PyLoT Robotics 2025 Team Description Paper

Koki Muramoto Katsuya Honda Riku Kinoshita
Towa Yamashita

January 31, 2025

Abstract. This paper introduces Runa, a mobile manipulator newly developed for RoboCup@Home 2025 OPL by middle and high school students from Kaijo Junior and Senior High School PyLoT Robotics, and describes the software development and contributions made by PyLoT Robotics.

We have created our own affordable, production-ready and flexible robotic hardware. Furthermore, based on software using ROS2, we have focused on object recognition for object grasping, voice recognition and output, and task planning using LLM.

Using these, we aim to build a robot system that can handle more general-purpose tasks.

1 Introduction

PyLoT Robotics is a RoboCup team affiliated with Kaijo Junior and Senior High School. The team was founded in 2023 by high school students and all activities related to the team, including administration, development and outreach, are done by junior and senior high school students. The team has participated in regional RoboCup@Home leagues and is actively involved in educational activities for middle school students. We have focused on software development and have continually enhanced it and have continued to enhance it in recognition, grasping, and planning. and have continually enhanced it

This paper focuses on the progress of PyLoT Robotics's development for RoboCup@Home 2025, including spatial object detection using image processing and point cloud processing, improved robot arm control, and improved navigation accuracy for the 2025 competition.

We are also working on various activities make robotics more accessible and to expand the robotics community middle and high school students. to expand the robotics community

2 Hardware

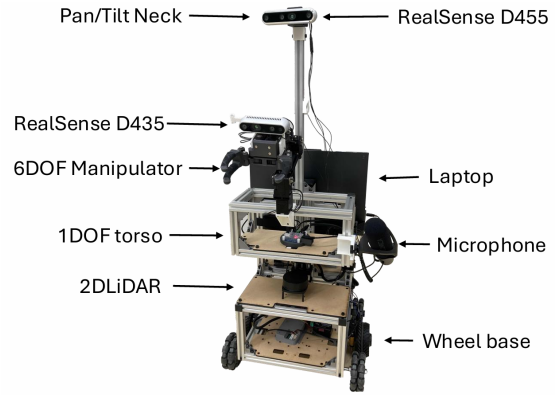

Fig. 1. Runa robot platform

For RoboCup@Home 2025 participation we intend to use a Runa robot which is equipped with a differential two-wheel drive, 6-DoF arm, 1-DOF torso and a pan-tilt-unit with a RGB-D camera. The laptop is equipped with an RTX 3050ti on the back. It uses 2D LiDAR and RGB-D camera to avoid obstacles. The fuselage is made of an aluminum frame, and its high expandability will allow for the addition of monitors and microphones in the future to enhance the robot's interaction.

3 Software

3.1 Software Overview

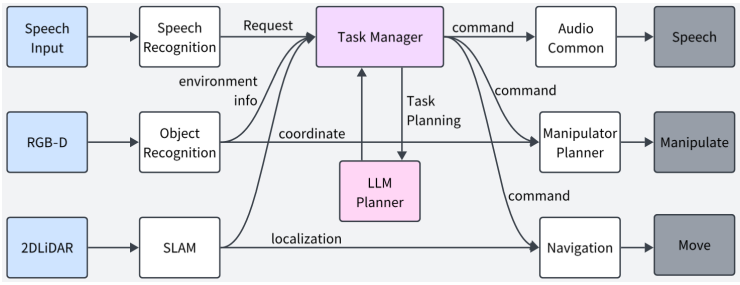

Fig. 2. System Overview

Fig.2 shows the overview of our foundation model-centric system. The system recognizes commands given by humans using voice recognition such as whisper, and plans using LLMs such as GPT. It recognizes objects using information from the RGB-D camera with CLIP[4] and Detic[3], constructs coordinate information with semantic information, and recognizes environment information.

After integrating this information, executable functions are assigned to a state machine and executed.

3.2 Speech Recognition

It uses Whisper[1] to convert speech into text. When you start a conversation, your spoken input is converted into text based on to pre-entered prompts, which are then fed into an LLM such as GPT4[2].

This enables the robot to simultaneously understand what to say from the user's input and what tasks to perform.

3.3 Object Detection

In RoboCup, the object is not known until just before the task, so it is difficult to have CNN learn the object directly. Therefore, we do not use CNN, but use two VLMs for recognition.

First, we perform segmentation of all objects using Detic. This does not narrow down the objects to be recognized, but captures all objects that appear on the screen. Next, we use CLIP to perform zero-shot recognition using prompts. This allows us to change the object to be recognized depending on the prompt we input, and by devising the prompt, we can significantly improve the recognition accuracy.

In addition, by reflecting the semantic information contained in these VLMs in the map, the robot can plan smoothly.

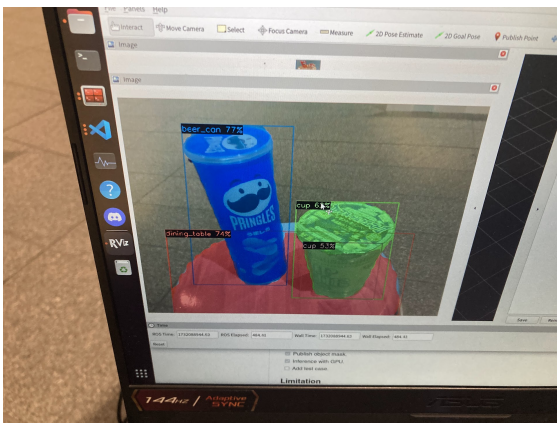

Fig. 3. Object Detection(Detic)

3.4 Human Recognition

The Yolov8[5] Pose package performs person detection and identifies key points of an individual. This allows coordinate recognition of body parts such as nose, eyes, ears, shoulders, hips, elbows, hands, knees, and ankles. It provides a confidence value for each key point. The package also includes the ability to detect the index of a person using the Yolo v8 Pose tracking system. Furthermore, it can determine the position of a person on the map and in relation to the robot, so that the specific room of the house the person is in can be identified. The package analyzes the pose of each person and distinguishes between standing, sitting, arms up, lying down, etc. There are several high-level configurable parameters to customize the results.

In particular, a person is only detected if its feet are visible, which is an important consideration due to the limitations of the RoboCup@Home arena. In addition, there are conditions such as a minimum confidence and a minimum number of key points required to consider a person. The package also incorporates a proximity condition, which ensures that a person is only detected if they are close to the robot. This feature is particularly useful in scenarios such as Receptionist, Sticker for the rules, and Follow you task.

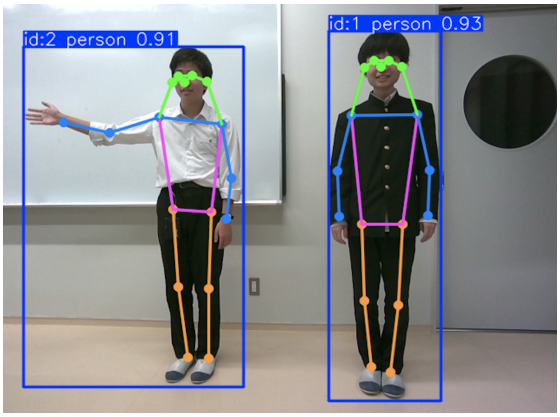

Fig. 4. Person pose detection

3.5 Grasp

We implemented our own inverse kinematics to control the robotic arm.

After recognizing the target object using the aforementioned object recognition system and information from the point cloud, the coordinates are converted to relative coordinates of the robot arm, and the movements of both the cart and the robot arm are planned simultaneously, taking into account surrounding obstacles.

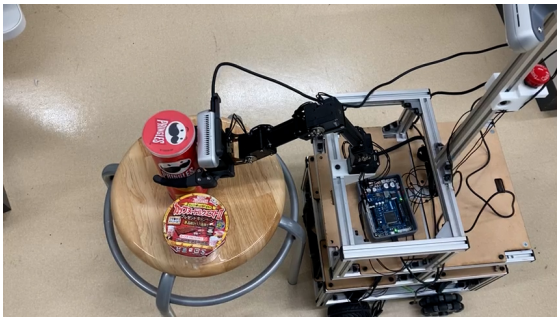

Fig. 5. Grasping object

3.6 Task Planning

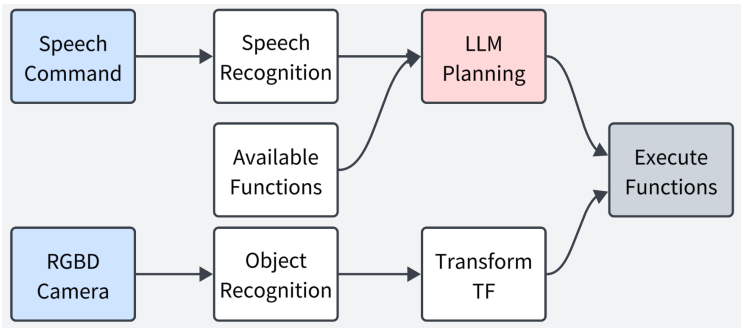

Fig. 6. Overview of a system that uses LLM to perform robot planning from speech

Task planning systems are essential to execute various requests from users in real-world environments. They use LLM to correctly understand linguistic instructions from humans and create action plans.

Taking into account pre-registered personal names, proper nouns, action functions, and other relevant information, the system outputs the action to be executed from the input sentence. This is achieved by stacking int sequence functions and objects in sequence, which operate as a single entity. This is essential for solving GPSR and EGPSR.

3.7 Visual Localization

Fusion of RGB-D camera and 2DLiDAR environment information creates a three-dimensional map by recognizing the environment in three dimensions such

as RTAB-MAP[6]. This allows the robot to deal with protruding desks and other objects that cannot be avoided with a normal 2D map. This is essential for the robot to perform localization and navigation with high accuracy. And for navigation, we use the navigation2 algorithm.[8][9][10] In addition, this map can be used to create a semantic map at the same time. By simultaneously running the object detection mentioned above, navigation and semantic map can be integrated and managed for smooth task planning. This is essential for solving GPSR, EGPSR.

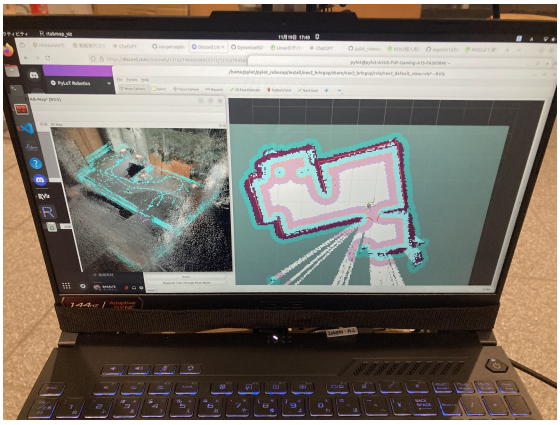

Fig. 7. Visual Localization

4 Our Research

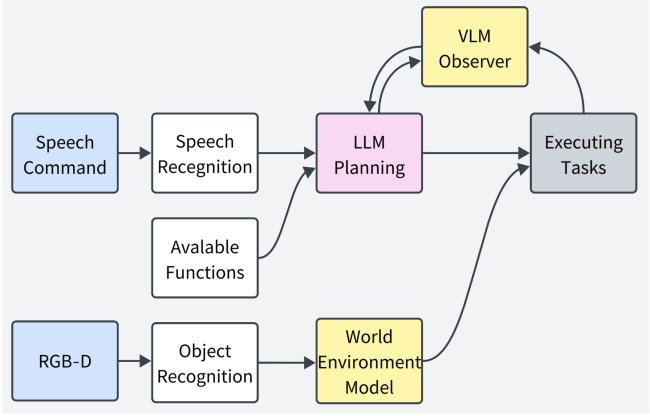

Fig. 8. Diagram of a robot planning system using VLM

We are working on the following research topics to improve the performance of the robot planning system. As mentioned above, we are planning through LLM. In addition, we are aiming to strengthen planning using VLM.

Based on the action plan indicated by the LLM, the Task Executor executes the function. When a state is completed, the LLM sends a prompt to the VLM asking about the robot's status, and the VLM checks whether the task was completed successfully using input from the robot's camera, and provides feedback to the LLM. LLM uses this information to perform replanning, enabling self-recovery. Since the VLM does not have any semantic information about the robot, it focuses on the target object, the target location, and generates appropriate prompts to get the status of the running function. A paper on this subject is currently under conference review.

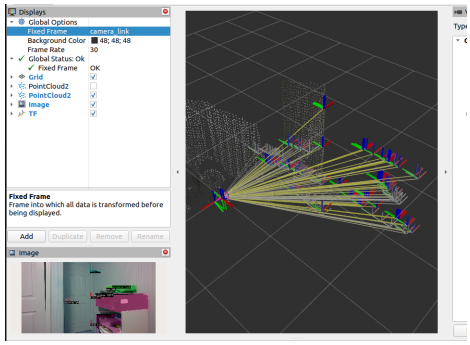

Fig. 9. Semantic map from using VLM

Furthermore, we aim to construct a world environment model with semantic information so that robots can execute general tasks efficiently. The world model, which combines location and semantic information, is inspired by the clip-field[7] and can organize information in a cohesive space.

5 Contribute

At PyLoT Robotics, our goal is not only to build robotic systems for generic tasks, but also to make robotics more accessible to people and to become a broad robotics community for middle and high school students. To this end, we are focusing on external social events and educational activities for junior high school students. For example, we organize publicity and hands-on events for elementary school students, and we create an environment where middle school students can learn practical robot development through step-by-step materials. In this way, making a contribution as a robotics team.

6 Conclusions

This paper describes the robot platform, scientific contributions, and approach of PyLoT Robotics, a RoboCup team from Kaijo Junior And Senior High School, to participate in RoboCup 2025 at RoboCup@Home OPL. By participating in RoboCup@Home, our team aims to develop autonomous mobile home service robots and make robotics more accessible to the public. In accordance with the principles of open source, we are publishing our various activities and results related to RoboCup. We plan to continue contributing to RoboCup by publishing our code and other materials even after the competition ends.

Acknowledgements

This paper is supported by Shimon Ajisaka, Dai Komukai, Yoshihiro Shibata and Takumi Nakamura.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pages 28492–28518. PMLR, 2023.
[2] OpenAI. GPT-4 Technical Report. Technical report, 2023.
[3] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In ECCV, 2022.
[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.
[5] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (Jan 2023), https://github.com/ultralytics/ultralytics
[6] Labbé, M., Michaud, F.: RTAB-Map as an open-source lidar and vi- sual simultaneous localization and mapping library for large-scale and long- term online operation. https://arxiv.org/pdf/2403.06341
[7] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory. In Robotics: Science and Systems, 2023.
[8] S. Macenski, T. Moore, DV Lu, A. Merzlyakov, M. Ferguson, From the desks of ROS maintainers: A survey of modern and capable mobile robotics algorithms in the robot operating system 2. https://github.com/ros-navigation/navigation2
[9] Steve Macenski and Matthew Booker and Josh Wallace. Open-Source, Cost-Aware Kinematically Feasible Planning for Mobile and Surface Robotics. https://arxiv.org/pdf/2401.13078
[10] Macenski, Steve and Martín, Francisco and White, Ruffin and Ginés Clavero, Jonatan. The Marathon 2: A Navigation System. https://github.com/ros-planning/navigation2

Robot Runa Hardware Description

Robot Runa has the patented for garbage recollection. Specifications are as follows:

- Base: wheel base (differential pair), 2.5m/s max speed.
- Arm: 7DOF(1 DOF torso, 6DOF manipulator)
- Neck: 2DOF
- Head: Depth Camera
- Robot dimensions: height: 1.4m (max), width: 0.6m depth 0.8m
- Robot weight: 20kg.
- RGB-D Sensors: Intel RealSense D435, Intel RealSense D455.
- LiDAR: RPLiDAR A1M8
- Microphones: CVM-V30PRO

Fig. 10. Robot Runa

Robot's Software Description

For our robot we are using the following software:

- Platform: ROS2 Humble
- Navigation: Navigation2,AMCL
- Face recognition: Yolov8-Pose,Detic,GPT-4o
- Speech recognition: Whisper,Vosk.
- Speech generation: Audio-Common.
- Object recognition: Detic,CLIP,Yolov8
- Arms control: In-house arm motion planner

The following are the specifications of the laptop mounted on our Robot

- CPU: Ryzen7
- GPU: NVIDIA Geforce RTX 3050ti
- Memory: 16GB