RSS Project Proposal

Team:

Name	Email Address
Advaith Kandiraju	kandiraju.v@northeastern.edu
Poorna Chandra Vemula	vemula.p@northeastern.edu
Sanjay Prabhakar	prabhakar.sa@northeastern.edu
Samavedam Manikhanta Praphul	samavedam.m@northeastern.edu

Interesting Task

1. What interesting/cool task would your team like the "robot" to do?

An interesting/cool task our robot would do - is to point towards/grasp a particular object in a closed space and dance if some music is put on. Extension to it would be to take speech input and perform corresponding actions.

We plan our task in three phases with the third phase as a stretch task. First phase will be to follow the user's pre-defined instructions. For instance, the user might give an input 'point/grab a red bottle' from the ROS interface. The robot will detect and find the red bottle in its environment, navigate to the location, and then point towards/grab the red bottle using the arm.

Next phase is to make the robot dance as per the tune of the music fed. For example, the user might provide a song (specific format) as an input, then we use the beat, tempo of the song to plan the movements. (This is a cool task we are thinking about). This goes beyond the first phase of fixed set of instructions to little open ended music option

The third phase, i.e. the extension we are planning, is to give the input as "speech" instead of providing it from a ROS interface (as in phase-1). The input speech is processed by an ASR module and the text is passed to an LLM. We promt-tune/fine-tune an LLM to produce necessary actions needed to be taken by the Robot to achieve the given task.

Accomplishing the task

2. How would you accomplish the task? This should provide sufficient details and be at least 2 paragraphs.

To enable the LoCoBot WX250 to point towards or grasp a particular object in a closed space, the first step involves identifying and locating objects. This system would use a combination of computer vision algorithms to recognize specific objects within the robot's environment. Once an object is detected, the robot would calculate its position relative to the robot. The robot would then determine the necessary movements to either point at or grasp the object, taking into account factors such as the object's position, size, shape, and orientation for precise manipulation.

Incorporating the ability for the robot to dance to music and respond to speech commands introduces another layer of complexity. By integrating a microphone and employing audio processing algorithms, the robot could analyze music to detect beats and rhythms, enabling it to perform pre-programmed dance moves in sync with the music. For speech recognition, the robot would use speech-to-text technology to convert spoken commands into text, which would then be processed to understand the intended actions and map it to the action space.

Challenge in task

3. Why is the task challenging and of interest to your team?

- **Sensor fusion:** We intend to superimpose the map produced by stereovision and the point-cloud produced by LiDAR. In this case, attempting to superimpose would be difficult because LiDAR would produce a 2D point cloud whereas stereovision would produce a 3D mapping. Another anticipated problem is that, because the camera(s) on the bot are set at a specific height, the 3D map produced by the Realsense camera will only be optimal in level ground. In other words, uneven terrain will result in an uneven 3D map.

- **Text to Task planning mapping:** For the extension, we are planning to use LLM to obtain the task plan for a desired action. This is an active area of research, which can be a big blocker for our project extension success. Given that large language models consume a lot of resources for the inference, and training, using these models on the robot can be very challenging due to limited compute power resources on the robot. So this might require our team with a two step process of fine tuning existing open source models and then pruning/distilling the fine tuned model for the robot which is both time taking and compute intensive but fun to explore these options.

Our interests about the project:

The task of enabling the robot to point towards or grasp objects, dance to music, and respond to speech commands encapsulates a fascinating blend of robotics and artificial intelligence technologies, including computer vision, audio processing, speech recognition, and machine learning. This multidisciplinary challenge showcases the potential for integrating advanced computational methods to create highly interactive and adaptive robots capable of engaging in complex human environments. The project's focus on sensor fusion, environmental interaction, and the innovative use of large language models for command interpretation represents a cutting-edge approach to overcoming the limitations of robotic communication and action, making it a highly engaging endeavor in the field of robotics.

Robot to use

4. What "robot" will you use and how will you access it?

LoCoBot WX250

Our plan is to attain a perfectly running simulation on Gazebo or Meta Habitat 3.0 before we get our hands dirty with the robot and test our implementation. Once we are ready with the simulation, we will work directly on the LoCoBot and perform the physical tasks. So, our plan would be to perform a simulation as well as a physical demo.

We are willing to take prior permission from PI or any other administrative staff before we get into direct contact with the bot.

Two capabilities of project:

5. What are the two or more capabilities your project will tackle?

The Robot will have the following capabilities:

1. **Perception:**
   The robot should be able to detect objects in the environment. Sensor fusion enables the robot to use both stereo vision and LiDAR to achieve a high level of environmental awareness, allowing the robot to accurately identify and interact with objects.

2. **Mapping:**
   Combining the data from stereovision and LiDAR sensors enables the robot to create detailed 3D maps of its environment. This mapping capability is critical for navigation and interaction within complex spaces, allowing the robot to understand its surroundings and move accurately towards specific objects or locations.

3. **Planning:**
   This involves the robot's ability to understand its environment, determine the best route to a specific location or object, and the movements of the arm to point/grasp the object.

4. **Speech Recognition:**

Incorporates speech recognition to understand and process verbal commands, enhancing human-robot interaction.

Team’s Success

6. How will you measure or demonstrate your team's success?

The robot's ability to demonstrate all of its functions live in front of an audience will be our ultimate achievement. However, our primary goal is to succeed in the simulation phase. Achieving the goal of having the robot operate to its full potential on the simulation platform would be a success. Even so, our goal is to effectively supply both the actual model and a simulation.