On RoboCup@Home — past, present and future of a scientific competition for service robots

Dirk Holz[1], Javier Ruiz del Solar[2], Komei Sugiura[3], and Sven Wachsmuth[4]

[1] Autonomous Intelligent Systems Group, University of Bonn, Germany
[2] Department of Electrical Engineering & AMTC, Universidad de Chile, Chile
[3] National Institute of Information and Communications Technology, Japan
[4] Center of Excellence Cognitive Interaction Technology, Bielefeld University, Germany

Abstract. RoboCup@Home is an application-oriented league within the annual RoboCup events. It focuses on domestic service robots and mobile manipulators interacting with human users. Participating robots need to solve tasks ranging from following and guiding human users to delivering objects, e.g., in a supermarket.
In this paper, we present the @Home league and how it evolved over the last seven years since its existence. We place particular emphasis on how we evaluate the teams' performances over the years and how we use the obtained statistics to drive the development of the league. This process is shown in detail on two examples—following human guides and object search and retrieval. Finally, we will outline possible future directions and developments.

1 Introduction

Scientific competitions are becoming more and more common in many research areas of Artificial Intelligence and Robotics. They provide a common testbed for comparing different solutions and enable exchange of research results. Furthermore, they are interesting for general audience and industries. Particularly interesting in a many of these competitions, is the opportunity of defining standard benchmarks for solving specific problems, comparing different solutions, and making the best solutions available to the community. Moreover, the tasks and the results of the competitions are often used in scientific papers to compare new approaches with respect to existing ones.

One of the classic competitions in AI and robotics research is RoboCup. RoboCup started in 1997 with the ambitious goal of bringing forth a team *"of fully autonomous humanoid robot soccer players"* that *"shall win a soccer game, complying with the official rules of FIFA, against the winner of the most recent World Cup"*. It incorporates many interesting problems from both AI and robotics research. The central point in the first years (up to now), was to have highly sophisticated gameplay in abstracted environments. That is, from year to year the complexity is increased by increasing the complexity of environment (e.g., color-coded goals etc.) and setup (e.g., no shot on goal without previous passing).

RoboCup@Home follows a different (and, in fact, inverted) approach: it started with very simple tasks in a complex (real) environment as opposed to the highly

complex task of playing soccer in a simplified environment. RoboCup@Home is a competition where domestic service robots are performing several tasks in a home environment, interacting with people and with the environment in a natural way. Each test requires a combination of different functionalities (like navigation, object perception and manipulation, person detection and tracking, etc.) and the score is related to the accomplishment of the task. RoboCup@Home started in 2006 and has the main characteristic of changing tasks every year while maintaining the same basic functionalities. By changing the difficulty and the combinations of the functionalities to be integrated, we aim at pushing the teams to develop general and robust solutions. Indeed, with this setting of the competitions, it is too difficult to implement many specific systems to solve each of the tasks instead of one general solution for all tasks.

This paper presents the RoboCup@Home league in detail discussing its history and present rules as well as the system used to drive changes and possible future developments.

2 History of the RoboCup@Home league

The first ideas on RoboCup@Home have been proposed in 2005 by Tijn van der Zant and Thomas Wisspeintner [1]. The competition was held as a demonstration at RoboCup 2006 in Bremen, Germany. In 2007, RoboCup@Home became an official league and was featured in both the main RoboCup competition in Atlanta and in local RoboCup competitions. Eleven teams participated in this first competition. At that time, the competition consisted of simple tasks testing only single basic functionalities such as navigation, following a person or finding an object. Every test was conducted in two phases. In the first phase, teams could customize (and simplify) the test setup, for example, by using own objects, artificial markers or own team members who knew how to interact with the robot. In the second phase, tasks were conducted as they had been planned by the Technical Committee (TC): the robot was operated by an independent referee, an official set of objects (not known to the teams before the competition) was used and no simplifications to the environment were allowed. Obviously, teams who had a general solution could score in both phases with the same approach.

After these first two years of RoboCup@Home competitions, we discussed how to evaluate the performance of the competition over the years in terms of improvement demonstrated by the teams. Three problems were identified: 1) improvement is difficult to measure because the tasks change every year, 2) performance is difficult to evaluate if situations differ, and 3) the score system, based on Boolean scores (either success or failure of the entire test), was not adequate for this analysis. While the first problem is inherent of a dynamic yearly competition, the second and third could be addressed. Therefore, since 2008, the score system was changed to have scores not per task but per sub-task. Moreover, since most teams in 2007 came up with general solutions (as expected and intended), the first phase was removed and, nowadays, all tasks are run as defined by the Technical Committee and in the very same fashion

Fig. 1. Teams and robots at the 2013 RoboCup@Home competition in Eindhoven.

(involved objects, people, situations, etc.) for all teams. The latter aspect provides a better testbed for benchmarking the presented approaches. Since 2008 up to the upcoming RoboCup 2014 in João Pessoa, this new score system was used together with a systematic analysis carried out every year in order to find out how the rules should be changed (e.g., keeping tasks, making (sub-)tasks more complex, removing or adding sub-tasks etc.).

With becoming a regular league and with the improved scoring system, the number of participating teams quickly grew. Since 2009, RoboCup@Home has a stable average number of participating teams around 19 to 20 teams from all over the world, with a peak attendance of 24 teams (maximum number of teams allowed) in Singapore, 2010.

The development of the league can best be seen in the evolution of tests conducted within the competition. In 2007, tests have been as simple as navigating in a static apartment (with enough time to build static maps beforehand), following a human guide gently walking in front of the robot without further disturbances, or finding an object in the apartment. Nowadays, the league features, amongst others, following a human guide through a crowded public place with various disturbances like people blocking path and/or sight between robot and guide, navigating and manipulating objects in previously unknown public places like supermarkets, or perform any task being asked for on demand where the task is only restricted to belong to the expected capabilities in the league (over all tests). The development can also be seen in details such as the complexity of object recognition: over the years not only the number of objects throughout the tests has been steadily increased (set of objects known to the teams), but we also introduced unknown objects which pose particular problems on object perception pipelines.

Naturally, developing a robotic system integrating all the needed capabilities is very challenging. For example, finding and manipulating an object is not only

Fig. 2. Typical RoboCup@Home arena (left) and objects (right).

complex (especially in uncontrolled environments) but was also solved by only very few teams in the beginning. However, in the past years we could observe an increase in the performance in both individual capabilities and integrated systems being capable of all functionalities and able to score in all tasks.

3 The Current Competition—Tasks and Sub-tasks

The RoboCup@Home competition runs in a realistic setting where an apartment with different functional rooms and typical furniture and objects is realized (see Figure 2). Since it is not completely specified beforehand, it may differ in its implementation and teams do not know any information (such as number of rooms, dimensions, material of the floor, colors of the wall, kinds of furniture and objects, etc.) before arriving at the competition venue. This ensures the development of general solutions. Moreover, the arena is subject to constant changes: especially minor changes such as moved furniture or objects lying on the ground may happen right before or even during tests. Furthermore, some tasks are conducted outside this area in a public space such as a restaurant or a shopping mall (that is known to the teams beforehand).

The competition is formed by about eight predefined tasks (called *tests*), two open demonstrations, a Technical Challenge and the finals. A stage system is used: Stage I is performed by all participating teams. Then the (better) half of them—or at least the best ten teams—advance to Stage II, and finally only the best five teams reach the finals. Previously achieved scores and a jury evaluation in the finals determine the winner of the competition.

The main functionalities required in the tests are the following (chosen as to reflect common capabilities a general purpose service robot in a domestic environment should posses): navigation and mapping, person recognition and tracking, object recognition and manipulation, and speech and gesture recognition. In addition, integration of the functionalities and higher level cognitive skills are of particular interest.

Each test requires a combination of some of these functionalities. For example, *"Follow me"* is a test in which the robot has to follow a person in a crowded area of the venue of the competition, enter an elevator with the person in order reach

a location far away and on a different floor with respect to the starting position. The guide is not known in advance by the robot and a quick automatic calibration procedure must be done at the beginning of the test, when the person appears in front of the robot. During the test, other persons are allowed to pass between the robot and the guide and at some point the guide hides away from the view of the robot, that must be able to reacquire his/her position. Finally, entering and exiting the elevator is guided by speech or gestures. This test integrates navigation, person tracking, person recognition and speech/gesture recognition.

Another test is *Cocktail Party*, in which the robot has to welcome unknown guests in the apartment. Five people (unknown to the robot) are in a room of the apartment either sitting or standing. When the robot enters the room, three of these people (one after the other) call the robot by waving and order a drink by speech command when the robot gets close to them. The robot has to go to the kitchen, grab the drink ordered by the person and bring it back to him/her. This test integrates navigation, person and speech recognition, object detection and manipulation.

The *Restaurant* test is executed in a real restaurant (in previous years in a real supermarket). The robot is guided by a user (a team member) through the environment (unknown to the robot) and some locations (e.g., tables and shelves with drinks and food) are described by the user to the robot during this visit. Then the robot receives an order to bring specific food or drink items to some of the locations previously visited and the robot is expected to reach the shelves, grasp the correct items and bring them to the correct locations. This test integrates navigation, mapping, person tracking, speech recognition, object detection and manipulation.

Finally, the test *General Purpose Service Robot* focuses on the ability of the robot to understand its goals and reason on them. The task is not specified beforehand, but generated by a random command generator and it is given to the robot through a spoken command. The robot has to understand the desired goal and to accordingly plan actions to execute an appropriate behavior. For example, a user request may be "bring me a drink" and the robot has to acquire possibly missing information and then plan a sequence of actions to go to a location in which drinks are, grab one and bring it to the user. In this test, all the functionalities may be required (the task is unknown and each possible task actually requires different functionalities), but in particular the cognition-related functionalities must demonstrate the ability of the robot of understanding the current situation as well as the user request and of performing a complex task not specified beforehand. Moreover, the user requests may feature missing information (e.g., underspecified aspects that the robot has to ask for), erroneous information (e.g., wrong aspects the robot has to cope with), and sequences of tasks (i.e., giving a sequence of three commands).

The tests in the competition are of two kinds: *standard tests*, in which the functionalities and their combination are decided by the Technical Committee, and *open tests*, for which each team can decide which functionalities to show. Standard tests are evaluated by a partial score system described in the next

section, while open tests are evaluated in a peer-to-peer fashion with different juries (jury of team leaders, technical committee jury, executive committee jury, and external jury).

It is important to notice that the these tests are improved (or replaced) every year, by making them more difficult and with unpredicted and difficult situations occurring. This evolution is important in order to prevent development of local optima solutions that specialize too much on a particular instance of the problem without general applicability.

4 Score System and Analysis

Tests in the RoboCup@Home competition are divided into multiple phases (or sub-goals) and each phase, when accomplished, provides the team with a score. The total score of the test is thus given by the sum of the scores of all the accomplished phases. If all the phases are correctly performed a full score is gained, otherwise only a partial score is collected. Each phase in a test is evaluated in a Boolean manner: it is either fully accomplished or not accomplished at all.

This definition of the score is used to compare and rank the teams and thus to provide the final results of the competition. However, in order to analyze more in detail the results of teams during the competition and to compare results over the years, we need a method to measure the performance of the teams in the tests with respect to the desired functionalities. This further analysis does not affect the final results of the competitions, but it is used to evaluate the performance of the entire competition, as discussed in the next section.

To this end, we associate to each phase of a standard test a set of functionalities that are required to achieve it. When a phase is successfully accomplished, we can state that the functionalities associated with it have been successfully implemented. On the other hand, if the phase is not accomplished, we can state that at least one of the associated functionalities was not successful, but we cannot say exactly which one, since we do not have any access to the internal state of the system under test.

To relate phase scores with the functionalities, we also define a *weight* for each functionality in each phase of a test. This weight is a value in $[0, 1]$ and the weights of all the functionalities associated to a phase sum to 1. These weights can be intuitively explained as the percentage of contribution of a given functionality to achieve a phase of the test, or, in other terms, the probability that it has been the cause of failure, if the phase is not accomplished.

Obviously, determining the weights is not a straightforward task. In practice, we are using an estimation based on our experience and discussed within the league's Technical Committee. Although approximated, we believe that the results obtained in this way are useful to evaluate the overall progress of the competition based on average performance of the teams.

The score system allows analyzing the results of the teams in the different functionalities; moreover, it can be used to evaluate the progress of the entire

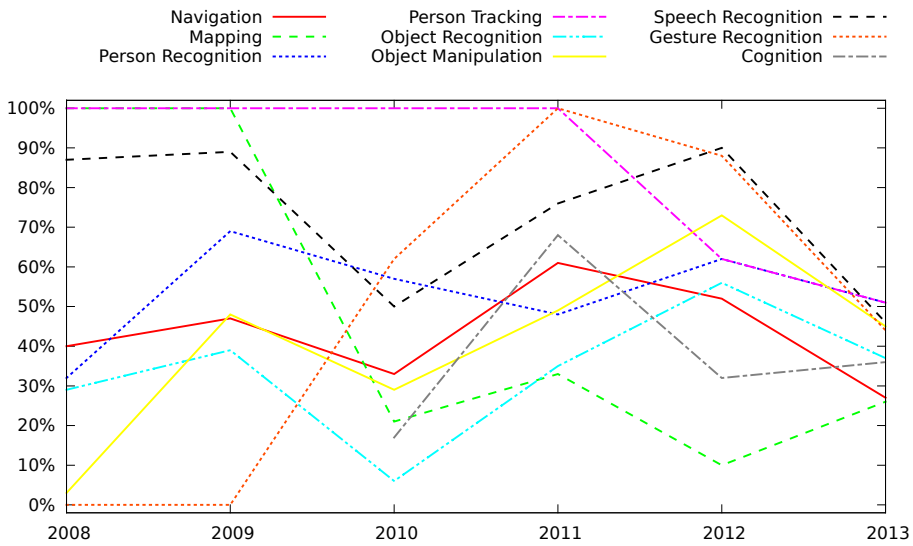

Fig. 3. Part of the statistics carried out after every competition. Looking at the scores achieved by the best teams (per capability) gives a good impression on what was solved well and there still problems exist.

competition. The results are important for the organizers and technical committees of a competition, since they show how the competition is progressing and possibly how to modify the rules in order to drive the development of the competitions. Some general rules that have been applied in RoboCup@Home are:

- if a functionality has high score (e.g., Mapping in 2009), then increase its difficulty in the next year;
- if a functionality has low score (e.g., Object Recognition in 2010), keep the difficulty unchanged;
- if a functionality is not developed at all (e.g., Gesture Recognition in 2008 and 2009), make it mandatory and increase the value and the phases in which it is used.

The main goal of this analysis and of the consequent measures is to keep a reasonable level of difficulty over the years and to balance the development on all the functionalities. Referring to Fig. 3, a functionality that has a high score for many years is an indicator that the problem addressed is too simple and that specialized solutions for the particular setting have been found. Moreover, in development of integrated research in AI and robotics, a good balance of all the functionalities is required. The low standard deviation obtained over the years in RoboCup@Home, demonstrate our efforts in developing systems that properly integrate many functionalities.

For more details on the conducted statistics and methods to derive changes, the interested reader is referred to [2]. In the following, we will give two particular

examples for how tests evolved over the years and how both complexity and overall performance increased.

4.1 Following a Human Guide

In addition to (direct) human-robot interaction and mobile manipulation, we consider following (and guiding) human operators an important capability of a domestic service robot. In RoboCup@Home, we began testing human detection, tracking and following from the beginning in the so-called "Follow me"-test.

2007: The first implementation provided a proof-of-concept in a simplified setup, e.g., in the first phase special markers on the guide have been allowed. Many teams succeeded (or in case of failure could at least show that the robot possesses the ability of detecting and following a human guide). However, some teams also managed to present a general solution that already worked for arbitrary operators (and did not require customization of the test setup). Following was considered an important aspect, but had to be increased in complexity.

2008: In order to stabilize in performance and form a common test bed, artificial markers were no longer allowed. Instead, the teams could use a previously known person (e.g., a team member) and calibrate on this very guide, e.g., by tracking shirt color etc. Furthermore, the test was made competitive in order to integrate a measure of time. In this first implementation, two teams were competing at the same time, and the team finishing the track faster received a bonus. In addition, a demo run outside the arena was conducted where guides and robots walked through the venue (not part of the competition).

2009: Instead of having guides known beforehand, independent referees were used, but the test stayed competitive (two teams in parallel). The same guide(s) have been used for all teams. At the beginning of the test, every robot got one minute for calibration on the guide right after being commanded to follow that person. To integrate new aspects like collision avoidance, the tracks of the robots were chosen to cross such that the robots had to avoid each other while still following their respective guides.

2010: The robustness of the approaches (not only for person following) in the arena considerably improved. In 2010, we introduced two tests that take place outside of the known RoboCup@Home arena and in (possibly) crowded public places—one of them being "Follow me". That is, the environment is 1) no longer controllable and 2) the environment is not known beforehand and, consequently, no map is available.

2011: After following (even in previously unknown environments) has matured and enough teams succeeded, in 2011, we introduced pre-defined interferences to increase the complexity of detecting, tracking and following even further. The track was split into waypoints. Whereas at one waypoint people passed in between guide and robot), at another the guide disappeared. When coming back to the robot, the guide was accompanied by another person effectively requiring that the robot recognized its guide and did not continue following the wrong person.

Since 2012: These interferences have been made more complex to incorporate a person directly passing between robot and guide and a situation in which the guide sneaks through a crowd so that the robot has to circumvent the crowd to continue following.In addition, we introduced an elevator. The elevator required that guide and robot leave in opposite order of entering. It also requires guide and robot to interact in order to coordinate.
In this setup, the test is conducted since 2012 since no team has so far been able to complete the track without failures.

Possible Extensions: The split into sub-tasks in this test allows for easily removing solved aspects of tracking and following while adding new more complex problems. More general extensions that are planned are 1.) endurance and 2.) guidance.

1.) Right now, the maximum time for this test is10 minutes and the overall track does not exceed 100 m. In order to foster the endurance aspect and to get closer to a real-world application, a possible future development is to extend the duration of the test and to let robots follow their guides through crowded public places for a longer period of time, e.g., 30 minutes.
2.) Another possible future development is guidance, e.g., for one part of the track the robot takes over the role of the guide and brings the operator to a certain position while keeping track of the operator (again, with interference as in following).

Another possible future test could include person search in public, crowded environments such as restaurants or supermarkets.

4.2 Finding and Manipulating Objects

Object search and manipulation were not limited to one recurring test, but evolved over a tree of tests over the years.

2007: A first implementation was "lost'n'found" in the 2007 rulebook where robots had to find an object somewhere in the apartment. Manipulation was not involved. In another test (called "manipulate"), an object should be grasped but the location of that object was known.

2008-2009: In order to obtain a more realistic setup, the procedure was changed in the "fetch'n'carry" test where teams could pick one object out of five (defined by the TC) together with five possible approximate locations (e.g. "on the couch table") of that object. The TC chose and placed the object. The robots were given commands for retrieving the object from the chosen location, but had to find the exact location and grasp the object themselves.

2010-2011: Since the approximate location of the objects was known, real search was not involved. Also, since there was only one object at the approximate location, solely grasping any object without recognition led to success because of the simplified setup. In 2010, these aspects were emphasized in the "go get it"-test by distributing four objects to grasp and four objects to ignore in the whole apartment (all from the set of known objects). Two robots were operating simultaneously and the first robot to find, grasp and deliver one of

the objects got a bonus. In the same year, we also introduced the "Shopping Mall" test in which objects had to be found and manipulated in previously unknown public areas.

Since 2012: Up to 2011, manipulation only considered grasping but not placing objects. In 2012, the "cleanup"-test was introduced in which robots had to find objects, recognize them and bring them back where they belong to (e.g., beverages on the kitchen table). We also introduced unknown objects that had to be brought to the trash bin.

Possible Extensions: Over the years, the number of objects that the robots needed to know and recognize was steadily increasing (to now 25 known objects). It is planned to further increase this number every year simply to increase recognition complexity. Also, objects are so far easy to manipulate (e.g., easily graspable). It is planned to introduce more complex objects, e.g., requiring for two-handed manipulation, being large or heavy, or even fragile. Another interesting aspect is dealing with unknown objects, e.g., categorizing them, retrieving missing information from the Internet etc. A possible future test may include object search in complex setups, such as the ones found in a kitchen or supermarket shelves or corridors.

5 Possible Future Directions

Just as the RoboCup soccer leagues, RoboCup@Home has a long-term goal: robots that find their way in the real world and cope with everyday problems in the real world. A possible application scenario are robots assisting people depending on help while they are traveling to the RoboCup competition. The robots would need to take over many responsibilities from planning the trip over packing everything needed to guiding and assisting their operators on the road, e.g., using public transport to the airport, finding check-in and gate, boarding the airplane and at the destination, find their way to the competition. Obviously, such systems are far from ready to be implemented, but there are many sub-problems on the way that can already be addressed, e.g.,

- **Interaction:** intuitively interacting with people (not used to robots), e.g., asking for the way and taking care of necessary communications (check-in etc.) possibly in different languages.
- **Finding the way:** obtaining information on the Internet, reading signs, maps and pictograms or interpreting gestures or hand-drawn maps.
- **Long-term operation:** being able to run for longer periods of time.

Some problems on the long run towards *real* real-world applicability are already integrated into the RoboCup@Home competition or planned for the next years.

Endurance and long-term operation: In contrast to soccer games that have a defined duration, real-world service robot tasks are open-ended. Moreover, tasks being coped with may take considerably longer than what is currently doable with a single charge of the batteries. That said, to enforce endurance,

teams need to work on both hardware and software side for making robots longer (reliably) operable and at the same time address the problems arising in long-term operation (e.g., knowledge management). Already implemented in RoboCup@Home is the "Enhanced General Purpose Service Robot"-test in which robots need to operate for 30 minutes (in contrast to the usual test duration of 10 minutes) and solve tasks on demand. It is planned, to have more and more parts of the competition running for a longer period of time possibly with several robots being tested simultaneously. We believe that this would also make the competition more interesting for spectators as there is always something happening.

Real-world application: In contrast to the RoboCup@Home arena in which most tests take place, the real-world is unpredictable, may be very crowded, and may require the robot to interact with people not used to see, hear, or operate robots. Right now, the league features two tests that take place in public areas. However, these areas are still modified by controlling direct access of the audience or simplifying task and environment, e.g., avoid arising problems that—from experience of the Technical Committee—are unsolvable right now. However, it is planned to 1) have more and more tests taking place in the real world and 2) to perform fewer customizations and simplifications.

Semantic Perception and Mapping: Up to now, semantic perception and mapping capabilities are not explicitly being tested in RoboCup@Home. It is planned to foster the development of such capabilities by introducing new tests that also aim for cognitive skills in general. A possible test could be searching for a place in unknown environments like looking for the kitchen or the toilette in a house or restaurant.

Intuitive (multimodal) interaction: When the league started, the only allowed way of interacting with the robot was by natural speech (i.e., spoken commands). While this was the most obvious and natural way of interacting, it is not always the most convenient or successful way. For example, in places where it is really loud like, for example, the RoboCup venue during an interesting soccer match, giving speech commands is condemned to failure. Instead, we are aiming at multimodal intuitive interaction by having, for example, a combination of speech and gestures, buttons/displays on the robot for direct cooperation and intuitive touch pad interfaces allowing to remotely command and operate the robots.

Shortcuts in test implementations: Many tasks in RoboCup@Home build upon one another or depend on certain user inputs or other events. For the first time in 2014, we integrate more and more shortcuts and workarounds that can be used to continue a test in case of failures, e.g., command being misunderstood or not understood at all, referees and operators acting wrong or giving wrongs commands etc. If successful, we are going to foster these workarounds as they 1) allow for evaluating components otherwise inaccessible in the test and 2) make the competition more attractive for the audience.

In addition, we are going to consider the following aspects to be incorporated in the upcoming rulebooks:

- Fostering the benchmarking character: in order to better assess the performances of teams and establish RoboCup@Home as a widely accepted benchmark for domestic service robots, we aim at improving its benchmarking character, e.g., by having
  - tests that are easy to set up and reproduce,
  - outcomes that are easy to evaluate and compare, and
  - more tests per capability and team for better statistics.

  This will, for sure, be accompanied by the introduction of new tools, e.g., for system monitoring or automated task assignment and evaluation.
- Semantic interpretation and categorization of complex, unknown environments, such as houses never seen before.
- Improving the cognitive and social skills of robots by integrating adequate tests. This may include the language skills, but also social behaviors.
- Improving safety and security aspects, especially for interaction and cooperation with non-experts.
- Human-robot cooperation and inter-team robot-robot cooperation
- Development and provision of a standard platform for RoboCup@Home to be more attractive for teams not willing to develop and maintain their own hardware.

6 Conclusion

In this paper we have presented the RoboCup@Home league—a league for domestic service robots—which addresses (in the long run) everyday problems in the real world. In particular, the competition is implemented as an annual benchmark for different robot capabilities ranging from intuitive human-robot-interaction to mobile manipulation. Part of this benchmark is evaluating and keeping track of the teams' performances over the years and to use this information to drive the league's development. We could show that the implemented changes in the last years considerably increased complexity while keeping the overall performance nearly stable. That is, the participating teams not only robustified existing capabilities but also improved on them and implemented new ones. We detailed some of the capabilities within the competition and gave an outlook on future developments that—in the long run—will hopefully allow domestic service robots to find their way in the real world and cope with everyday problems in the real world.

References

1. Tijn van der Zant and Thomas Wisspeintner. RoboCup X: A proposal for a new league where RoboCup goes real world. In *Proceedings of the RoboCup International Symposium*, pages 166–172, 2005.
2. Dirk Holz, Luca Iocchi, and Tijn van der Zant. Benchmarking intelligent service robots through scientific competitions: the RoboCup@Home approach. In *Proceedings of the AAAI Spring Symposium Designing Intelligent Robots: Reintegrating AI II*, 2013.