NimbRo@Home 2013 Team Description

Jörg Stückler, David Droeschel, Kathrin Gräve,
Dirk Holz, Michael Schreiber, and Sven Behnke

Rheinische Friedrich-Wilhelms-Universität Bonn
Computer Science Institute VI: Autonomous Intelligent Systems
Friedrich-Ebert-Allee 144, 53113 Bonn, Germany
{ stueckler | droeschel | graeve | holz | schreiber | behnke} @ ais.uni-bonn.de
http://www.NimbRo.net/@Home

Abstract. This document describes the RoboCup@Home league team NimbRo@Home of Rheinische Friedrich-Wilhelms-Universität Bonn, Germany, for the competition to be held in Eindhoven, Netherlands, in June 2013. Our team uses self-constructed humanoid robots for mobile manipulation and intuitive multimodal communication with humans. The paper describes the mechanical and electrical design of our robots Cosero and Dynamaid. It also covers our approaches to object and environment perception, manipulation and navigation control, and human-robot interaction.

1 Introduction

Our team NimbRo competes with great success in the @Home league since 2009, winning the last two RoboCup@Home competitions in 2011 in Istanbul and in 2012 in Mexico City. In the competitions, we successfully participated in most of the tests in stages I and II, and reached the Finals with the highest score. Our final demonstrations achieved the best scorings by the juries. We also participate successfully in local RoboCup GermanOpen competitions, winning in 2011 and 2012.

Our robots, Dynamaid and Cosero, have been designed to balance indoor navigation, mobile manipulation, and intuitive human-robot interaction. We equipped the robots with omnidirectional drives for robust navigation, two anthropomorphic arms for object manipulation, and with communication heads. In contrast to many other service robot systems, our robots are lightweight, inexpensive, and easy to interface.

We investigate methods for real-time object and environment perception using 3D sensors such as laser scanners and RGB-D cameras. Efficient perception is integrated for manipulation of objects and safe navigation in 3D. For manipulation, we developed real-time segmentation of objects on table-tops and shelf layers. To estimate the 6-DoF pose of objects, e.g., to approach larger objects during mobile manipulation, we developed 3D modelling and real-time pose tracking of the objects. To grasp objects in complex scenes, e.g., to pick objects from bins, we integrate motion planning with efficient grasp planning on the

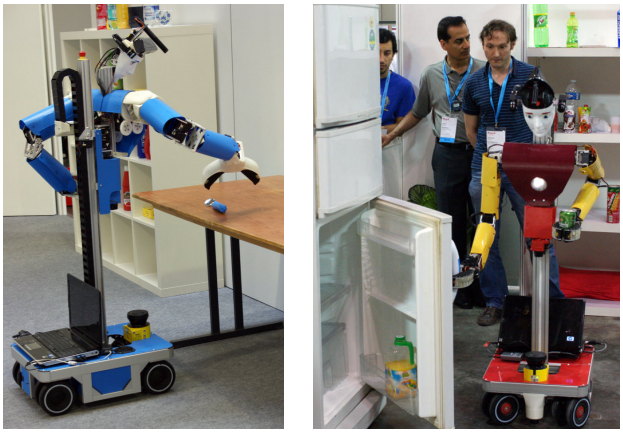

Fig. 1. Left: Cognitive service robot *Cosero* grasps a spoon. Right: *Dynamaid* manipulates the fridge.

detected objects. Simple situations, in which the direct reach to the object is not obstructed, are rapidly detected and tackled using fast parametrized motion primitives.

In the next section, we detail the mechanical and electrical design of our domestic service robots. Sections 3 and 4 cover perception and behavior control, respectively.

2 Mechanical and Electrical Design

We equipped our robots Cosero and Dynamaid (see Fig. 1) with omnidirectional drives to maneuver in the narrow passages found in household environments. Their two anthropomorphic arms resemble average human body proportions and reaching capabilities. A yaw joint in the torso enlarges the workspace of the arms. In order to compensate for the missing torso pitch joint and legs, a linear actuator in the trunk can move the upper body vertically by approx. 0.9 m. This allows the robots to manipulate on similar heights like humans.

The robots have been constructed from light-weight aluminum parts. All joints are driven by Robotis Dynamixel actuators. These design choices allow for a light-weight and inexpensive construction, compared to other domestic service robots. While each arm of Cosero has a maximum payload of 1.5 kg (Dynamaid: 1 kg) and Cosero's drive has a maximum speed of 0.6 m/sec (Dynamaid: 0.5 m/sec), Cosero's low weight of ca. 32 kg (Dynamaid: ca. 20 kg) requires only moderate actuator power. This makes the robots inherently safer than a heavy-weight industrial-grade robot.

Compared to its predecessor Dynamaid [10], we increased payload and precision of Cosero by stronger actuation. Cosero is mainly driven by Dynamixel EX-106+ (10.7 Nm holding torque, 154 g) and RX-64 (6.4 Nm holding torque,

116 g) actuators. The strongest joints in the robot are the shoulder pitch joints with a holding torque of 42.8 Nm. Each of these joints is actuated by two EX-106+ in parallel via a 2:1 transduction. We also improved safety and appearance of the robot with 3D-printed covering for joints and an energy chain in the torso.

The robots perceive their environment with a variety of complementary sensors. A SICK S300 laser scanner measures the distance to objects in a height of approx. 24 cm within 30 m maximum range and with a 270° field-of-view. It is primarily used for 2D mapping and localization. In order to detect small obstacles on the floor in front of the robots, a Hokuyo URG-04LX laser scanner is mounted between the front wheels. It scans in a height of 3 cm. The robots also sense the environment in 3D with a tilting Hokuyo UTM-30LX in their chest (max. range 30 m) and a Microsoft Kinect RGB-D camera in their head that is attached to the torso with a pan-tilt unit in the neck. A second URG-04LX laser scanner is attached through a roll joint to the torso. In horizontal alignment, its scan plane is adjusted to be 2 cm above the surface height when the robot manipulates on tables or in shelves. Its height above the ground can be adjusted from ca. 0.13 m to 1.03 m with the linear joint in the trunk.

We mounted the RGB-D camera on the head for several reasons: First, since the robots have a similar body height (1.6 m default height) like humans, faces can be viewed from the front. The fact, that we as humans design our environment to be easily perceivable with our own sensing capabilities, further supports to perceive the world from human eye height. The placement of the sensor on a pan-tilt neck enables the robot to point its sensors towards targets in a human-like way, i.e., humans can easily interpret the robot's gaze. We use all laser scanners and the depth camera for obstacle detection. For robust manipulation, the robots can measure the distance to obstacles directly from the grippers.

Finally, the sensor head also contains a shotgun microphone for speech recognition. By placing the microphone on the head, the robots point the microphone towards human users and at the same time direct their visual attention to her/him.

3 Perception

3.1 Perception of Human Interaction Partners

For human-robot interaction, a key prerequisite for a robot is awareness of the whereabouts of people in its surrounding. We combine complementary information from laser range finders (LRFs) and vision to continuously detect and keep track of people [11]. Using the VeriLook SDK, we implemented a face enrollment and identification system. In the enrollment phase, our robots approach detected persons and ask them to look into the camera. The extracted face descriptors are stored in a repository. If the robot meets a person later, it compares the new descriptor to the stored ones, in order to determine the identity of the person.

Gestures, like pointing or showing are a natural way of communication in human-robot interaction. A pointing gesture, for example, can be used to draw

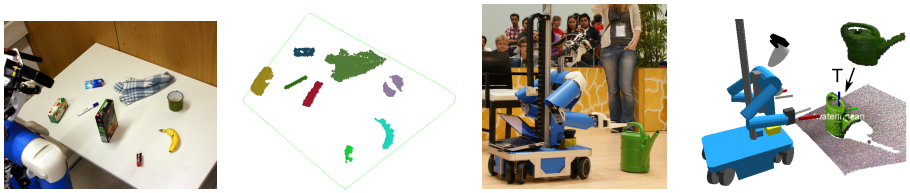

Fig. 2. From left to right: Table-top scene; segmentation into objects; Cosero approaching a watering can during the RoboCup 2012 final in Mexico City; a 3D multi-resolution surfel map of the watering can is aligned with RGB-D images in real-time to estimate the relative pose of the object.

the robot's attention to a certain object in the environment. We implemented the recognition of pointing gestures, showing of objects, waving, and stop gestures. The primary sensor in our system for perceiving a gesture is the RGB-D camera mounted on the robot's pan-tilt unit. We determine the position of the head, hand, shoulder, and elbow which allows us to interpret gestures [3]. The perception is based on the detection of body parts in amplitude images as well as body segmentation in three-dimensional point clouds of the camera. Some gestures such as pointing are further interpreted for their parameters, e.g., for the pointing direction.

We apply the commercial Loquendo [7] system for speech recognition and synthesis. Loquendo's speech recognition is grammar-based and speaker-independent. Its grammar definition allows rules to be tagged with semantic attributes. For instance, one can define keywords for actions or attributes like "unspecific" for location identifiers such as "room". When Loquendo recognizes a sentence that fits to the grammar, it provides the recognized set of rules together with a semantic parse tree. Our task execution module then interprets the resulting semantics and generates appropriate behavior.

3.2 Self-Localization and Mapping

We use state-of-the-art methods for simultaneous localization and mapping in 2D representations of the environment [5]. We use adaptive Monte Carlo Localization (MCL) to estimate the robot's pose in a given occupancy grid map. We also developed localization and mapping in 3D using surfel grid maps [6]. Mapping in 3D allows for fully judging traversability for navigation, while 3D localization can be made more robust against dynamic changes in the environment than localization in a 2D horizontal plane, since it can exploit relevant information from all heights in the environment.

3.3 Perception of Objects

For object perception we develop approaches that combine depth sensing and vision (see Fig. 2). From Kinect depth images, we extract the surface on which

the objects are located through efficient RANSAC methods [14]. We cluster the remaining measurements to obtain a segmentation into objects and keep track of these detections.

Our robots recognize objects by matching SURF features [1] in RGB images to an object model database [10] and by enforcing spatial consistency between the features. In addition to the SURF feature descriptor, we store feature scale, feature orientation, relative location of the object center, and orientation and length of principal axes in the model. During recall, we efficiently match features between an image and the object database according to the descriptor using kd-trees. Each matched feature then casts a vote to the relative location, orientation, and size of the object. We consider the relation between the feature scales and orientation of the features to achieve scale- and rotation-invariant voting. When unlabelled object detections are available through planar RGB-D segmentation (see above), we project the detections into the image and determine the identity of the object in these regions of interest.

During mobile manipulation, the pose of objects needs to be retrieved and tracked in real-time to robustly compensate for the motion of the robot. We thus developed model learning and real-time tracking of objects [15] (see Fig. 2). We successfully applied our method for tracking a table during human-robot cooperative carrying of the table [13]. Core to our approach is the compact representation of RGB-D images in multi-resolution surfel maps [15]. We extract such maps from 640×480 VGA images in just a few milliseconds. We devised a robust registration method that allows for registering two VGA images in real-time at about 20 Hz, and that tracks objects at full 30 Hz frame rate. Full-view models of the objects are acquired in a SLAM approach by moving the camera around the object.

To speed up the process of detecting collisions during manipulation planning [8], we employ a multiresolution height map that extends our prior work on multiresolution path planning [2]. Our height map is represented by multiple grids that have different resolutions. Each grid has $M \times M$ cells containing the maximum height value observed in the covered area (Fig. 3). Recursively, grids with quarter the cell area of their parent are embedded into each other, until the minimal cell size is reached. With this approach, we can cover the same area as a uniform $N \times N$ grid of the minimal cell size with only $\log_2((N/M)+1)M^2$ cells. Planning in the vicinity of the object needs a more exact environment representation as planning farther away from it. This is accomplished by centering the collision map at the object. This approach also leads to implicitly larger safety margins with increasing distance to the object.

4 Behavior Control

The autonomous behavior of our robots is generated in a modular control architecture. We employ the inter process communication infrastructure and tools of the Robot Operating System (ROS) [9].

We implement task execution, mobile manipulation, and motion control in hierarchical finite state machines. The task execution level is interweaved with human-robot interaction modalities. For example, we support the parsing of natural language to understand and execute complex commands.

Tasks that involve mobile manipulation trigger and parametrize sub-processes on a second layer of finite state machines. These processes configure the perception of objects and persons, and they execute motions of body parts of the robot. The motions themselves are controlled on the lowest layer of the hierarchy and can also adapt to sensory measurements.

4.1 Motion Control

We implemented omnidirectional driving controllers for the mobile base of our robots [10]. The driving velocity can be set to arbitrary combinations of linear and rotational velocities. We control the 7-DoF arms using differential inverse kinematics with redundancy resolution. The arms also support compliant control in task-space [12].

4.2 Robust Indoor Navigation

For navigation, we implemented path planning in occupancy grid maps and 3D obstacle avoidance using measurements from the LRFs and the depth camera [4]. To enlarge the narrow field-of-view of the depth camera, we implemented active gaze control strategies.

4.3 Mobile Manipulation

To robustly approach mobile manipulation tasks we integrate object perception, safe navigation, motion primitives, and motion planning. Our robots can grasp objects on horizontal surfaces like tables and shelves efficiently using fast grasp planning [14]. We derive grasps from the top and the side directly from the raw object point clouds. The grasps are then executed using parametrized motion primitives, if the direct reach towards the object is not obstructed. In complex scenarios, such as in bin-picking, the robots plan collision-free grasps and reaching motions [8].

The robots can also carry the object, and hand it to human users. When handing an object over, the arms are compliant in upward direction so that the human can pull the object, the arm complies, and the object is released. We also developed solutions to pour-out containers, to place objects on horizontal surfaces, to dispose objects in containers, to grasp objects from the floor, and to receive objects from users. Based on compliant control, we also implemented mobile manipulation controllers to open and close doors, when the door leaf can be moved without the handling of an unlocking mechanism.

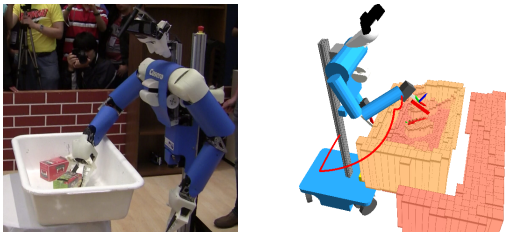

Fig. 3. Motion planning in a bin-picking scenario. We extend grasp planning on object segments with motion planning (reaching trajectory in red, pregrasp pose as larger coordinate frame) to grasp objects from a bin. For collision avoidance, we represent the scene in a multi-resolution height map. We decrease the resolution in the map with the distance to the object. This reduces planning time and models safety margins that increase with distance to the object.

4.4 Intuitive Human-Robot Interfaces

Domestic service robots need intuitive user interfaces so that laymen can easily control the robots or understand their actions and intentions. Speech is the primary modality of humans for communicating complex statements in direct interaction. For speech synthesis, we use the commercial system from Loquendo. Loquendo's text-to-speech system supports natural and colorful intonation, pitch and speed modulation, and special human sounds like laughing or coughing. We also implemented pointing gesture synthesis as a non-verbal communication cue for the robot. Cosero performs gestures like pointing or waving. Pointing gestures are useful to direct a user's attention to locations and objects.

5 Conclusion

The described system has been evaluated for four years now at RoboCup German Open and RoboCup competitions in 2009 to 2012. In all competitions, it performed very well, winning the last two competitions in 2011 and 2012. We developed and integrated several state-of-the-art approaches to perception and control for navigation, mobile manipulation, and human-robot interaction. Our robots avoid collisions using 3D measurements, perform simultaneous localization and mapping, and plan paths in the maps. They detect objects on planar surfaces, recognize them, and track their 6-DoF pose in real-time. For grasping objects, the robots plan grasps on the detected objects efficiently, and decide between a fast direct reach or motion planning in complex scenes. Compliant control enables safe human-robot interaction and extends manipulation capabilities, e.g., to close and open doors without explicit articulation models.

We plan to equip Dynamaid and Cosero with more expressive communication heads. We will continue to improve the system for RoboCup 2013 and to integrate new capabilities, such as on-line learning of novel objects and tool-use. The most recent information about our team (including videos) can be found on our web pages www.NimbRo.net/@Home.

Team Members

Currently, the NimbRo@Home team has the following members:[1]

- Team leader: Jörg Stückler, Prof. Sven Behnke
- Staff: David Droeschel, Kathrin Gräve, Dirk Holz, and Michael Schreiber
- Students: Manus McElhone

References

1. H. Bay, T. Tuytelaars, and L. Van Gool. SURF: speeded up robust features. In *9th European Conference on Computer Vision*, 2006.
2. Sven Behnke. Local multiresolution path planning. *Robocup 2003: Robot Soccer World Cup VII, Springer LNCS*, pages 332–343, 2004.
3. D. Droeschel, J. Stückler, and S. Behnke. Learning to interpret pointing gestures with a time-of-flight camera. In *In Proc. of the 6th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2011.
4. D. Droeschel, J. Stückler, D. Holz, and S. Behnke. Using time-of-flight cameras with active gaze control for 3D collision avoidance. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.
5. G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Trans. on Robotics*, 23(1), 2007.
6. J. Kläß, J. Stückler, and S. Behnke. Efficient mobile robot navigation using 3d surfel grid maps. In *Proc. of the 7th German Conference on Robotics (ROBOTIK)*, 2012, to appear.
7. Loquendo S.p.A. Vocal technology and services. http://www.loquendo.com, 2007.
8. M. Nieuwenhuisen, D. Droeschel, D. Holz, J. Stückler, A. Berner, J. Li, R. Klein, and S. Behnke. Mobile bin picking with an anthropomorphic service robot. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, to appear 2013.
9. M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. ROS: an open-source Robot Operating System. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2009.
10. J. Stückler and S. Behnke. Integrating Indoor Mobility, Object Manipulation, and Intuitive Interaction for Domestic Service Tasks. In *Proc. of the IEEE Int. Conf. on Humanoid Robots (Humanoids)*, 2009.
11. J. Stückler and S. Behnke. Improving people awareness of service robots by semantic scene knowledge. In *Proc. of the RoboCup Int. Symposium*, Singapore, 2010.
12. J. Stückler and S. Behnke. Compliant task-space control with back-drivable servo actuators. In *Proc. of the RoboCup Int. Symposium*, Istanbul, Turkey, 2011.
13. J. Stückler and S. Behnke. Following human guidance to cooperatively carry a large object. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Bled, Slovenia, 2011.
14. J. Stückler, R. Steffens, D. Holz, and S. Behnke. Efficient 3D object perception and grasp planning for mobile manipulation in domestic environments. *Robotics and Autonomous Systems*, 2012.
15. Jörg Stückler and Sven Behnke. Model learning and real-time tracking using multi-resolution surfel maps. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

[1] Funding for the project is provided by Deutsche Forschungsgemeinschaft (DFG) under grant BE 2556/2 and Rheinische Friedrich-Wilhelms-Universität Bonn.