# Heuristics-Based Vision Language Navigation

Caleb Pong | caleb.pong.2024@gmail.com
GLAMOR Lab
Gretchen Whitney High School, Class of 2024
USC Viterbi | Thomas Lord Department of Computer Science, SHINE 2023

## Introduction

Vision Language Navigation (VLN) has been an increasingly important topic in the intersection of AI and Robotics, and aims to help robots navigate the physical world given natural language instructions. Integrating aspects of computer vision for feature extraction and natural language processing to extract executable instruction from natural language instructions, then utilize a model to execute navigational tasks. Although there have been many attempts to use deep neural networks to perform VLN, much of the time, heuristic rule-based models have proven to work very well, and deep learning models are often compared to a rule-based model as a good indicator of baseline performance.

## Objective & Impact of Professor's Research

My Professor's lab is about the intersection between natural language processing and robotics, and looks to connect language to agent perception and action through interaction. In the world of today's robotics, the integration of natural language processing (NLP) and robotics has allowed robots to understand natural language, enabling them to receive instructions from humans, provide informative responses, and even engage in more sophisticated interactions which may involve ambiguity.

## Acknowledgements

I would like to thank my mentor Abrar for giving me guidance whenever I needed it, and coming with me to the robotics warehouse to work on the robot. Not only did I learn a lot about the work conducted at this lab specifically, I also learned a lot about the world of computer science research in general. He was very helpful and patient with me when explaining concepts that I hadn't learned before.

## Methods

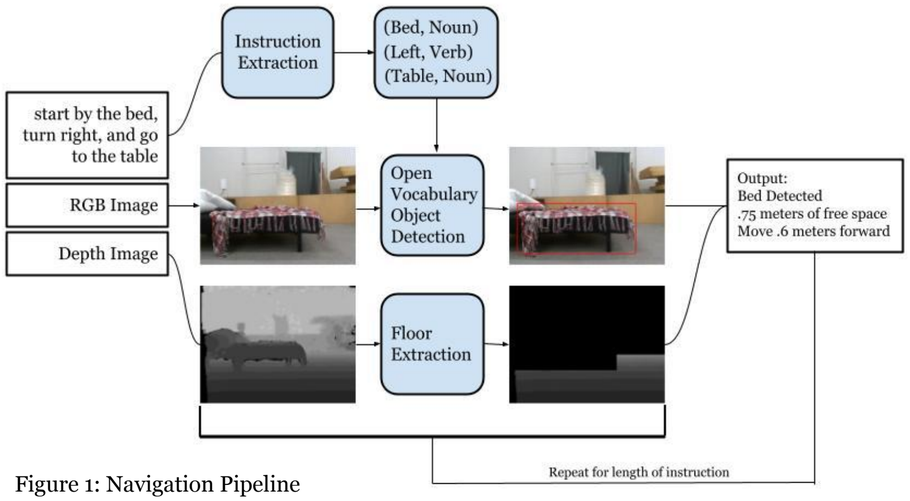


Figure 1: Navigation Pipeline

Overall, the heuristic that the robot used to navigate was as depicted in Figure 1 above. Three key pieces of information were used: user instruction, RGB images, and depth images taken from the robot's camera.Following the sequential order of subgoals in the instruction, the robot utilizes object detection and free space detection to navigate to each object in order.

Figure 2: Object Detection with OWL-VIT

The object detection module implemented open vocabulary object detection using the OwlVIT[1] model, taking an image and a list of target texts as input, and returning bounding boxes for detected objects.


Figure 3: Floor Extraction

Utilizing the Object Detection function, a binary mask for the image can be created by assigning every array value within detected "floor" bounding boxes to 1, then every array value within detected "furniture" or "wall" to 0, which leaves behind only the floor.

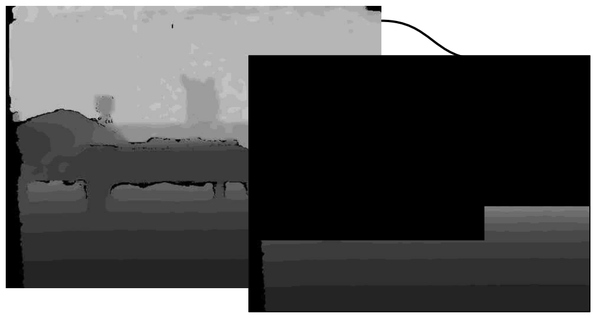

Figure 4: Instruction Extraction

Figure 4 depicts the "translation" of natural language instruction to a list of objects and verbs, which is easier to work with.

"start by the bed, turn right, and go to the table"

```
[]
[['start', 'N'], ['bed', 'N'], ['right', 'V'], ['table', 'N']]
[['bed', 'N'], ['right', 'V'], ['table', 'N']]
```

## Results

The implementation of this model will be tested in the upcoming future. The individual modules of the system, such as floor extraction, instruction extraction, and object detection work with relatively high accuracy, with the instruction extraction providing the desired output 10 out of 10 times, and the object detection finding the target object in 8 out of 10 images.

## Next Steps for You & Advice to Future SHINE participants

This was the first time that I had worked on integrating different aspects of machine learning together into something tangible (the robot), whereas previously I had only worked within the computer. I thought doing this was a fantastic experience, and I plan to continue in the future by continuing to see what interesting problems can be addressed with this approach. For future SHINE participants, my biggest advice would be to make friends with people in or out of your cohort. Making friends will help you stay motivated while working on your project, and help keep you accountable for your work.

## Citations

1. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., ... & Shen, Z. Simple open-vocabulary object detection with vision transformers. arXiv 2022. arXiv preprint arXiv:2205.06230.

2. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., & Lee, S. (2020). Beyond the nav-graph: Vision-and-language navigation in continuous environments. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16 (pp. 104-120). Springer International Publishing.