a larger dataset could aid in expanding our object classification, it would likely slow down the model and still be incomplete. In addition, simple things like bumps or breaks in a path are not objects and therefore will not be detected. Eventually, our model could use anomalies in the depth estimates to identify hazards like these, but for now, NaviGatr should probably be used in concert with more conventional tools, like the white cane.

Information Parsing:

In busy scenes NaviGatr will be handling large sets of data and paring down that data into a human-friendly audible prompt may be difficult or impossible. For example, walking down a busy street in New York City could result in dozens of human, bicycle, and pet detections in any given scene. Additionally, these objects will likely be in constant motion. NaviGatr will likely be unable to provide a useful output to the user as it won't necessarily be able to prioritize the proper objects and by the time it finishes with its audio description, the scene may have changed completely. While keeping our models as fast and nimble as possible is paramount to speedy information transmission, this constraint will likely always present a significant problem for NaviGatr.

Peripheral Objects:

Any objects outside the camera's field of view will remain unclassified, so what might be in a user's peripheral vision, may not be within the camera's range, potentially leading to users not learning about critical objects. This concern can be mitigated by using wide-lensed cameras and creating a "scanning mode," in which the user can move their head about to capture multiple frames for analysis before receiving direction. Once again these concerns may be mitigated by using other assistive devices like the walking cane.

Boxed Depth Estimates:

Since our object detection model provides bounding boxes, rather than object masks, our depth estimates may not be accurate; given an object's bounding box, there are many ways to calculate the depth of the object, none of which are guaranteed to be accurate due to the inclusion of

background in the box as well as potential other objects. For example, if NaviGatr captures a scene with a person and a basketball being thrown past them, NaviGatr may accidentally transmit the depth of the basketball instead of the depth of the person if the ball is undetected and within the person's bounding box. While this constraint could potentially be mitigated by averaging over masked objects, this may increase inference time resulting in lagging output.[6]

Social Stigma:

NaviGatr's current physical construction is quite obtrusive, possibly inviting unwanted attention to its users. The general public is unused to seeing people with battery packs and cameras strapped to their heads, which may ostracize users. Additionally, many people may have privacy concerns or simply be uncomfortable being present in front of a camera, especially in sensitive areas like restrooms. Lastly, some areas have restrictions on videography, which may prevent users from navigating through them. Future prototypes of NaviGatr could potentially be less noticeable, through technologies like cameras integrated into glasses and bone conduction audio. However, given our limited resources, our product will simply be unable to blend in.

Accuracy:

NaviGatr runs on various existing Machine Learning algorithms and hence, its reliability depends on the accuracy of the current machine learning algorithms. In the case of object detection, and depth sensing, the models are the more reliable ones. Having said that, they are not perfect and can lead to data that might not relate to actual scenarios. In terms of the emotion models, due to how different human beings are and how expressions are very person-dependent, the accuracy is even lower. They rely on huge datasets of images of different people. Even then, these models can have reliability issues in terms of predictions.

Engineering Standards

Software:

- Our depth estimation and object detection models are both open source and have accompanying documentation on the structures and mathematics that power the models in accordance with IEEE 7000-2021, Annex F.[4,15,21]
- As our current project plan has all computation occurring offline, we avoid a myriad of privacy concerns that come with transmitting live images over the internet.

Hardware:

- We followed thermal design best practices by including heatsinks on the Raspberry Pi 4 and placing a cooling fan within the 3D-printed enclosure to prevent thermal throttling or damage during continuous inference workloads.[22]
- The enclosure was printed using PLA, which is recognized as a safe, non-toxic material.[23] This allows for safe contact with the user's head.

- Our enclosure design avoids sharp edges and includes rounded, ergonomic contours to avoid injuring the user.
- For mobile power, we selected a trusted Anker power bank using Lithium-Ion cells with built-in protections including overcharge, short circuit, and thermal protections.[24]
- The USB 3.0 standard was followed to ensure sufficient bandwidth and power for the Coral TPU, allowing efficient communication between the Raspberry Pi and the Coral.[25]

Alternative Designs and Design Choices for Software

Object Detection Component:
Our object detection model must be deployable on a portable edge device and able to process a live video feed in real-time. Key requirements include low memory demands, power efficiency, and high detection accuracy - especially for nearby objects. Notably, correct labeling is secondary to detecting all relevant nearby entities.

We evaluated the following candidate models:
- NanoDet+
- YOLOX-Nano
- YoloFastestV2
- YOLOv8-Nano

Each model was assessed not only for benchmark performance but also for implementation practicality and documentation support. This includes installation guidance, demo availability, and framework support for Python deployment on Raspberry Pi.

Each model follows a 3-sectional, single-stage (YOLO) structure:
- **Backbone**: Feature extraction from the image
- **Neck**: Feature enhancement through network processing
- **Head**: Provides formatted output

The components of each model are as follows:

Model	Backbone	Neck	Head
NanoDet+	Ghost PAN	PAFPN	Depth-wise separable convs, anchor-free
YOLOX-Nano	DarkNet53	FPN	Decoupled, anchor-free
YOLOv8-Nano	CSPDarkNet53	PAN	Anchor-boxes
YoloFastestV2	ShuffleNetV2	Modified FPN	Decoupled, anchor-free

Table 2 - Image Detection Model Components[7, 8, 9, 10, 13, 14]

Model	Size	mAP	Jetson Nano	RPi 4 1950	RPi 5 2900	Rock 5
NanoDet	320x320	20.6	26.2 FPS	13.0 FPS	43.2 FPS	36.0 FPS
NanoDet+	416x416	30.4	18.5 FPS	5.0 FPS	30.0 FPS	24.9 FPS
YoloFastestV2	352x352	24.1	38.4 FPS	18.8 FPS	78.5 FPS	65.4 FPS
YOLOv8-Nano	640x640	37.3	14.5 FPS	3.1 FPS	20.0 FPS	16.3 FPS
YOLOX-Nano	416x416	25.8	22.6 FPS	7.0 FPS	34.2 FPS	28.5 FPS

Table 3 - Image Detection Model Benchmarks[11]

To compare these models, we utilized the below Pugh matrix.

Criteria	Weight	NanoDet+	YoloFastesyV2	YOLOv8n	YoloXn
Speed	60	5.0 FPS	18.8 FPS	3.1 FPS	7.0 FPS
Accuracy	40	30.4	24.1	37.3	25.8
Documentation	50	20	2	15	10
	Total:	2516	2192	2428	1952

Table 4 - Pugh Matrix for Selection of Object Detection Model

Despite YoloFastestV2 scoring highest, its lack of Python support and sparse documentation made implementation impractical. The model itself is Python-based, but the deployment demonstration in the documentation uses the NCNN framework which is solely (C++)-based.[14, 26] Therefore, NanoDet+ was selected for its balance of performance and deployability.

Depth Detection Component:

In our research, we found very few models that were able to produce metric depth estimates. Two state-of-the-art models stood out after coming through our options:

- ZoeDepth (built on MiDaS)
- Apple's Depth Pro

After evaluating their respective papers and benchmarks, Depth Pro outperformed ZoeDepth in accuracy, boundary precision, and inference time, making it the preferred model.[4]

Pugh Matrix – Depth Model Comparison:

Criteria	**Weight**	**ZoeDepth**	**Depth Pro**
<1 sec. inference	35	1[12]	1
Metric Output	50	1	1
Boundary Adherence	1	1	14
Metric Accuracy	40	1	2
	Total:	126	179

Table 5 - Pugh Matrix for Selection of Depth Estimation Model[4]

Non-binary criteria scores were estimated by rounded averages of scores from [4].

Emotion Detection Component:

The emotion detection module interprets facial expressions from individuals detected in the camera feed to convey their emotional state to visually impaired users. It is only triggered when a person is detected and is within a close range, based on depth data. The system processes the face, classifies their emotion, and audibly reports it.

Model Candidates

Two models were evaluated:

- FER2013 CNN: A custom-trained convolutional neural network using the FER2013 dataset.

- EfficientNetB0 FER: A lightweight, pre-trained model optimized for edge inference and compatible with TFLite and Coral TPU.

Model Comparison Table:

Criteria	FER2013 CNN	EfficientNetB0 FER
Model Size	~1MB (custom)	~5.3MB (quantized)
Accuracy on FER Dataset	~65%	~72–75%
Inference Speed (Raspberry Pi)	~4–6 FPS	~1–2 FPS
TPU Compatibility	Yes (convertible)	Yes (quantized)
Deployment Ease	Easy (TensorFlow)	TFLite conversion req.
Documentation & Support	High (basic CNN)	High

Table 6 - Model Comparison

Pugh Matrix Criteria

Criteria	Weight	FER2013 CNN Model	EfficientNet
Inference Speed[16]	30	1	2
Model Size[17]	10	2	1
Accuracy on FER-2013[18]	30	1	2
Edge Compatibility[19]	20	2	2
Training Simplicity[20]	10	1	2
	Total:	130	190

Table 7 - Pugh Matrix for Emotion Detection Model Selection

FER2013 CNN was trained using grayscale 48×48 images with data augmentation using ImageDataGenerator. While performant and compact, it has slower inference on Raspberry Pi without a TPU. EfficientNetB0 FER offers higher accuracy and TPU compatibility but requires quantization for efficient deployment. For CPU-only setups, FER2013 CNN is preferred[29]. For Coral TPU acceleration, EfficientNet is ideal[30].

Real-World Deployment Considerations

- Subjects will be within 6–10 feet of the device.
- Lower FPS is acceptable since blind users perceive information audibly, not visually.
- Faces are detected, cropped, and resized before emotion inference.

- Grayscale processing ensures consistency across lighting conditions.

EfficientNet is chosen for final deployment due to its balance of performance and compatibility with the Coral TPU. FER2013 CNN remains an efficient and accessible fallback for prototyping or hardware-constrained scenarios.

Alternate Design Considerations and Selections for Hardware

Microcontroller/Computer: Raspberry Pi 4 vs Raspberry Pi 3

- Raspberry Pi 4 was selected due to:[28]
    - Increased RAM (up to 8GB)
    - USB 3.0 support (critical for Coral TPU compatibility)
    - Faster processor for real-time inference
    - Improved thermal control with heatsinks and fan
- Raspberry Pi 3 was rejected due to:[28]
    - Insufficient processing power
    - Lack of USB 3.0 (which limits Coral TPU functionality)

AI Acceleration: Coral TPU Inclusion

- Coral USB Accelerator was included to:
    - Enable fast edge inference using TensorFlow Lite models
    - Reduce CPU load on the Raspberry Pi
    - Maintain real-time performance with lightweight CNNs
- Without Coral TPU:
    - The Pi would struggle to run even moderately sized models like EfficientNet[32]
    - FPS would drop, compromising real-time navigation

3D Printing Material: PLA vs PETG[31]

- PLA was selected:
    - Easy to print, readily available
    - Does not warp easily
    - Sufficient for low-heat electronics if airflow is managed
- PETG was considered:
    - More durable and heat-resistant
    - However, requires more print tuning and time
    - Given we added heatsinks and fans, PLA was acceptable.

Operating System: Ubuntu vs Arch Linux vs Windows

- Ubuntu 24.04 LTS (Lite/Server) was chosen:[32]
    - Stable, well-supported community
    - Easier driver and library support
- Compatible with Coral TPU and TensorFlow Lite
- Arch Linux:[32]
    - Initially considered due to its minimal nature
    - Rejected due to package and driver stability concerns
- Windows IoT:[33]
    - Too heavy
    - Not practical for real-time low-latency inference

Power Design: Battery Placement & Supply

- On-head battery (chosen):
    - Sleeker design
    - No cord running from the headset to user's pocket
- Anker Power Bank in Pocket was selected (rejected):
    - Could heat up in pocket
    - USB cable runs to Pi from headset to pocket
    - Cord could get disconnected
    - Bulky

Model Deployment: Onboard vs Server-Based

- Onboard Model Inference (chosen):
    - Allows real-time processing without network reliance
    - Chosen models are light enough to run on Pi + Coral
- Remote Server Inference (rejected):
    - Too slow and unreliable for navigation
    - Introduces latency and dependency on internet
    - Privacy concerns with transmitting picture data over the internet

Component	Alternatives Considered	Final Choice	Justification
Microcontroller	Raspberry Pi 3, Raspberry Pi 4	Raspberry Pi 4	More power, USB 3.0, future-proof[28]
AI Acceleration	No TPU, Coral USB TPU	Coral USB TPU	Required for real-time inference[32]
3D Material	PLA, PETG	PLA	Easy to print, safe with cooling[31]

OS	Ubuntu, Arch, Windows	Ubuntu Lite	Stability, TensorFlow support[32]
Power Supply	On-head, Pocket battery	Pocket Power Bank	Ergonomics, heat management
Model Deployment	Remote server, Local	Local	Speed, independence

Table 8 - Summary of Hardware and Operating System Alternatives

Technical Analysis of the NaviGatr System

System Level

On a system level, NaviGatr accepts picture frames from the Pi camera and outputs an audio soundbite describing the user's surroundings. To accomplish this higher level implementation, the system utilizes several subsystems to parse and generate information. The image, captured in a Python script, is sent to the object detection, depth estimation, and emotion detection subsystems for parallel processing. The formatted output of each model is then fed into a large language model to turn the raw data into human-understandable text, which is then read outloud to the user through a Python text-to-speech library.

At the system level, the sources of error range from the sharpness of camera input to the real time accuracy of information conveyed to the user. If a given frame is blurry, it cannot reliably output useful information which could result in no or inaccurate information. Additionally, in very dynamic environments, the information conveyed may be irrelevant by the time it's conveyed to the user.

Camera Subsystem

Our camera subsystem is composed of a Raspberry Pi camera with a 5MP still picture resolution, 30 frames/second capture speed, and a JPEG file format output.[27]

Object Detection Subsystem

The object detection subsystem takes a given frame from the camera and processes it with the NanoDet model returning a list of detected classes and bounding box coordinates to the data handler. Sources of error will likely be objects that deviate from the model's conception of that object, resulting in either low confidence or no detection.

Depth Estimation Subsystem

The depth estimation subsystem is built on Apple's Depth Pro. When an image is dispatched to the subsystem, it processes the frame with the Depth Pro model and returns a Numpy array of pixel depth values to the data handler.

Emotion Detection Subsystem

Built on the FER2013 dataset using an EfficientNet model, the emotion detection subsystem takes in any bounding boxes associated with the person class that the depth estimation model found to be within the user's relevant radius. Using these coordinates, the subsystem will narrow the area of focus to a 256x256 pixel facial subregion. This frame is then sent to the EfficientNet model for emotion classification, which returns a list of person locations and associated emotions to the data handler.

Data Handler Subsystem

The data handler subsystem is in charge of dispatching tasks and coordinating the flow of information. It will be written largely in Python, utilizing libraries like TensorFlow Lite. The data handler will request a frame input from the camera and then dispatch it to the object detection and depth estimation subsystems in parallel. Using the results of the two models, if there are people in the user's relevancy radius, it will then dispatch the people and their coordinates to the emotion detection subsystem. Once all the information is received back, the data handler will parse the information into a human-understandable output and then dispatch an audio output for the user.

Computer Subsystem

Our computational subsystem is formed of the hardware components of our Raspberry Pi 4. The Pi is running Ubuntu 22.04 and is powered by a portable power bank. The Pi has access to 8GB of RAM and the Google Corral TPU for added processing power. To prevent overheating, the Pi has integrated heat sinks and cooling fans. This system is fundamentally limited as it's an edge computing device and has limited computing power available.

Design Validation for System and Subsystems

Depth Estimation Preliminary Results:

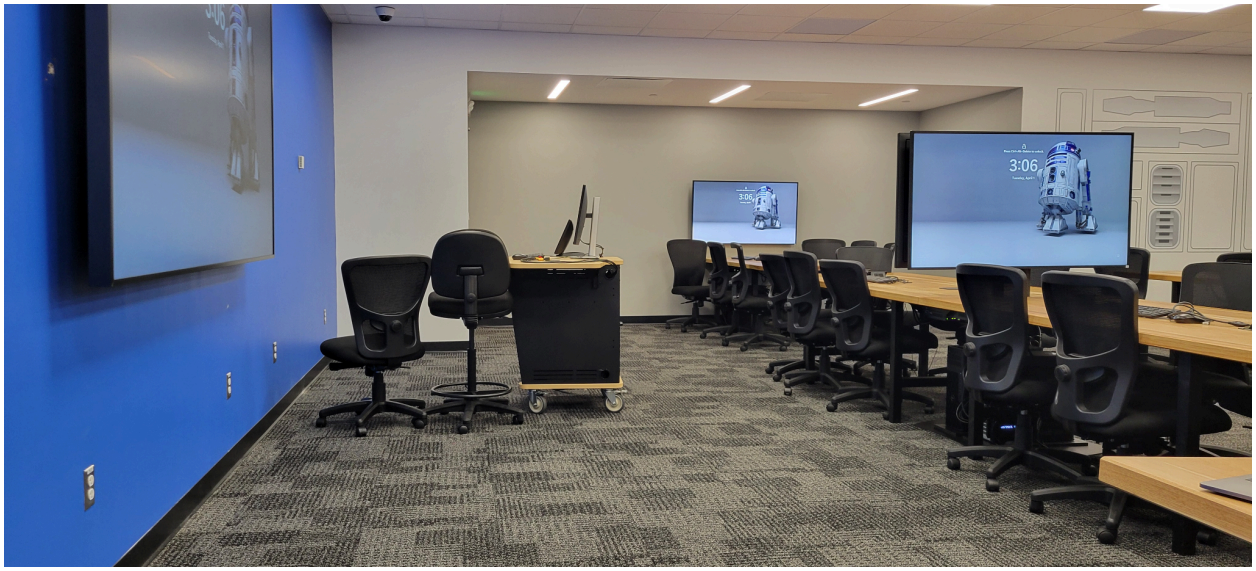

Figure 2 - Image of the ENEE408N Classroom

Figure 3 - Depth Pro Normalized Depth Heatmap of the ENEE408N Classroom (brighter colors are closer)

From figures 2 and 3, it's evident that the depth estimation model is producing promising early results. The relative distances between objects by and large match our expectations, with the

table next to the photographer being the brightest/closest spot on the depth map and the far wall being the darkest/farthest on the depth map. Further testing with ground truth measurements will be needed to assess the metric accuracy of the model.

Object Detection Preliminary Results:

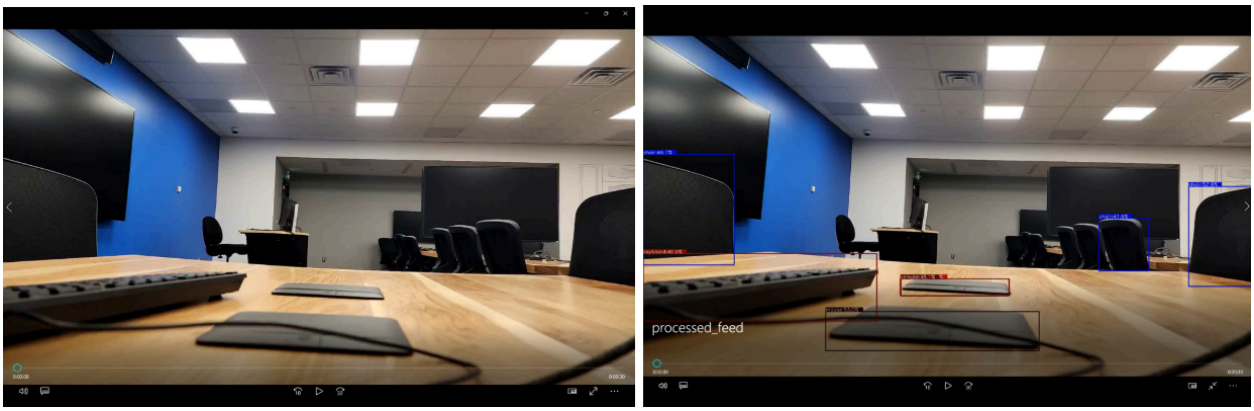

Figure 4 - Image of NanoDet+ Object Detection of ENEE408N Classroom

The object detector test results are shown in Figure 4. On the left is the raw imaging. On the right is the same raw frame after model processing. We see some misclassification of near objects where a docking station is detected to be a ‘remote’. All near objects are detected except for the table since it is not within the field of view for proper detection.

Emotion Detection Preliminary Results:

Photos used in experiment:

Figure 5 - Test Inputs for Emotion Detection Model

```
Predicted emotion scores for image1(happy): [[0.00358334 0.86432326 0.00684604 0.00490304 0.0601556  0.01925882
  0.0409298 ]]
<class 'numpy.ndarray'>
Disgust: 0.86
Neutral: 0.06
1/1 [==============================] - 0s 37ms/step
Predicted emotion scores for image2(neutral): [[0.00882955 0.74789566 0.01837    0.00956405 0.15679091 0.02863535
  0.02991445]]
Disgust: 0.75
Neutral: 0.16
1/1 [==============================] - 0s 32ms/step
Predicted emotion scores for image3(sad): [[0.02249728 0.7597282  0.0064179  0.00809032 0.0919518  0.05594445
  0.05537   ]]
Disgust: 0.76
Neutral: 0.09
1/1 [==============================] - 0s 32ms/step
Predicted emotion scores for image4(disgust): [[0.00741694 0.8006515  0.01456363 0.00699583 0.06528971 0.04815996
  0.05692251]]
Disgust: 0.80
Neutral: 0.07
```

Figure 6 - Model Results Prior to Cropping the Images

```
-----------------------------------Cropped Images-----------------------------------
1/1 [==============================] - 0s 47ms/step
Predicted emotion scores for image1(happy): [[0.01726356 0.44116342 0.13189997 0.11381362 0.17295179 0.09900355
  0.02390403]]
<class 'numpy.ndarray'>
Disgust: 0.44
Neutral: 0.17
1/1 [==============================] - 0s 41ms/step
Predicted emotion scores for image2(neutral): [[0.00983356 0.2757767  0.05559461 0.4559796  0.08998469 0.07134686
  0.041484  ]]
Happy: 0.46
Disgust: 0.28
1/1 [==============================] - 0s 27ms/step
Predicted emotion scores for image3(sad): [[0.00700952 0.15347147 0.02322756 0.63241756 0.0409252  0.11172887
  0.03121989]]
Happy: 0.63
Disgust: 0.15
1/1 [==============================] - 0s 29ms/step
Predicted emotion scores for image4(disgust): [[0.00537685 0.26772988 0.08593299 0.35264322 0.07181924 0.19865449
  0.01784332]]
Happy: 0.35
Disgust: 0.27
```

Figure 7 - Model Results After Cropping the Images

While we can see that the model is not perfectly accurate in the first demonstration, the images have not been put through face detection models that crop the face specifically. This means, the entire photo is converted to 256x256 and might get very blurry, given its very zoomed out. I then decided to crop the images and run the tests again and the results were better than the previous ones. Previously, all images returned disgust but now, emotions were scattered across the board. This model took a total of 26 seconds to run (CPU) since it's already trained. We expect a faster performance on the Coral TPU.

Test Plan

To evaluate the effectiveness of the NaviGatr as a real-world assistive tool for visually impaired individuals, we will conduct a structured test demonstration during the 2025 UMD Capstone

Fair. This demonstration will serve as a brief field test for object detection, depth estimation, and emotion recognition capabilities of the system. We will have a blind-fold that can be worn to simulate visual impairment and have the person try to make it from the start line to the finish line for our course.

Goals of the Demonstration

- Validate real-time object detection and avoidance.
- Evaluate depth estimation for spatial awareness.
- Test emotion detection when a human subject interacts at close range.
- Ensure fluid system integration and accurate text-to-speech guidance.

Physical Setup – Obstacle Course

The demonstration space will be configured as a mini obstacle course with:

1. Start Line – Marked entry point for the test subject.
2. Chair Obstacle – To test static object detection and avoidance.
3. Human Subject – Standing approximately 6–8 feet away and saying “hello” to trigger emotion detection.
4. Low Profile Object – A broomstick on the ground, to test low-lying hazard detection.
5. Dynamic Object – A rolling cart or box slowly moving across the path.
6. Final Station – A QR code or printed signpost to confirm path completion.

Component	Input	Expected Output	Pass Criteria
Object Detection	Real-time camera feed	Audio warning: "Chair ahead", "Obstacle on right"	Object detected with correct spatial reference
Depth Sensing	Scene frame with distance variation	Correct distance tags assigned per object	Distances accurate within ±20%
Emotion Detection	Human face detected < 10 ft	Audio output: 'Person ahead, they look happy/sad'	Emotion detected from facial expression
TTS Output	Parsed alert string	Clear, understandable speech	Speech within 2 seconds
Data Handling	Combined object + depth + emotion frame	Consolidated object array	No loss or delay in frame handoff

Overall System	Video stream input	Real-time audio guidance	Subject can complete course using only system feedback

Table 9 - Evaluation Criteria

The system is considered successful if:

- All 5 obstacles are detected and announced correctly.
- The emotion of the human subject is interpreted accurately
- Text-to-speech audio cues are accurate and timely.
- No system crashes or noticeable delays occur during the demonstration.
- The subject is able to navigate the course without touching or colliding with obstacles.

Desired Performance Metrics

- Response Time (image frame to voice output): < 2 seconds
- Detection Accuracy: ≥ 85% for all objects
- Speech Clarity: Subjective scoring during demo (rated 1–5 by observers)
- Frame Processing Rate: ~10–15 FPS minimum

Additional Notes

- The Coral TPU is used to accelerate inference to support real-time feedback
- The emotion detection model is only invoked when a person is within 10 feet
- The Pi Camera v2 provides 1080p input, but frames are resized for inference
- Safety personnel will be available to intervene if necessary

Project Planning and Management

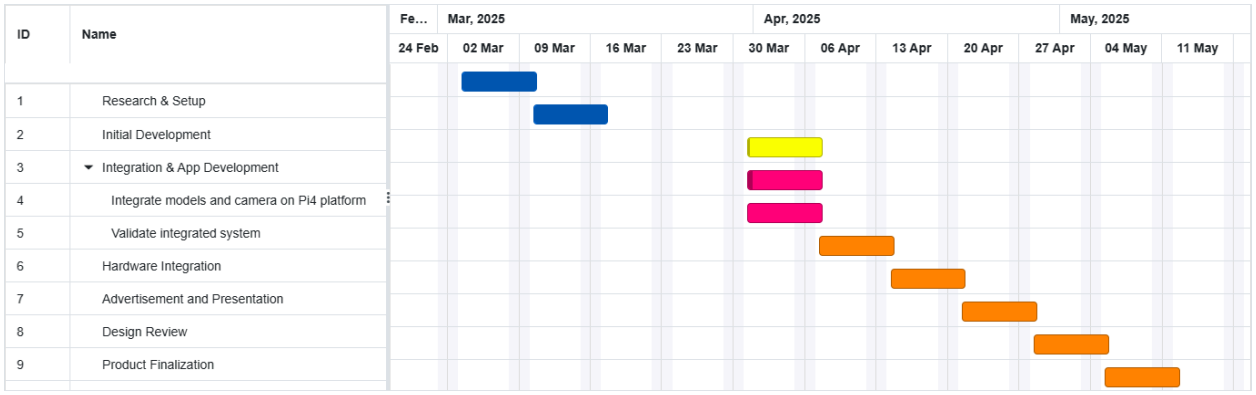

Figure 8 – Gantt Chart of Project Management (Using Gantt Chart Maker Software)[34]

The project chart represents 4 possible tasking stages: 1) Completed (blue), 2) in progress (yellow), 3) overdue (pink), and 4) not started (orange). Shading blocks represent the percentage of progress on a task.

Conclusions

There is a need to assist visually impaired people in navigating foot travel beyond current capabilities. We propose a machine learning solution in a portable, wearable device. The device inputs live-camera feed to our Raspberry Pi which runs model inference on the video input to produce an environment mapping of the space around a person. The mapping then informs the person of obstacles and aids in circumventing them through audible navigation. Our inferencing includes depth detection, object detection, and emotion recognition.

To guide our development we established goals which are to assist visually impaired individuals with real-time environmental awareness; detect, classify, and estimate the depth of surrounding objects; recognize human facial emotions for social context; provide audible guidance to the user; and ensure low-latency, portability of the system. Our test progress serves as validation for the specifications which are still in development.

The formulated constraints are potential challenges in the practical application of this product. We identified 6 key constraints: 1) Environment objects having no corresponding classification labels, 2) adaptation of audio output when the environment changes swiftly and suddenly, 3) limited scene assessment through the camera's narrower field of view, 4) interior bounding box pixels holding background or foreground information aside from the object in which the box pertains to, 5) model outputs are not guaranteed and their results have possible deviations from ground truth, and 6) regulatory restrictions and societal response to the product. These constraints create unique challenges that must be addressed throughout product development.

We evaluated many potential models that align with the established goals. By using the derived specifications, priorities to model features/metrics were assigned. The justification for the model choices was based on those priorities. The models were part of a wider system in which subsystems connect through interfaces and achieve the system-level goals.

Moving forward, we intend to connect all subsystems and run system-wide tests to assess the validation of the project specifications.

References

[1] World Health Organization, “Blindness and vision impairment,” World Health Organization, Aug. 10, 2023. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[2] CDC, “Prevalence Estimates for Vision Loss and Blindness,” Vision and Eye Health Surveillance System, May 24, 2024. https://www.cdc.gov/vision-health-data/prevalence-estimates/vision-loss-prevalence.html

[3] M. D. Messaoudi, B.-A. J. Menelas, and H. Mcheick, “Review of Navigation Assistive Tools and Technologies for the Visually Impaired,” Sensors, vol. 22, no. 20, p. 7888, Oct. 2022, doi: https://doi.org/10.3390/s22207888

[4] A. Bochkovskii et al., “Depth Pro: Sharp Monocular Metric Depth in Less Than a Second,” arXiv.org, 2024. https://arxiv.org/abs/2410.02073

[5] Ultralytics, “COCO,” docs.ultralytics.com. https://docs.ultralytics.com/datasets/detect/coco/

[6] L. Ueno, “Ultimate Guide to Converting Bounding Boxes, Masks and Polygons,” Roboflow Blog, Aug. 15, 2023. https://blog.roboflow.com/convert-bboxes-masks-polygons/

[7] RangiLyu. (2021, December 26). The ultra-simple auxiliary module accelerates the training convergence and greatly improves the accuracy! The real-time NanoDet upgrade for mobile, the NanoDet-Plus, is here! zhuanlan. https://zhuanlan.zhihu.com/p/449912627

[8] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430. https://arxiv.org/pdf/2107.08430

[9] Torres, J. (2024, March 19). YOLOv8 architecture explained: Exploring the YOLOv8 architecture. YOLOv8. https://yolov8.org/yolov8-architecture-explained/

[10] Qiuqiuqiu. (2021, August 16). Yolo-FastestV2: Faster and lighter, up to 300 FPS on mobile, with only 250k parameters. zhuanlan. https://zhuanlan.zhihu.com/p/400474142

[11] Qengineering. (2021). YoloX Jetson Nano [Software]. GitHub. Retrieved from https://github.com/Qengineering/YoloX-ncnn-Jetson-Nano

[12] vtalpaert, “GitHub - vtalpaert/ros2-monodepth: ROS2 Monocular depth estimation using ZoeDepth https://github.com/isl-org/ZoeDepth,” GitHub, 2024. https://github.com/vtalpaert/ros2-monodepth (accessed Apr. 07, 2025).

[13] Qiuqiuqiu. (2022, July 7). FastestDet: Faster than yolo-fastest! Stronger! Simpler! Newly designed ultra-real-time anchor-free object detection algorithm. zhuanlan. https://zhuanlan.zhihu.com/p/536500269

[14] dog-qiuqiu. (2021). dog-qiuqiu/Yolo-FastestV2: V0.2 (V0.2). Zenodo. https://doi.org/10.5281/zenodo.5181503

[15] RangiLyu. (2021). NanoDet-Plus: Super fast and high accuracy lightweight anchor-free object detection model. [Software]. GitHub. Retrieved from https://github.com/RangiLyu/nanodet

[16] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, May 2019. https://arxiv.org/abs/1905.11946

[17] "EfficientNet Models | TensorFlow Hub." [Online]. Available: https://tfhub.dev/s?module-type=image-classification&q=efficientnet

[18] State-of-the-art Accuracy on FER2013 Dataset
"Facial Emotion Recognition: State of the Art Performance on FER2013" – arXiv| https://arxiv.org/pdf/2105.03588.pdf

[19] EfficientNet-EdgeTPU Performance & Speed on Coral TPU
Google AI Blog – EfficientNet-EdgeTPU: Creating Accelerator-Optimized Neural Networks with AutoML
https://ai.googleblog.com/2019/08/efficientnet-edgetpu-creating.html

[20] EfficientNet Design & Training Simplicity
EfficientNet on Wikipedia – Model scaling and simplicity
https://en.wikipedia.org/wiki/EfficientNet

[21] "IEEE Standard Model Process for Addressing Ethical Concerns during System Design," IEEE Std 7000-2021, pp. 1–82, Sep. 2021, doi: https://doi.org/10.1109/IEEESTD.2021.9536679.

[22] "The Pi Hut," The Pi Hut, 2019. https://thepihut.com/blogs/raspberry-pi-roundup/how-to-keep-your-raspberry-pi-4-cool (accessed Apr. 08, 2025).

[23] "Is PLA Safe & Non-Toxic?," FOW Mould, Jan. 16, 2024. https://www.immould.com/is-pla-safe/

[24] "Anker," Anker, 2021. https://www.anker.com/products/a1268 (accessed Apr. 08, 2025).

[25] "USB 2.0 vs 3.0: A Comparative Guide for Beginners 2023 - Anker US," Anker. https://www.anker.com/blogs/hubs-and-docks/usb-2-vs-usb-3

[26] Ni, H., & The ncnn contributors. (2017). ncnn [Computer software]. https://github.com/Tencent/ncnn

[27] "Amazon.com: Arducam 5MP Camera for Raspberry Pi, 1080P HD OV5647 Camera Module V1 for Pi5, Pi 4, Raspberry Pi 3, 3B+, and Other A/B Series : Electronics," Amazon.com, 2025. https://www.amazon.com/dp/B012V1HEP4 (accessed Apr. 08, 2025).

[28] C. L. Lynch, "Raspberry Pi 3 vs. 4: What are the differences?," *MakeUseOf*, Jan. 7, 2024. [Online]. Available: https://www.makeuseof.com/raspberry-pi-3-vs-4-differences/.

[29] J. Ramírez López, "Facial emotion recognition in real-time video using deep learning," B.S. thesis, Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya, Barcelona, Spain, Jan. 2023. [Online]. Available: https://upcommons.upc.edu/bitstream/handle/2117/394168/178133.pdf. [Accessed: Apr. 8, 2025].

[30] Y. Bhalgat, J. Lee, M. Nagel, T. Blankevoort, and N. Kwak, "Comparison of all configurations of quantization with EfficientNet-B0," *ResearchGate*, Jun. 2020. [Online]. Available: https://www.researchgate.net/figure/Comparison-of-all-configurations-of-quantization-with-EfficientNet-B0-FP-accuracy-761_tbl2_343271152. [Accessed: Apr. 8, 2025].

[31] "PLA vs PETG - Which filament is right for me?," FormFutura, [Online]. Available: https://formfutura.com/blog/material-guide-pla-vs-petg/. [Accessed: Apr. 8, 2025].

[32] Ultralytics. "Coral Edge TPU on Raspberry Pi." *Ultralytics Documentation*, https://docs.ultralytics.com/guides/coral-edge-tpu-on-raspberry-pi/?utm_source=chatgpt.com. Accessed 8 Apr. 2025.

[33] Raspberry Pi Foundation. "Can I Run Windows on Raspberry Pi?" *Raspberry Pi Documentation*, https://www.raspberrypi.com/documentation/computers/os.html#can-i-run-windows-on-raspberry-pi. Accessed 8 Apr. 2025.

[34] Gantt Chart Maker. Free online Gantt chart software. https://ganttchartmaker.com/gantt/#/online. [Accessed on April 8, 2025].

Appendices

Bill of Materials

ITEM	PURPOSE	COST	PROVIDED BY	LINK
RaspberryPi4	Commanding and controlling	84.89	Naitik	Amazon.com: Raspberry Pi 4 Computer Model B 8GB Single Board Computer Suitable for Building Mini PC/Smart Robot/Game Console/Workstation/ Media Center/Etc. : Electronics
Coral TPU	Model inferencing	74.99	N/A	USB Accelerator | Coral
PiCam	Data input	10.00	Eliav	Raspberry Pi Camera Module
Battery Powerbank	SBC power supply	17	Naitik	
Headband Filament	Hardware harnessing	10	Naitik	
Ribbon cable	PiCam-Pi connection	5	Naitik	
M2.5 Screws	Needed to attach pi	10		

Technical Drawings