NimbRo@Home 2024 Open Platform League Team Description

Raphael Memmesheimer, Jan Nogga, Alena Savinykh, Evgenii Kruzhkov, Jonas Bode, Bastian Pätzold, Helin Cao, Simon Bultmann, and Sven Behnke

Autonomous Intelligent Systems, Computer Science, Univ. of Bonn, Germany
`nimbroathome@ais.uni-bonn.de`
`https://www.ais.uni-bonn.de/nimbro/@Home/`

Abstract. This team description paper outlines the setup, contributions, and efforts for team NimbRo@Home of the Autonomous Intelligent Systems group from the University of Bonn for their participation at RoboCup@Home Open Platform League taking place in 2024 in Eindhoven, Netherlands. We plan to attend the competition with a modified PAL Robotics TIAGo++ omnidirectional, two-armed robot platform. Further, we describe our intended approaches for object pose and grasp estimation, semantic mapping and human-robot-interaction. Our software contributions can be found at: `https://github.com/AIS-Bonn/`.

1 Introduction

The NimbRo team has a well established track record of successful participations in various robotic competitions ranging from domains like humanoid soccer in the RoboCup AdultSize league, unstructured environments like the DARPA Grand Challenge 2016 to autonomous bin picking challenges like the Amazon Picking Challenge. Recently, in 2022 the NimbRo team won the ANA Avatar XPRIZE challenge. The team already successfully took part in the RoboCup@Home league and won three consecutive international RoboCup@Home competitions (2011 Istanbul [22], 2012 Mexico City [21], 2013 Eindhoven [20] and also won numerous RoboCup@Home German Open challenges. We focused on two-armed manipulation and tool usage in our demonstrations. After reinitiating our domestic service robotics activities, we participated at RoboCup@Home 2023 Bordeaux and ended up in the 4th place. An excerpt from our performance during the RoboCup@Home 2023 in Bordeaux is given in Figure1.

We developed methods for real-time environment and object perception, 3D object pose and grasp estimation using 3D sensors such as laser scanners and RGB-D cameras. We further describe our approaches for object segmentation, mapping and navigation, grasping, audio and natural language processing and behaviour control.

In this paper, we briefly outline the intended robotic platform. Further, we describe our proposed approaches for the RoboCup@Home tasks and give a coarse overview of our behaviour control. Finally, we summarize our domestic service robotics related research.

(a) Receptionist (b) Serving Breakfast (c) Restaurant

Fig. 1: RoboCup@Home 2023 impressions.

2 Hardware

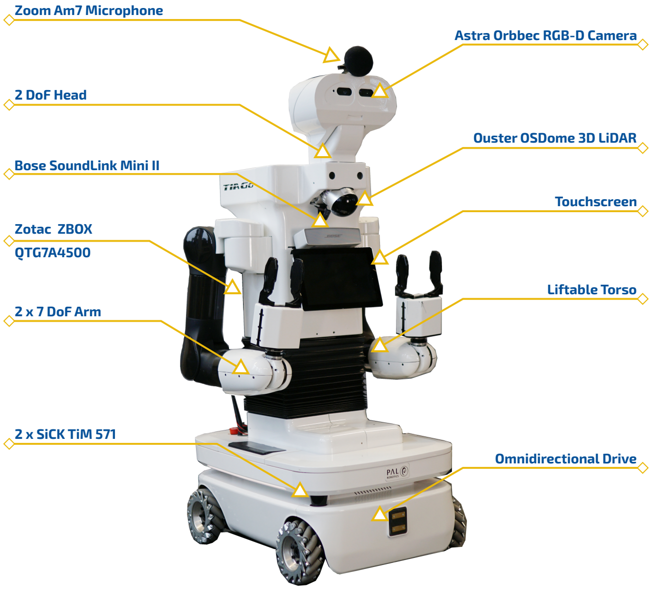


Fig. 2: TIAGo++ omnidirectional robot platform.

For RoboCup@Home 2024 participation we intend to use a TIAGo++ robot (see figure 2) which is equipped with an omnidirectional platform, a linear liftable torso with two 7-DOF arms and a pan-tilt-unit with an RGB-D camera. A ZBOX QTG7A4500 with an NVIDIA RTX A4500, which is used for model inference, is mounted on the back of the robot's torso. An Ouster OSDOME-128 with a 180° FOV is used to gather a wide frontal view of the robot for small obstacle avoidance and precise estimates for distant vision perception. The LiDAR is calibrated against a Logitech Brio webcam with a wide-angle lens. A 10-inch IPS touch screen at the front of the robot and a Zoom Am7 microphone are for human-robot interaction. We aim at using a second, almost identical setup to distribute the development and testing and may integrate distributed robot-robot interaction and extend the robot's individual views by integration into a sensor-edge network [2].

3 Approaches

In this section, we present the approaches of the team related to perception and behaviour control.

3.1 Object Segmentation

Our object detection system relies on a Mask DINO [12] model to generate pixel-wise segmentation for each object instance visible in the color channels of our RGB-D camera, as depicted in Fig. 3. The model is pre-trained on COCO [13] and briefly fine-tuned on annotated data captured in our arena, at RoboCup 2023 and collected from web image searches. Rapid manual data annotation is achieved by interaction with Segment Anything [10] in CVAT [4]. To adapt to the requirements of different RoboCup@Home tasks, tailored datasets featuring pertinent object classes and negative examples are created on the fly using fiftyone [15] to remap labels or remove dataset sections entirely. Used together, these components yield performant detectors which are readily adapted to novel object classes on-site. Our datasets, dataset curation process and training script are available[1].

3.2 Person Recognition

We detect and recognize humans on multiple levels. For human pose estimation we utilize OpenPifPaf [11] which infers a set of 2D keypoints from multiple persons. These keypoints are then projected in 3D using the depth channel of the RGB-D camera. Similarly, we extract human faces using the RetinaFace approach [5]. To remember and re-identify, we augment the faces with descriptors gathered by a learned metric [19] and the analysis of facial attributes. For person tracking, we combine tracked legs from 2D LiDAR data and person detections from camera images, where we augment the person detections with a person descriptor. This allows us to re-identify persons once their track is lost, e.g. by leaving the robot's FOV. We plan to improve the fusion of different detections and modalities and use the wide-angle camera in combination with the Ouster OSDOME 3D LiDAR for accurate person estimation of more distant persons, e.g. in the Restaurant task.

[1] `https://github.com/JanNogga/athome23_detection`

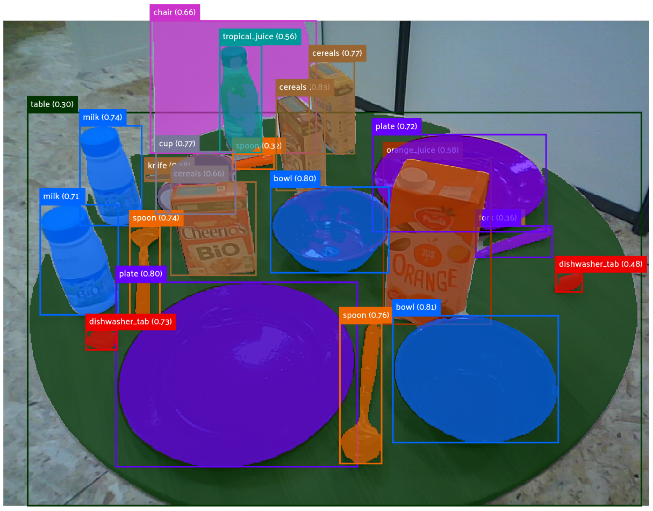


Fig. 3: Object instances as segmented by our fine-tuned Mask DINO [12].

3.3 Mapping and Navigation

We employ the SLAM Toolbox [14] to perform mapping and utilize AMCL for localization. The SLAM toolbox is a graph-based approach with high mapping accuracy due to loop closure support. In the future, we plan to employ the localization of the SLAM Toolbox and expect to improve localization in known environments with many changes. We use two modes, a localization mode for use in known environments and a mapping mode for use in unknown environments. In known environments, we utilize pose location markers to encode poses of interest. In addition, we can use those markers to define regions that can be used to distinguish between different areas (rooms), to determine which area people, the robot or objects belong, or to restrict specified areas for the robot. A visualization of the underlying map is given in Figure 4.

Our navigation approach relies on ROS 2 Navigation. For global path planning, we use the standard A* algorithm that updates the path once per second. Since our robot is omnidirectional, we employ a Timed Elastic Band (TEB) controller [18] that is designed for omnidirectional robots. This controller is aware of the robot dynamics and thus allows replanning trajectories frequently without abrupt motions. We filter the raw LiDAR data from the Ouster OSDOME and the SiCK LiDARs and then construct an aggregated costmap. This costmap is used for planning and integrates dynamic obstacle avoidance. The Ouster OSDOME gives us the ability to avoid hard-to-see obstacles for the 2D LiDARs, e.g. tables, chairs, barriers or small objects. The 3D LiDAR is processed by a crop box, pass through and statistical outlier filters before aggregation into the costmap.

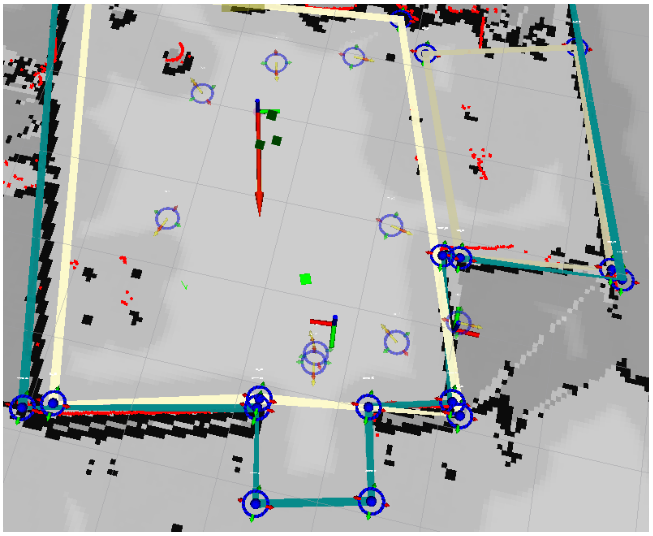


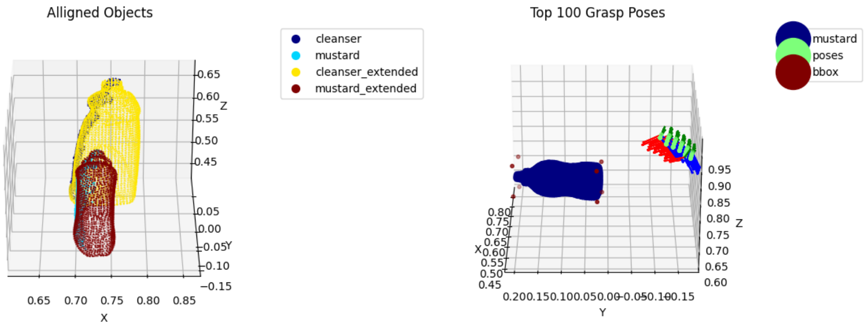
Fig. 4: A constructed map including location markers and annotated regions. Markers show locations of interest. Regions define different rooms, area borders or task-specific regions, e.g. for searching persons.

3.4 Grasping

Grasp planning begins with partial object pointclouds of recognized objects, which are lightly post-processed and registered to a 3D object model when available. This is illustrated in Figure 5 (a). Then, grasp proposals are sampled on a pill-shaped structure around the detected object and filtered to exclude poses which are outside the robot's workspace or in collision with the nearby environment. Additionally, poses in which the object does not fit the gripper width are rejected. Finally, a cost function considering the best available pre-grasp poses, desired object approach modes, a safe distance from obstacles and the surface supporting the object as well as problematic regions in the workspace is employed to find the best grasp proposal. An example is depicted in Figure 5 (b). This cost function is symmetrical for both robot arms, allowing for flexible selection of the best suitable arm and also dual-arm manipulation for grasping two objects at once.

3.5 Audio and Natural Language Processing

The foundation of our audio processing pipeline is the JACK Audio Connection Kit [8], which provides capabilities for real-time audio processing and interfacing to connected audio hardware. To cope with challenging acoustic conditions in downstream tasks, the microphone signal is being pre-processed using the NVIDIA Maxine toolkit [3], applying denoising and dereverberation to isolate speech from environmental noise. To retrieve speech commands at specific times



(a) Object Registration

(b) Grasp Proposals

Fig. 5: (a) If available, 3D models can be registered to partial pointclouds of all detected objects. (b) The grasp proposals minimizing the grasping cost function for approaching a bottle of mustard lying on a surface.

during task execution, we use a voice activity detection model [1] to determine beginning- and end-of-speech boundaries. Speech segments are then forwarded directly to a speech recognition model [17], which is capable of transcribing 99 different languages and translating them into English. Thus, our speech recognition pipeline can be characterized as robust, grammar-free, and multilingual.

For text-to-speech synthesis, we utilize the Coqui.ai library [6], which implements the end-to-end approach of Jaehyeon et al. [9]. We embed this model between custom pre- and post-processing modules for text normalization of numerals and punctuation, as well as loudness normalization between passes and loudness maximization to cut through loud environmental noises.

To comprehend and utilize complex natural language instructions, where naive scanning for specific keywords is not sufficient, we use large language models [16][7]. This allows us to extract relevant information from arbitrary text (and images). In addition, we instruct these models to generate JSON output to autonomously choose between multiple viable functions to advance task completion.

3.6 Behaviour Control

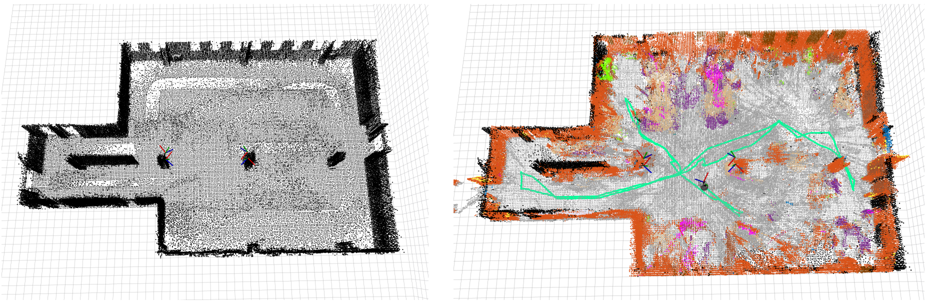
We developed an abstraction layer for integration of novel robot behaviours based on state machines encapsulating basic behaviours and functionalities, keeping in mind that these will be replaced by learned behaviours in the future. We employ an abstraction layer for simplifying higher-level behaviour development of complex state machines by re-using generalized sub-statemachines on various levels of complexity on a functionality and behaviour level. These state-machines are designed such that their execution and individual parameters can be mapped from natural language descriptions. In its current state the state machines are sequences of manually defined behaviours and functionality sequences, however our system is designed to learn these behaviours and functionalities e.g. by demonstrations.

4 Research

In this section, we briefly describe our domestic service robotic related research.

4.1 Visual Pose Estimation with Smart Edge Sensors and Collaborative Semantic Mapping

We developed an external camera-based mobile robot pose estimation approach for collaborative perception with smart edge sensors. Our approach allows for the initialization and correction of a mobile robot's pose from N external static cameras. Furthermore, robot observations from changing viewpoints are fused into the allocentric scene model to extend the view of the static cameras. The robot pose is estimated using a robot detection and keypoint estimation approach trained on a combination of synthetic and real data. The robot pose is then recovered by minimizing the reprojection errors in multiple views. We verify the performance of our approach with various experiments using a Toyota HSR robot in an approximately 250 m^2 lab. First, we demonstrate the external visual pose estimation for initialization of the robot pose in the map, given no or a coarse initial localization. Our approach is shown to perform well for initial camera-based localization. Second, we evaluate the accuracy of our pose estimation approach using an HTC Vive tracking system as a reference. Through the continuous correction via localization feedback from the external cameras,



(a) Initialization

(b) Explored Map

Fig. 6: Resulting semantic map of the proposed collaborative semantic mapping approach with improved consistency by external pose refinement. The approach is initialized with an empty map (a) and a mobile smart edge sensor node explores the environment, resulting in an explored final semantic map (b).

the pose error remains below a few centimeters. We integrate the external pose estimation approach in a collaborative semantic mapping scenario. The external localization improves the consistency of the resulting semantic map. Finally, we demonstrate long-term robustness in a highly cluttered environment (see Figure 6). The laser-based robot-internal localization accumulates a high localization error over longer paths, which leads to unreachable targets or high positional errors at the target locations that could lead to collisions. Our external pose estimation can correct and compensate for the localization errors for better robustness. Recently, we integrated the PAL Robotics TIAGo++ platform with this framework.

5 Conclusions

In this team description paper, we presented the intended robot platform, scientific contributions and intended approaches for an intended RoboCup 2024 participation of team NimbRo in the RoboCup@Home Open Platform League. Our team previously participated successfully in the RoboCup@Home competition. With our intended RoboCup@Home participation, we are aiming to enforce our autonomous mobile domestic service robot research activities. Our most recent developments about our team can be found on our team webpage `https://www.ais.uni-bonn.de/nimbro/@Home/`.

References

[1] Hervé Bredin and Antoine Laurent. "End-to-end speaker segmentation for overlap-aware resegmentation". In: *Proceedings of Interspeech*. 2021, pp. 3111–3115.

[2] Simon Bultmann, Raphael Memmesheimer, and Sven Behnke. "External Camera-based Mobile Robot Pose Estimation for Collaborative Perception with Smart Edge Sensors". In: *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[3] NVIDIA Corporation. *NVIDIA MAXINE Audio Effects*. Nov. 2023. URL: `https://github.com/NVIDIA/MAXINE-AFX-SDK`.

[4] CVAT.ai Corporation. *Computer Vision Annotation Tool (CVAT)*. Version 2.8.2. Nov. 2023. URL: `https://github.com/opencv/cvat`.

[5] Jiankang Deng et al. "Retinaface: Single-shot multi-level face localisation in the wild". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5203–5212.

[6] Gölge Eren and The Coqui TTS Team. *Coqui TTS*. Version 1.4. Jan. 2021. URL: `https://github.com/coqui-ai/TTS`.

[7] Georgi Gerganov. *llama.cpp*. Nov. 2023. URL: `https://github.com/ggerganov/llama.cpp`.

[8] *JACK Audio Connection Kit*. Nov. 2023. URL: `https://jackaudio.org`.

[9] Jaehyeon Kim, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech". In: *Proceedings of the 38th International Conference on Machine Learning*. 2021, pp. 5530–5540.

[10] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: `2304.02643 [cs.CV]`.

[11] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association". In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2021), pp. 13498–13511.

[12] Feng Li et al. *Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation*. 2022. arXiv: `2206.02777 [cs.CV]`.

[13] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: `1405.0312`. URL: `http://arxiv.org/abs/1405.0312`.

[14] Steve Macenski and Ivona Jambrecic. "SLAM Toolbox: SLAM for the dynamic world". In: *Journal of Open Source Software* 6.61 (2021), p. 2783.

[15] B. E. Moore and J. J. Corso. "FiftyOne". In: *GitHub* (2020). URL: `https://github.com/voxel51/fiftyone`.

[16] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: `2303.08774 [cs.CL]`.

[17] Alec Radford et al. "Robust speech recognition via large-scale weak supervision". In: *ArXiv* abs/2212.04356 (2022).

[18] Christoph Rösmann, Frank Hoffmann, and Torsten Bertram. "Timed-elastic-bands for time-optimal point-to-point nonlinear model predictive control". In: *2015 european control conference (ECC)*. IEEE. 2015, pp. 3352–3357.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[20] Jörg Stückler et al. "Increasing Flexibility of Mobile Manipulation and Intuitive Human-Robot Interaction in RoboCup@Home". In: *RoboCup 2013: Robot World Cup XVII [papers from the 17th Annual RoboCup International Symposium, Eindhoven, The Netherlands, July 1, 2013]*. Ed. by Sven Behnke et al. Vol. 8371. Lecture Notes in Computer Science. Springer, 2013, pp. 135–146. DOI: `10.1007/978-3-662-44468-9\_13`. URL: `https://doi.org/10.1007/978-3-662-44468-9%5C_13`.

[21] Jörg Stückler et al. "NimbRo@Home: Winning Team of the RoboCup@Home Competition 2012". In: *RoboCup 2012: Robot Soccer World Cup XVI [papers from the 16th Annual RoboCup International Symposium, Mexico City, Mexico, June 18-24, 2012]*. Ed. by Xiaoping Chen et al. Vol. 7500. Lecture Notes in Computer Science. Springer, 2012, pp. 94–105. DOI: `10.1007/978-3-642-39250-4\_10`. URL: `https://doi.org/10.1007/978-3-642-39250-4%5C_10`.

[22] Jörg Stückler et al. "Towards Robust Mobility, Flexible Object Manipulation, and Intuitive Multimodal Interaction for Domestic Service Robots". In: *RoboCup 2011: Robot Soccer World Cup XV [papers from the 15th Annual RoboCup International Symposium, Istanbul, Turkey, July 2011]*. Ed. by Thomas Röfer et al. Vol. 7416. Lecture Notes in Computer Science. Springer, 2011, pp. 51–62. DOI: `10.1007/978-3-642-32060-6\_5`. URL: `https://doi.org/10.1007/978-3-642-32060-6%5C_5`.

Fig. 7: NimbRo@Home team at RoboCup 2023, Bordeaux (France)

Name of team : NimbRo
Member : Raphael Memmesheimer, Jan Nogga, Jonas Bode, Bastian Pätzold, Helin Cao, Alena Savinykh, Evgenii Kruzhkov, Simon Bultmann, Michael Schreiber, Sven Behnke (TBC)
Contact information : memmesheimer@ais.uni-bonn.de
Website : `https://www.ais.uni-bonn.de/nimbro/@Home/`
Hardware :

- PAL Robotics TIAGo++ Omni Edition

Software :

- ROS
- OpenCV
- SLAM Toolbox
- PCL
- PyTorch
- Whisper
- Custom software:
  - 3D Semantic Scene Perception using Distributed Smart Edge Sensors `https://github.com/AIS-Bonn/SmartEdgeSensor3DScenePerception`
  - Online Marker-free Extrinsic Camera Calibration using Person Keypoint Detections `https://github.com/AIS-Bonn/ExtrCamCalib_PersonKeypoints`
  - Directional TSDF InfiniTAM `https://github.com/AIS-Bonn/DirectionalTSDF`)
  - Real-Time Multi-View 3D Human Pose Estimation using Semantic Feedback to Smart Edge Sensors `https://github.com/AIS-Bonn/SmartEdgeSensor3DHumanPose`
  - ROS transport for high-latency, low-quality networks `https://github.com/AIS-Bonn/nimbro_network`
- More open-source releases can be found here: `https://github.com/AIS-Bonn`