Artificial Intelligence 229 (2015) 258–281

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots ☆

Luca Iocchi [a,*], Dirk Holz [b], Javier Ruiz-del-Solar [c], Komei Sugiura [d], Tijn van der Zant [e]

[a] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Italy
[b] Autonomous Intelligent Systems Group, University of Bonn, Germany
[c] Department of Electrical Engineering & Advanced Mining Technology Center, Universidad de Chile, Chile
[d] Information Services Platform Laboratory, National Institute of Information and Communications Technology, Japan
[e] Cognitive Robotics Laboratory, University of Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form 15 July 2015
Accepted 8 August 2015
Available online 11 August 2015

Keywords:
Robotic competitions
Artificial intelligence and robotics
Benchmarking

ABSTRACT

Scientific competitions are becoming more common in many research areas of artificial intelligence and robotics, since they provide a shared testbed for comparing different solutions and enable the exchange of research results. Moreover, they are interesting for general audiences and industries. Currently, many major research areas in artificial intelligence and robotics are organizing multiple-year competitions that are typically associated with scientific conferences.

One important aspect of such competitions is that they are organized for many years. This introduces a temporal evolution that is interesting to analyze. However, the problem of evaluating a competition over many years remains unaddressed. We believe that this issue is critical to properly fuel changes over the years and measure the results of these decisions. Therefore, this article focuses on the analysis and the results of evolving competitions.

In this article, we present the RoboCup@Home competition, which is the largest worldwide competition for domestic service robots, and evaluate its progress over the past seven years. We show how the definition of a proper scoring system allows for desired functionalities to be related to tasks and how the resulting analysis fuels subsequent changes to achieve general and robust solutions implemented by the teams. Our results show not only the steadily increasing complexity of the tasks that RoboCup@Home robots can solve but also the increased performance for all of the functionalities addressed in the competition.

We believe that the methodology used in RoboCup@Home for evaluating competition advances and for stimulating changes can be applied and extended to other robotic competitions as well as to multi-year research projects involving Artificial Intelligence and Robotics.

☆ This paper was submitted to the Competition Section of the journal.

* Corresponding author.
E-mail addresses: iocchi@dis.uniroma1.it (L. Iocchi), holz@ais.uni-bonn.de (D. Holz), jruizd@ing.uchile.cl (J. Ruiz-del-Solar), komei.sugiura@nict.go.jp (K. Sugiura), tijn@ai.rug.nl (T. van der Zant).

http://dx.doi.org/10.1016/j.artint.2015.08.002
0004-3702/

1. Introduction

Artificial intelligence and robotics competitions have significantly increased their scope and visibility in recent years and today most research areas have some type of competition. Such competitions provide frameworks where research groups can compare the results of developed methods and give opportunities to define standard benchmarks for solving specific problems, comparing different solutions, and disseminating the best solutions available to the community. Moreover, tasks and results are often used in scientific papers to compare new approaches with existing ones.

Competitions go beyond typical experiments. Analogies and differences between robotic experiments and competitions have been investigated [1,2]. Experiments (and benchmarking) in robotics are characterized by the following: i) a common testbed (usually provided through detailed specifications), ii) specific performance metrics, iii) reproducibility, and iv) repeatability. On the other hand, scientific robotic competitions involve: i) specifications of competition environment, ii) specification of robot requirements and constraints, iii) specific performance metrics to rank the participants, iv) information about how they are organized.

The main difference is of course found in the scope of these two activities; the former aims at demonstrating and measuring the performance of a system or component to solve a particular problem, while the latter aims at directly comparing different solutions in a predefined testbed. A common element is the environment specifications that can be more or less detailed in both cases, depending on the design of the experiment/competition. For example, RoboCup@Home gives no specifications about the shape and the size of the apartment, the kind of furniture, and the objects used in the tests.

In competitions, since performance metrics are defined (in terms of a scoring system) to measure a system's overall ability, they are generally not intended to measure specific functionalities or internal features. This is not in contrast with robotics experiments, but the latter are more often used to measure the performance of a system's components to solve a specific problem rather than the entire system.

The following are some important differences between experiments and competitions: 1. in competitions, perfect reproducibility is not possible, since recreating the same scenario (including the same level of background noise) is typically not possible, and 2. repeatability is usually not considered within the competition (each test is normally conducted only once) for organization reasons.

Competitions have thus both similarities and differences with respect to robotic experiments, and the definition of a proper relation between them is an on-going effort. For example, a proposal has been presented for improving the scientific aspects of experiments within robotic competitions and challenges [1]. The proposed framework is based on a modular integrated system and the interoperability of components, where participants can focus on particular modules. Even though this idea has been partially implemented in some competitions, it has not been fully exploited yet. The RoCKIn project also aims for better integration between competitions and scientific benchmarking [3] by collecting benchmarking data during tests that combine the logged internal states of robots and the ground truth acquired through an external system. RoCKIn competitions will thus be an important contribution to this research area.

Another important aspect of a competition is its evolution. Many existing competitions are held periodically (e.g., annually), changing tests over time to address more difficult problems, to enlarge the variety of problems, to provide different experimental conditions, etc. When a competition is run for many years, it is important to evaluate how its design and organization affect the solutions implemented by the participating research groups. More specifically, the choices in defining the tasks to be addressed during a competition and the score that determines the rank of the participants impact how the participants develop their solutions to the specified problems. A good test and scoring system design allows for suitable development of corresponding solutions from participants.

In this article, we are mainly interested in analyzing the results of competitions over the years. In particular, we focus on how this analysis has been carried out in the last years within the RoboCup@Home competition, which is the largest competition for domestic and service robots. The evaluation of a competition's effectiveness based on the scores of the teams over the years is important for better fueling the competition and the novel approach introduced in RoboCup@Home for addressing this issue is the main contribution of this article. Although our analysis reported in this article is specific to the RoboCup@Home competition, we believe that its principles can be applied both to other existing competitions and to new ones, and it could even be used for evaluating multi-year projects or challenges in AI and robotics where some tasks are repeated and changed over the years.

RoboCup@Home is a competition where domestic and service robots perform several tasks in a home environment, interacting with people and with the environment in a natural way. Natural interaction means that a robot is expected to interact with the environment and with other people, as any person would do. So natural forms of human–robot interaction include speech and gestures, but not joysticks or keyboards.

During the competition, the teams are required to perform several tests. Since their total score is the sum of the scores obtained in each test, teams are motivated to perform well in every test in order to gain a high rank. Each test requires a combination of different functionalities (including 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 its main characteristic is that it changes tests every year while maintaining the same basic functionalities. By changing the difficulty and the combinations of the functionalities to be integrated, we aim

at encouraging the teams to develop general and robust solutions. Indeed, with this competition setting, it is difficult to implement specific ad-hoc systems to individually solve each task.

After the competition's first two years, it became clear that analyzing its results (i.e., how the teams performed) was very important in order to encourage development. But, the difficulty of this task was also evident. In the first two years (2006–2007), the task scores were Boolean: when the task was successfully performed the team received a score, otherwise the score was zero. Such a scoring system made it impossible to relate the performance of the teams to the development of functionalities. Moreover, in some cases, it was also impossible to properly reward partial accomplishments.

Because of that, the scoring system was changed in 2008 as described later in this article, in order to better reflect the functionalities underlying the tests. With this new scoring system, the contribution of a specific functionality can be estimated in the total score of the teams to evaluate their general performances during the competition.

Our analysis, which was conducted during 2008–2014, played a significant role in driving the competition toward desired directions. Since we believe that an intelligent domestic and service robot should optimally balance its capabilities, a good solution must demonstrate effective integration of many functionalities rather than optimal performance in some functionalities and low performance in others. In the first few years of the competition, it was clear that some teams had very good performance in particular functionalities and very bad performance in others. This unbalance created an inadequate system for actual deployment. On the other hand, a balanced system, which has adequately integrated all the required functionalities, has many advantages, both in terms of competition results and for public demonstrations and dissemination to general audiences and industries. Moreover, integrating several AI techniques into a robotic system has been a major goal of AI since the very beginning and is regaining interest.[1]

As described in the following sections, the method used in RoboCup@Home gradually reduced the standard deviation of the scores achieved by the teams over all the functionalities. This means that the teams have been working on integrating many functionalities to make a complete working system, rather than optimizing performance in a single functionality. This behavior has been guided by adopting rules based on the analysis presented here.

Therefore, RoboCup@Home supports the development of complex robotic systems that can: 1. suitably operate in a domestic environment and interact with people, 2. integrate many functionalities, and 3. actually execute many different tasks. These achievements are critical not only for RoboCup@Home competitions but also for research that integrates AI and robotics.

The remainder of this article is organized as follows. In Section 2, we describe several multi-year competitions relevant for AI and robotics and highlight their main characteristics as well as similarities and differences with RoboCup@Home. In Section 3, we briefly describe the overall goals and the organization of the RoboCup@Home competition and give greater details of its scoring system and the evolution of its tests. In Section 4, we present analysis of the competitions and discuss the corresponding results. In Section 5, we describe some of the best solutions adopted by RoboCup@Home teams in the desired functionalities that have interest for other AI systems. In Section 6, we address the lessons we have learned and the future plans of the competition. Finally, in Section 7, we draw some conclusions and discuss future directions.

2. Related work

The main goal of this article is to discuss how to evaluate the competition results over the life of the competition. In this section we present a set of competitions relevant to AI and robotics that have been held for multiple years. Other challenges are not listed are, for instance, the DARPA Challenges,[2] because they have only been organized a few times, making it impossible to analyze their evolution. As discussed later in this section, although we examined many multi-year competitions, to the best of our knowledge, no detailed analyses exist about their evolution.

We divide competitions in two groups. First we describe *software competitions* that are not run in a physical environment, and then we describe *robotic competitions* that involve actual robots acting in a physical environment.

The main difference between these two kinds of competitions is that in the first group, everything can be easily controlled and the tests and the entire competition can be easily replicated off-line at a different time and location with respect to the actual competition. In this case it is also always possible to directly compare the results of different systems. On the other hand, when competitions are based on real interaction in the physical world, replicating the results off-line and directly comparing different systems is no longer possible, since exact replication of the physical environment and the phenomena occurring during the competition is impossible and/or very expensive in terms of space, time, and economic cost.

2.1. Software competitions

Several competitions involving AI techniques implemented as software modules are illustrated in this section. They relate to the main components of a complex AI system.

[1] See for example the *AAAI Symposia on Designing Intelligent Robots: Reintegrating AI*—people.csail.mit.edu/gdk/dir2 and the "AI and Robotics" initiative: ai-robotics.wikispaces.com/.

[2] www.theroboticschallenge.org.

Planning competitions International Planning Competitions[3] have been held since 1998. They are related to solving planning problems, where planning domains are represented in a standard language (PDDL) and the implementations of algorithms for finding planning solutions are compared. Typical performance metrics are the computational time, the number of solved problems, and the quality of the solutions. Problems are changed every year; in the last few years, several tracks have been added for dealing with uncertainty, learning, and continuous time.

Another interesting multi-year competition is the Answer Set Programming Competition, which has been held biannually since 2007. It focuses on evaluating declarative knowledge representation systems that are used to solve complex AI problems. This competition also provides benchmarks that use a standard language (ASP-Core-2) and compares systems in terms of the number of solved problems and the time to compute solutions.

Since the results of these competitions can be easily reproduced the domains and the datasets used in the competitions are also used in scientific papers to evaluate new approaches and compare them with respect to existing ones. The best solutions are also typically made available to the research community.

Competitions related to vision using standard data sets There are several competitions whose goal is to evaluate the performance of image or video processing systems, by defining a challenge to be solved (e.g., video surveillance) and an associated dataset that normally includes evaluation procedures. Typically, the following are the key elements: (i) standardized databases, (ii) a common set of tools for accessing and managing the database annotations, (iii) a challenge for evaluating the performance of methods using the defined databases, and (iv) a scientific event (e.g., a workshop) at which the best solutions are presented. Well-known competitions include PETS[4] and PASCAL VOC.[5] Performance Evaluation of Tracking and Surveillance (PETS), which has been organized annually since 2000, addresses the performance evaluation of visual tracking and surveillance algorithms. The PETS organization defines benchmark data consisting of video datasets for video surveillance related topics every year and organizes a workshop where researchers present papers that describe the best solutions to the current problem. The PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning) VOC (Visual Object Class) challenge is "a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures" [4]. This challenge has been held annually since 2005. ImageNet Challenge,[6] which has been held since 2010, focuses on object detection and image classification like PASCAL VOC, but on a much larger scale, e.g., up to 1000 object categories.

In PETS, PASCAL VOC, and ImageNet, the evaluation of different algorithms is based on performance metrics, which depends on the exact challenge being solved (e.g., accuracy of detection and classification). These metrics usually consider the difference between a ground truth (typically manually labeled images and data) and the output of the system under test. The final score of a given algorithm depends on the computed metrics over some evaluation data sets. For instance, in PETS, the following metrics are used: Negative Rate, Misclassification Penalty, Rate of Misclassifications, and Weighted Quality Measure [5].

The Robot Vision Challenge,[7] which started in 2009, deserves a special mention as a computer vision competition. It focuses on semantic place classification using visual and depth information. Recently, object recognition tasks have also been included. This competition evaluates different solutions of image understanding applied to images collected by a mobile robot equipped with on-board sensors.

In this class of competitions, the results are easy to reproduce and the datasets produced for the competitions are used in scientific papers.

Competitions related to speech processing and dialogue systems In the speech processing community, shared tasks, corpora, and metrics have been the driving forces for the improvement of the accuracy and the quality of speech processing systems. One of the first competitions on Automatic Speech Recognition (ASR) was a series of benchmark tests supported by DARPA and the National Institute of Standards and Technology (NIST) [6]; however, they are no longer active. The IWSLT ASR track [7] started in 2004 and remains one of the most vibrant competitions in the ASR field. IWSLT is a yearly scientific workshop, associated with an open evaluation campaign mainly on ASR and spoken language translation (SLT). At IWSLT 2013, the task was the transcription of TED talks.[8] The participants were requested to submit the system's transcription results. The metric was the word error rate (WER), which can be calculated automatically. In 2013, the WER was in the range of 13.5% to 27.2% for English and 25.2% to 37.8% for German. In recent competitions, Deep Neural Networks (DNN) [8] generally achieved the best performance. One unique aspect of IWSLT is its "regression tests" in which systems are evaluated by the test sets of previous years. This prevents the possibility of having systems that overfit a particular task.

The Loebner Prize,[9] which is an instantiation of the Turing Test, can be categorized into a competition on (text-based) dialogue systems. It started in 1990 and will reward with a grand prize of $100,000 any system that can pass the Turing

[3] ipc.icaps-conference.org.
[4] pets2013.net.
[5] pascallin.ecs.soton.ac.uk/challenges/VOC.
[6] www.image-net.org.
[7] imageclef.org/2014/robot.
[8] www.ted.com.
[9] www.loebner.net/Prizef/loebner-prize.html.

Test. Unfortunately, no system has passed it yet. Each year a bronze medal and an annual prize ($4,000 in 2013) are given to the system that best exhibit human-like behavior. The REAL Challenge[10] is an evaluation of *spoken* dialogue systems. Its task is to provide bus schedule information for the general population of Pittsburgh by telephone interfaces. For example, if the caller says "I'd like to go to Forbes and Murray," then a desirable answer is: "To Forbes and Murray. Where are you leaving from?" This challenge's important aspect is that participating systems are deployed to interact with actual users. The WER and task completion rates of each system have been evaluated [9].

The Generating Instructions in Virtual Environments (GIVE) Challenge evaluates Natural Language Generation (NLG) systems that guide human users to solve tasks in virtual environments [10]. The first GIVE challenge, GIVE-1, started in 2008. In GIVE-1, each volunteer user played a treasure hunting task in a 3D virtual world. An NLG system was randomly selected from the participating systems and guided the user by generating instructions such as "Starting from left to right and from top to bottom, press the 3rd button from the row 2." The system was evaluated based on two aspects: objective (task success, duration, etc.), and subjective (questionnaires on understandability, timing, etc.). GIVE-2 took place in 2010, and GIVE-2.5 took place in 2011–2012; however, they are no longer active.

The largest academic competition in speech synthesis is the Blizzard Challenge [11], which started in 2005. Eleven teams participated in the Blizzard Challenge 2012. They were given an audiobook and annotations as a shared training set, and requested to submit synthesized voices as test sentences. The evaluation was conducted by hundreds of paid and volunteer listeners by web-browser interfaces. They gave subjective opinion scores on such aspects as overall impressions, pleasantness, and listening effort. In recent competitions, a hybrid approach of the unit selection and HMM-based methods achieved the best performance.

Competitions related to agents and multi-agent systems The agent and multi-agent competitions described in this section refer to competitions where an agent is either a pure software agent or a high-level representation of a physical agent. Competitions for robots and simulated robots are addressed in the next sections. These competitions can be divided into two categories: i) competitions for pure software agents, ii) and competitions for agents that simulate physical agents or systems (using a multi-agent simulator).

The first category includes the International Automated Negotiating Agents Competition [12], which started in 2010. In this competition, agents act in a bilateral negotiation scenario without any knowledge about the preferences and strategies of their opponents. The performance metrics are the utility scores achieved by each agent during all the negotiations.

RoboCup soccer 2D[11] and RoboCup Rescue Agents Simulation[12] are in the second category. The former uses a 2D simulator where each soccer agent is modeled by abstract perception and the action capabilities affected by simulated perception noise and non-deterministic actions. The latter also uses a 2D simulator where the perception and action abilities of the police, ambulance, and fire-fighter agents are again modeled at a very high level.

In these competitions, some random noise is introduced during any execution of a test. Therefore, although results can be reproduced off-line with respect to the competition, it is not guaranteed to achieve exactly the same results as an effect of non-deterministic execution. However, since we can execute many runs, it is easy to average performances over a set of runs. For example, in RoboCup soccer 2D, a new team can be compared with the best available one by running hundreds of soccer matches in a short time (e.g., by accelerated simulations) and evaluating the results.

Competitions using robotic simulators Many competitions have been implemented using robotic simulators. The level of fidelity with which the robots and the environments are modeled depends on the application scenarios. Competitions like RoboCup soccer 3D[13] and RoboCup Rescue Virtual Robots[14] are designed based on the corresponding real robot competitions, from which they inherit rules and performance metrics.

These competitions are usually used to fine tune methods and systems that are then used in the real robots. Indeed, many teams participate in both virtual and real competitions with similar implementations. Moreover, virtual environments allow for scaling to larger environments and more robots (e.g., 11 vs. 11 in soccer and large rescue scenarios), which is not possible or too expensive for real competitions.

As in the agent-related competitions described above, a single run is affected by randomness and non-determinism and executing many tests is necessary for statistically significant results. Unlike previous competitions, since running robotic simulators is more expensive in terms of computational resources and time, running hundreds of tests in a short time may not be feasible.

2.2. Robotic competitions

In contrast to software competitions, robotic competitions are run in a real physical environment with real robots and require a physical space in which the competition takes place. Robotic competitions involve much logistical effort and

[10] dialrc.org/realchallenge.
[11] wiki.robocup.org/wiki/Soccer_Simulation_League.
[12] www.robocuprescue.org/agentsim.html.
[13] wiki.robocup.org/wiki/Soccer_Simulation_League.
[14] www.robocuprescue.org/virtualsim.html.

cost for both the organizers and the participants. It is also important to observe that the competition environment is usually not perfectly specified and the participants can access it only a few days before the actual competition. Moreover, it is very difficult (or even impossible) to reproduce the setup after the competitions, and the runs cannot be exactly replicated afterwards either. Thus, evaluations in robotic competitions are biased by the impossibility of reproducing and replicating the results, although some efforts to improve the benchmarking aspects in the competitions are in progress (see also Section 3.5).

AAAI Mobile Robot competitions The AAAI Mobile Robot Competitions, which are the oldest competitions that address the integration of AI and robotics, started in 1992 and are typically hosted at the main artificial intelligence conferences: AAAI and IJCAI. The competitions are not focused on a specific topic; they include two or three tasks that have significantly changed application areas over the years: navigation and exploration in an unknown area, collecting objects in arenas, rescue robots, human–robot interaction, navigation in dynamic environments, etc. [13]. More recently, learning from demonstrations and robot chess challenges have been performed as well as many open demonstrations and exhibitions. Being co-located with the main AI conferences, these competitions guarantee great visibility within the scientific community and most of the scientific and technological achievements demonstrated during them have also been presented within conference events. On the other hand, the significant change of the aim and the focus of the tasks over the years renders analysis of the competition evolution impossible.

RoboCup Soccer and Rescue RoboCup Soccer, which began in 1996 and has held with official games since 1997, is probably the most famous robot competition. Its main focus is "the game of football/soccer, where the research goals concern co-operative multi-robot and multi-agent systems in dynamic adversarial environments".[15] RoboCup Soccer is organized into different leagues (humanoid, small size, middle size, and standard platform), and each league defines its own regulations, which include such aspects as field characteristics, game rules, and robot characteristics. Regional and worldwide competitions are held every year. Technical committees, whose members are elected from among the team members, analyze team performances and annually define rule changes to make the competitions more attractive and more difficult. In this way the leagues evolve over time, with more difficult competitions from year to year. The scoring system of RoboCup Soccer is goal-oriented, since it is based only on the results of matches, which are obviously determined simply by goals scored. Technical challenges are also organized to test specific functionalities, but these challenges determine a different rank.

RoboCup Rescue[16] started in 2002. It focuses on the development of robots, information infrastructure and decision support systems that can assist disaster rescue operations. As in the case of RoboCup soccer, technical committees analyze the team performances and improve the rules over the years. In RoboCup Rescue, the scoring system is also goal-oriented and is determined by the number of human victims found in the environment. However, teams also get points for the quality of their mapping, for the level of autonomy of the mission, and for delivered payload. 'Best in class' prizes for performances in specific functionalities (e.g., mobility, autonomy, etc.) are also awarded.

Micro Air Vehicle competitions International Micro Air Vehicle Conference and Competitions,[17] which have been held since 2009, focus on Micro Air Vehicle (MAV) and multi-MAV missions in indoor and outdoor environments. The tasks are related to such applications as surveillance and recognition and include functionalities like mapping the environment from aerial views and recognizing particular elements (people, buildings, etc.). Teams can address either the complete mission or sub-tasks of it, and the autonomy of MAVs can be chosen by the teams and affect the score. Evaluation is based on the autonomy of the MAVs, the elements found during the mission, and MAV size. These competitions are crucial for testing and assessing both the hardware devices and the control software. Due to limited weight, power, computation, and communication payload, software implementation addresses only simple control algorithms.

RoCKIn and euRathlon competitions The RoCKIn[18] and euRathlon[19] competitions are very recent efforts to organize robotic competitions in different scenarios. RoCKIn addresses indoor scenarios (@Home and @Work, similar to the corresponding RoboCup initiatives), while euRathlon focuses on outdoor scenarios for emergency-response applications. These competitions have an important role because, by building on the experience of previous ones, they encourage innovative elements. However, the limited duration of the corresponding projects does not allow for extensive analysis of evolution. Although the main competitions for these projects are being planned for late 2015, interesting outcomes already exist that are related to the analysis presented in this article. RoCKIn@Home is the most similar competition to RoboCup@Home, and its significant focus on benchmarking, that includes a different evaluation methodology and a deeper analysis of the interaction between functionalities and tasks, is certainly of great interest to RoboCup@Home. As discussed further in Section 6.2,

[15] www.robocup.org/robocup-soccer.
[16] www.robocuprescue.org.
[17] www.imavs.org.
[18] www.rockinrobotchallenge.eu.
[19] www.eurathlon.eu.

the ideas of multiple repetitions of a test for better statistical analysis and a new scoring system for a detailed analysis of the relation between functionality and task performance come directly from observations of the RoCKIn@Home competition.

2.3. Discussion

Based on the above analysis, we can summarize the features of the two categories. Software competitions have the following features:

1. possibility of reproducing the results many times and off-line with respect to the actual competition, thus providing opportunities to compare new methods with the state-of-the-art using benchmarks from the competitions;
2. organization of the competitions is simplified by the fact that they usually require only computational resources;
3. teams can participate by just sending their own implementations.

Software competitions are aimed at solving specific problems or small parts of a general complex problem, and although they are instrumental in the development of research, sometimes the best results developed in these contexts are not ready for real-world applications. Moreover, software and simulated competitions are also very effective for the research groups involved, but since they are typically less attractive to general audiences, they are not suitable for disseminating AI results and technologies.

Robotic competitions enjoy the following features:

1. possibility of testing and comparing physical systems in real environments;
2. development and testing of integrated multi-disciplinary research, where many functionalities must be properly combined;
3. possibility of disseminating AI and robotics research by making it attractive for general audiences and industries;
4. high organization cost and significant participation effort;
5. difficulty (or impossibility) of reproducing the competition scenarios.

With respect to other robotic competitions, RoboCup@Home addresses different scenarios with different kinds of robotic platforms and presents the following specific features:

1. only fully autonomous systems are tested;
2. human–robot interaction is a primary capability, and natural interaction (e.g., speech) is required;
3. cognition (i.e., robot's reasoning ability) is explicitly required and measured as part of a specific test;
4. there is a large variety of tests, discouraging the development of ad-hoc solutions.

On the other hand, some functionalities are shared with other robotic competitions. The modules tested in software competitions can also be used in RoboCup@Home. For example, RoboCup soccer robots have to be fully autonomous and recognize objects under natural lighting conditions. Object manipulation and the perception of emergency situations are common functionalities for rescue and domestic robots. Moreover, modules developed within planning, vision, and speech understanding competitions could be integrated into @Home robots.

With respect to scoring systems, RoboCup@Home's system is more sophisticated, as described in this article, because it integrates two modes: a goal-oriented one, which measures the ability of the robots to fulfill the tasks specified in the rules and is used to actually rank the participant teams and award prizes; a functionality-oriented one, which measures how the teams perform in each of the basic functionalities defined for the competition. This measure does not affect the rank of the teams, but it is used to analyze and drive changes over the years, as explained in this article. Since the two modes of the scoring system are applied simultaneously on the same tests, participants are not required to perform separate tests for measuring the goal- and functionality-oriented scores. The functionality-oriented scoring system allows for analysis of the competition's evolution. It is critical to assess the effectiveness of the overall competition over the years instead of individual participating systems.

As already mentioned, although other competitions that have been run for multiple years have similar evolving characteristics as RoboCup@Home, no detailed analysis of the evolution of their performances exists. More specifically, to the best of our knowledge, no reports have analyzed the evolution of competitions in detail (with a clear methodological approach) or described how to drive this evolution by measuring task complexity. On the other hand, we consider this as an important open issue that deserves future investigation to improve the scientific quality of competitions. Indeed, the analysis of a competition's evolution can bring many benefits and help driving its overall objectives. Consequently, we believe that the RoboCup@Home approach can make important contributions to it.

The evaluation of a competition's effectiveness based on the team scores is thus a novel approach introduced in RoboCup@Home and is the main contribution of this article. Moreover, the evaluation methodology applied in RoboCup@Home might help both existing competitions and new ones as well as multi-year projects or challenges in AI and robotics where some tasks are repeated and changed over the years.

3. The RoboCup@Home competition

In this section, we first describe the overall goals of the RoboCup@Home competition and its limits in the considered scenarios and then provide a brief overview of its organization. We describe in detail its scoring system that is very relevant for the analysis presented in the next section. We also discuss how the tests have been changed since the competition's inception and finally address benchmarking activity in RoboCup@Home.

A more detailed description of the competition is available [14], and additional information about it (including photos and videos of the execution of many tests over the years) are in the RoboCup@Home website.[20] In particular, the specifications of the tests and the changes made during the years are fully described in the rulebooks.[21]

3.1. Overall goal of competition

RoboCup@Home's overall goal is to provide a framework, in the form of a scientific competition, for testing and comparing solutions for the development of service and assistive robot technology with high relevance for future personal domestic applications. We aim at demonstrating robots that are able to understand and satisfy typical user needs in a domestic scenario. A set of benchmark tests is designed and used to evaluate robot abilities and performances in a realistic non-standardized home environment.

The RoboCup@Home competition has no constraints on the type of robots used, except for general rules about size and dimension that require the robot to move in a typical home environment. However, some difficulties in the environment have not been considered so far: stairs or uneven floor, glass walls or other objects that are difficult to perceive, objects for which an interaction is required (e.g., to be manipulated) placed in difficult-to-reach locations in the environment, etc. Moreover, in most of the tests the number of people in the environment is limited and their actions and movements are quite predictable, since they are included in the test's description. Some exceptions to these conventions have been encountered for tests performed outside the arena. At RoboCup 2012, the 'Restaurant' test was carried out in a facility having glass walls. All the participating teams successfully coped with them. During the 'Follow me' tests at RoboCup 2012 and 2013, a relatively large audience viewed the tests, creating a barrier for the sensor readings and introducing unexpected background noise (both for audio and video). No major problems occurred in these cases, either.

As a consequence of the environment setup and the above conventions, the form of all the robots participating in the competition is similar (Fig. 1): a wheeled mobile base, a torso with one or two arms and an upper part that resembles a head, including sensors and possibly the representation of a face. In other words, these robots tend to have an anthropomorphic upper body mounted on a wheeled mobile base. Therefore, legged humanoid robots[22] have been used just in a few tests, while no flying robots have been used so far. In general, currently, the RoboCup@Home competition does not favor legged or flying robots, and this direction is unlikely to change in the next few years.

In the design of the tests and in the analysis of the results, we consider the above overall goal of RoboCup@Home and the conventions limiting the difficulties of the environment. Thus, advances of RoboCup@Home are measured in terms of the effectiveness of the robotic systems (tested during the competition) in performing tasks and services that are relevant for helping non-expert users in daily activities in a domestic scenario, taking into account the discussed limitations of the difficulty of the environment.

3.2. Organization of competition

RoboCup@Home competitions include an annual international event within RoboCup and several regional events in Europe, Japan, USA, etc. Since 2006, one international and several regional competitions have been performed every year.

The competition runs in a realistic setting where an apartment with different functional rooms and typical furniture and objects is realized (Fig. 1). The environment is not completely specified in the rulebook, and the local organizers can decide its implementation. Thus the teams lack such specific information about this environment as the size, floor material, wall colors, kinds of furniture, objects, etc. before arriving at the competition venue. This strategy ensures the development of general solutions. Some tests are performed outside this area: either in the competition venue or in such public spaces as a restaurant or a shopping mall that are not known by the teams beforehand.

The competition is comprised of about ten tests, a Technical Challenge and the finals. A stage system is used. Stage I is performed by all the participant teams, usually half of them advance to Stage II, and five reach the Finals.

The following are the main functionalities required in the tests:

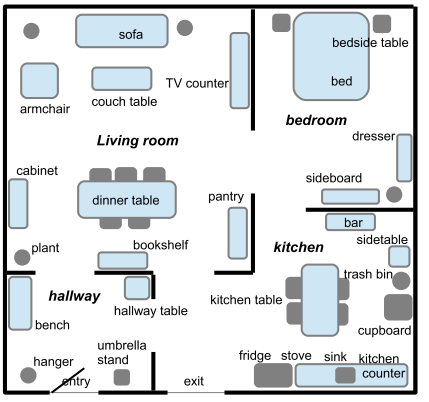
- *Navigation* is the ability to move safely in the environment and navigate to specific target positions, while avoiding static and dynamic obstacles.

[20] www.robocupathome.org.

[21] www.robocupathome.org/rules.

[22] In particular, the NAO robots from Aldebaran.

(a) Teams and robots



(b) @Home arena setup and appearance

(c) Set of known objects

Fig. 1. RoboCup@Home 2013: robots, people, arena and set of known objects.

- *Mapping* is the ability to autonomously build a representation of a partially known or unknown environment on-line or off-line.
- *Person Recognition* is the ability to detect and recognize people in the environment.
- *Person Tracking* is the ability to track a person's position over time.
- *Object Recognition* is the ability to detect and recognize (known or unknown) objects in the environment.

(a) 2007: Custom walker (b) 2011: Unknown walker, unknown environment, interferences (c) 2012: Crowded environment

Fig. 2. Evolution of 'Follow me' test. Starting with known guides in a known environment, the test is now conducted with unknown guides in an unknown environment with several interferences along the route.

- *Object Manipulation* is the ability to grasp, move, or place an object.
- *Speech Recognition* is the ability to recognize and interpret user spoken commands with no preliminary training about a particular user.
- *Gesture Recognition* is the ability to recognize and interpret human gestures.
- *Cognition* is the ability to understand the current situation of the environment and the user needs, and to reason about the world as well as the robot's skills for achieving user goals. In other words, the robot is expected to execute actions and demonstrate that it has understood the user needs and can properly satisfy them.[23]

Each test requires a combination of some of these functionalities. These tests are chosen by a Technical Committee that changes every year, following guidelines from the Executive Committee, the Trustees of the RoboCup Federation, and feedback gathered from the teams of the previous year. When multiple options arise about tests to be selected, after discussion, the Technical Committee votes. In some cases the teams are also involved in these decisions. The criteria for deciding which tests will be selected include (in order of importance): 1. overall goal of RoboCup@Home, 2. applicability for organization, 3. scientific value and challenge, 4. attractiveness for audience and media.

The competition include two kinds of tests: *standard tests*, in which the functionalities and their combinations are decided by the Technical Committee; *open tests*, for which each team can decide which functionalities they want to show. Standard tests are evaluated by the partial scoring system described in the next section, and open tests are evaluated by peer-to-peer evaluation, where each team leader assigns a score to the performance of the other teams.

As explained in Section 3.4, these tests change every year, to make them more difficult and to introduce more unpredictable situations. This evolution is important to prevent *overfitting*, which is the development of local optimal solutions that excessively specialize on a particular instance of a problem without providing general applicability.

Some standard tests are described in the following. Additional details and tests can be found in the competition rulebooks, while videos and additional images can be found on the RoboCup@Home website.

Follow me is a test in which the robot has to follow a person in a crowded area of the competition venue, and enter an elevator with the person to reach another location on a different floor. The leader person (walker) is not known in advance by the robot, and a quick automatic calibration procedure is required at the beginning of the test when the person appears in front of the robot. During the test, others are allowed to cross between the robot and the walker, and at some point the walker hides from the robot who must then reacquire the walker's 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.

Fig. 2 shows some images of the execution of tests over the years. In 2007–2008, the walkers were chosen by the teams and wore special markers to facilitate identification (Fig. 2a). Since 2009, the walkers have been unknown, and they stand in front of the robot at the beginning of the test for easy and fast calibration. Since 2010, the test is run outside the arena in an unknown space for the robot. Since 2011, some pre-defined interferences have been introduced, such as a person crossing between the walker and the robot (Fig. 2b). Finally, since 2012 the test has been conducted in a crowded environment and the robots are required to follow a person through a group and into an elevator to move to another floor (Fig. 2c). This

[23] This definition focuses on RoboCup@Home tasks and is aligned with more general definitions, such as the one given in the *Robotics 2020 Strategic Research Agenda*: "Cognition is the ability to interpret the task and environment such that tasks can be effectively and efficiently executed even where there exists environmental and/or task uncertainty. The ability to interpret human commands delivered in natural language or gestures. The ability to interpret the function and interrelationships between different objects in the environment and understand how to use or manipulate them. The ability to plan and execute tasks in response to high level commands. The ability to work interactively with people as if like a person."

(a) Robot in a supermarket

Category 1
Move to ⟨LOCATION1⟩, get ⟨OBJECT⟩ and put it at ⟨LOCATION2⟩. Go to ⟨LOCATION1⟩, find ⟨PERSON⟩ and follow him/her. ...
Category 2
Bring me a drink. Look for a person in the apartment. ...
Category 3
Find ⟨OBJECT1⟩ at ⟨LOCATION1⟩ (but there is no such object at the given location). ...

(b) GPSR command categories and examples

Fig. 3. Restaurant/Supermarket and General Purpose Service Robot (GPSR) tests. (a) Robot during supermarket test in Singapore 2011, (b) Examples of commands for GPSR test. Depending on the command category, the robot is expected to execute a sequence of actions, acquire missing information about the task if necessary, and detect and report errors in case of erroneous information in the task assignment.

evolution, which shows increasing levels of difficulty for this test, has been fundamental for improving the quality of the solutions implemented by the teams.

In *Cocktail Party* the robot welcomes new guests to the apartment. Five people, who are unknown to the robot, are either sitting or standing in a room of the apartment. When the robot enters the room, three (one after the other) call the robot by waving and order a drink by a speech command when the robot approaches 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 recognition, speech recognition, object detection, and manipulation.

The *Restaurant* or *Supermarket* test is executed in a real restaurant or supermarket (Fig. 3 left). 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 to the robot by the user during this visit. Then the robot receives an order through a speech command to bring food or drink items to some of the locations previously visited. The robot is expected to reach the shelves, grab the correct items, and bring them to the correct locations. This test integrates navigation, mapping, person tracking, speech recognition, object detection, and manipulation.

A special mention is needed for the *General Purpose Service Robot* (GPSR) test, which was introduced in 2010. This test is very different from all the others, since the actual task that the robot must accomplish is not specified before hand, but is given to the robot on-line during the test. The command specifying the task is randomly generated by a 'command generator' program that chooses among a set of templates (not known by the participants), and these templates are instantiated with locations, persons, and objects specified during the set up days. Some examples of templates for the three command categories used by the command generator are reported on the right side of Fig. 3. The sentence generated by the 'command generator' is read by a referee and spoken to the robot. The robot has to understand the desired goal and execute an appropriate behavior. For example, if a user requests "bring me a drink", the robot has to go to a location at which the drinks are stored, grab one, and bring it to the user.

Unlike the other tests, here it is not possible to encode the solution as a deterministic predefined program that combines basic primitive actions and sensing routines; some form of reasoning is needed to properly compose these primitive actions and sensing routines on-line. This test focuses on the ability of the robot to understand its goals and to reason about them to execute an overall behavior and to accomplish them. 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 functionality must demonstrate the robot's ability to understand the current situation and the user requests to perform a complex task that was not specified beforehand. The robot has to show cognitive abilities in user command understanding, planning, and perception, since the tasks may be under-specified and common-sense reasoning is needed to generate effective plans and behaviors. Note that only an external evaluation of this ability is performed. Thus, the robot is evaluated only on the basis of its actual accomplishment of what was required; no internal representations of the knowledge, of the reasoning system, or of the solution are required or evaluated.

It is important to observe that, although cognitive abilities are clearly useful in other tasks, it is difficult to measure them, since it is also possible (for these other tasks) to completely hand-code the problem's solution with minimal or no cognition at all. The current organization of the competition and the scoring system does not allow cases to be distinguished in which cognition is used in other tasks. This is the main reason why we have designed a specific test to measure cognition.

To the best of our knowledge, this is the only existing test or benchmark for measuring the cognitive abilities of a real robot that is performing actions in a real environment and naturally interacting with people.

3.3. Scoring system

The competition's scoring system is designed to measure the capability of the robots to perform tasks related to the overall goals of RoboCup@Home. Scores are assigned for accomplishing these tasks, to measure the effectiveness of the robotic systems in understanding and satisfying typical user needs in a domestic scenario.

After the first two years of RoboCup@Home competitions (2006–2007), we discussed how to evaluate the performances in terms of measuring the improvements demonstrated by the teams in the tests over the years. Two problems were identified: 1. improvements are difficult to measure because the tests change every year, 2. the scoring 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 in a dynamic competition, the second could have been addressed. Therefore, since 2008, the scoring system was changed as described in this section.

In the current scoring system, every standard test is divided into multiple sequential phases (or sub-goals). Each phase, when accomplished, provides the team with a score and enables the next phase. The total score of the test is comprised of the sum of the scores of all the accomplished phases. If all the phases are correctly performed, a full score is gained, and otherwise only a partial score is given. Each phase in a test is evaluated in a Boolean manner: either fully accomplished or not accomplished at all. A partial score refers only to the score of the entire test that can be partial (with respect to the maximum possible score) if not all the phases are accomplished.

As mentioned in the previous section, each test requires proper integration of a subset of the described functionalities, and in standard tests, which are used for the analysis reported in this article, the combination of the functionalities for each test is predefined and identical for all teams. Each phase typically integrates different functionalities, and the sequence of phases impacts the total score, since a phase can be started only if the previous one is accomplished. Given that tests integrate the functionalities in different ways and in different orders, a system with only one extremely good functionality and poor performance in the remaining ones would generally perform worse than a system having average level functionalities. Consequently, although the score is the sum of the scores of single phases, we are still evaluating integrated systems as a combination of different functionalities and not each individual functionality.

This definition of the score reported above is used to compare and rank the teams and to provide the competition's final results. However, to analyze the results of teams during the competition in more detail and to compare results over the years, we need a method to measure the performances of the teams in the tests with respect to the desired functionalities. This further analysis does not affect the rank of the competition; it evaluates 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 accomplish it. When a phase is successfully accomplished, the functionalities associated with it have been successful. On the other hand, if the phase is not accomplished, 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 being tested.

To relate the phase scores with the functionalities, we also define a *weight* for each functionality in each phase of a test. These weights are normalized to 1 and can be intuitively explained as the percentage of the contribution of a given functionality to achieve a test phase.

More specifically, consider test T, divided into n phases $(p_1, \ldots, p_n)$, and denote with $F(p_i)$ the set of functionalities associated to each phase p_i. Then define w_{f,p_i} as the weight of functionality $f \in F(p_i)$ in phase p_i. For each phase p_i, maximum score M_{p_i} is available. When team δ executes a test, it receives score M_{p_i} if phase p_i is accomplished or 0 otherwise. When a phase is accomplished, the gained score M_{p_i} can be assigned to functionalities $F(p_i)$ that were used in this phase by using weights w_{f,p_i}. In the following, we denote with S^{δ}_{f,p_i} the score gained by team δ in phase p_i for functionality f. $S^{\delta}_{f,p_i} = w_{f,p_i} M_{p_i}$ if p_i is accomplished, or 0 otherwise. For example, consider a test phase that requires two functionalities, *Navigation* and *Person Tracking*, with weights 0.7 and 0.3. If the phase is accomplished and a score of 100 points is given, then we can assign 70 points to *Navigation* and 30 points to *Person Tracking*.

The total score of team δ in a given functionality f is thus defined by $\sum_{p_i \in P} S^{\delta}_{f,p_i}$, where P is the set of all the phases of all the competition tests. The normalized score of team δ in functionality f is defined as

$$\sigma^{\delta}_{f} = \frac{\sum_{p_i \in P} S^{\delta}_{f,p_i}}{\sum_{p_i \in P} w_{f,p_i} M_{p_i}}$$

If we consider set of teams Δ, we can compute the average and maximum normalized scores of teams Δ in functionality f as the average and the maximum of values σ^{δ}_{f} for $\delta \in \Delta$.

Obviously, determining weights w_{f,p_i} is not a straightforward operation. In practice, these weights depend on the specific implementations (so in general they should not be the same for each team) and cannot be defined a priori without any knowledge of the system being tested. However, since these information are not available, for completing the analysis reported below, we used an estimation based on our experience. Although approximate, we believe that the results obtained in this way are useful to evaluate the overall progress of the competition based on the average performances of the teams.

3.4. Evolution of tests

To drive the competition over multiple years, the definition of the tests must change. If this is not the case (e.g., a test does not change for several years), ad-hoc solutions to particular instances of the problem defined in the test will be found instead of general solutions. The evolution of tests allows both for avoiding the overfitting of solutions and for leading the competitions to interesting research directions. In the case of conflicts between these two objectives (e.g., pushing a functionality in some direction, but its performance analysis suggests instead a different type of change), we give priority to the avoidance of overfitting. In this way we motivate teams to develop robust solutions to simpler problems before they tackle more difficult ones.

To properly implement a methodology for the evolution of tests, it was necessary to change their definition in such a way to increase or decrease the difficulty of each functionality based on the overall goal of the competition. For example, as described in the next sections, when a functionality was considered too easy, we wanted to increase its difficulty in subsequent years.

The evaluation of the difficulty or the complexity of a task is thus an important feature for our analysis. Some studies have measured task complexity [15] and two main classes of approaches have been considered: *subjective* methods, where the difficulty of a task is measured by involving users who are asked to accomplish the task, and *objective* methods, where such features as the number of components and the relationship between them are identified and measured. Some of these methods require the task to be actually executed to measure its complexity. This approach is not feasible for our purposes, since we need to evaluate task complexity at design time (i.e., before the competition). Moreover, some of these methods provide only task-specific measures, which again are not appropriate in our case. Consequently, no technique is available for a proper quantitative measurement of the complexity of tasks that is adequate for the scope of RoboCup@Home. Nor are we aware of any attempt to quantitatively measure the complexity of tasks in other competitions. Finally, for our purposes of evaluating the evolution of competitions, we do not need absolute measures of a test's complexity, but rather relative measures between its versions that have been changing.

Our approach of measuring task complexity resembles a subjective method. In our case the subjects who were asked to evaluate this complexity are members of the Technical Committee of the competition who define the rules for the next year. They are experts in the field and usually are members of a team. Thus they have experience to evaluate whether a change in the definition of the test would increase or decrease the difficulty of a given functionality. Although this does not allow for a quantitative measure of a task's difficulty, it does allow for a partial ordering of different versions of a test based on the difficulty of its functionalities. We believe that this partial ordering is satisfactory to state that a functionality in a test has been made more difficult in a new version of the test.

An increase of the difficulty of a functionality in a test can be obtained, for example, by implementing one or more of the following changes:

- adding a phase in a test that requires this functionality;
- increasing the number of times or the duration in which the functionality must be used (for example, the robot has to detect 5 persons in the environment instead of 3, it has to reach more target points, it has to navigate in the apartment for a longer time, etc.);
- generalizing a constraint of the test (for example, the people in the environment can either stand or sit with respect to having only standing people, an object that must be found might be located anywhere in the apartment instead of being confined to a single room, the person who speaks can be any random person instead of a predefined team member, etc.);
- increasing the environmental clutter or the background noise (for example, finding an object on a table that contains other objects rather than on an empty table, allowing more people around the individual whom the robot has to identify, etc.).

In the definition of the test extensions, we take into account the overall goals of RoboCup@Home (as stated above) and the environment limitations. For example, adding stairs to the environment to make navigation more challenging is not being considered at this moment, but placing a carpet on the floor might be discussed since it is within the scenario limits we have defined. Another example is given by the objects to be manipulated. We may consider having objects with different shapes to make the grasping phase more challenging, but we are not currently interested in placing objects in difficult locations (e.g., on a high bookshelf), because reaching such a location would require a complex hardware mechanism.

3.5. Benchmarking and collection of data sets

An important feature of a robotic competition is the setup of an environment in which it is possible to test integrated robotic systems in actual operational conditions. Such competition environments are typically difficult to fully reproduce in a laboratory. Therefore the competition set up and the tests performed by the robots are important data sources for benchmarking.

Previous collections of data sets for benchmarking in other RoboCup leagues have been reported: a ground truth system for the RoboCup Small Size League was obtained by overhead cameras mounted on infrastructure above the field [16]; a col-

Table 1
Achieved normalized scores for all functionalities gained by finalist teams: first value is the maximum normalized score, and second value is the average normalized score.

Functionality	Year						
	2008 [%]	2009 [%]	2010 [%]	2011 [%]	2012 [%]	2013 [%]	2014 [%]
Navigation	40/25	47/40	33/20	61/26	52/23	27/18	76/47
Mapping	100/44	100/92	21/10	33/10	10/4	26/15	97/31
Person recognition	32/15	69/37	57/23	48/16	62/15	51/30	66/36
Person tracking	100/81	100/69	100/72	100/76	62/33	51/33	73/64
Object recognition	29/8	39/23	6/1	35/10	56/20	37/13	97/26
Object manipulation	3/1	48/23	29/8	49/21	73/27	45/13	93/27
Speech recognition	87/37	89/71	50/38	76/59	90/56	46/33	60/47
Gesture recognition	0/0	0/0	62/26	100/49	88/37	44/25	84/40
Cognition	–	–	17/3	68/24	32/8	36/15	31/22
Average	48.9/26.4	61.5/44.4	41.6/22.4	63.3/32.5	58.2/24.8	40.3/21.6	75.2/37.8
Standard dev.	41.3/27.3	34.8/30.6	22.1/28.8	25.1/23.4	26.8/16.0	9.4/8.7	19.9/13.3

lection of images in the RoboCup soccer domain was labeled for benchmarking object perception techniques [17]; ground truth position of robots was acquired by laser range finders [18] and by RGBD cameras placed around the field [19]; an automated referee system was described for the RoboCup Logistics League that is also suitable for producing benchmarking data [20].

Moreover, the RoCKIn project and the competition organized within it clearly relate competitions and benchmarking. Teams are required to provide data about the internal states of their robot during the test through a predefined common logging system. Ground truth information are collected during the test with devices and software provided by the organizers. These data are then combined for a more precise measure of the robot performance and for benchmarking specific functionalities.

Using the competition set up to collect data sets might improve their effectiveness, since they would contain the same level of noise that one would encounter at the actual competition. In RoboCup@Home, the collection of data sets to measure the performance of specific functionalities has been performed mostly in research labs. For example, a data set for comparing different methods of person detection and tracking was collected in a lab [21]. HuRIC[24] [22], a data set for benchmarking speech understanding techniques, was partially acquired at competition venues (in particular, RoboCup 2013), but not during the execution of tests. In this work, team members were involved in the acquisition of spoken sentences and they used devices (i.e., microphones) similar to those used on the robots and with a level of background audio noise resembling the one actually faced by the robots during the competition.

Increasing the collection of data sets using competition set ups and introducing benchmarking technology that resembles to the one introduced in RoCKIn are future goals of RoboCup@Home. Gathering information about the internal states of robotic systems being tested during the competition would allow for additional important data sets for benchmarking and for improved evaluation of the competition results. For example, they may provide a better way of determining the weights w_{f,p_i} that relate to the functionalities and phases of a test, as discussed in Section 3.3.

4. Analysis and results

The scoring system described above allows for analysis of the results of the teams in the different functionalities and can also be used to evaluate the progress of the entire competition, as explained in this section. Our analysis aims at understanding the progress of the competitions through the years and at providing the Technical Committee with a tool to decide rule changes for the next year.

The main goal of this analysis is to decide measures to keep a reasonable level of difficulty over the years and to balance the development of all the functionalities. The scores of the functionalities naturally tend to increase, because of the improvements of the participant teams. However, a functionality that has a high score for many years indicates that the addressed problem has become too simple and that specialized solutions for a particular setting have been found. Thus, in these cases, we increase the difficulty of some phases of the tests where this functionality is required, based on the guidelines discussed in Section 3.4.

In Table 1, we show the maximum and average normalized scores in each functionality (as defined in Section 3.3), considering the results of the teams reaching the finals for each indicated year. The maximum value is also reported as a graph in Fig. 4. This view shows at a glance the evolution of the performance in the desired functionalities. We believe both metrics are interesting: best scores give an intuition of level of performance needed to win the competition, while average scores describes how much performance is expected to be an average good team reaching the final. As stated below, the standard deviation over all the functionalities in a year (last row of the table) is also important to assess the balance of the achievements of robotic systems with respect to different functionalities.

[24] sag.art.uniroma2.it/HuRIC.html.

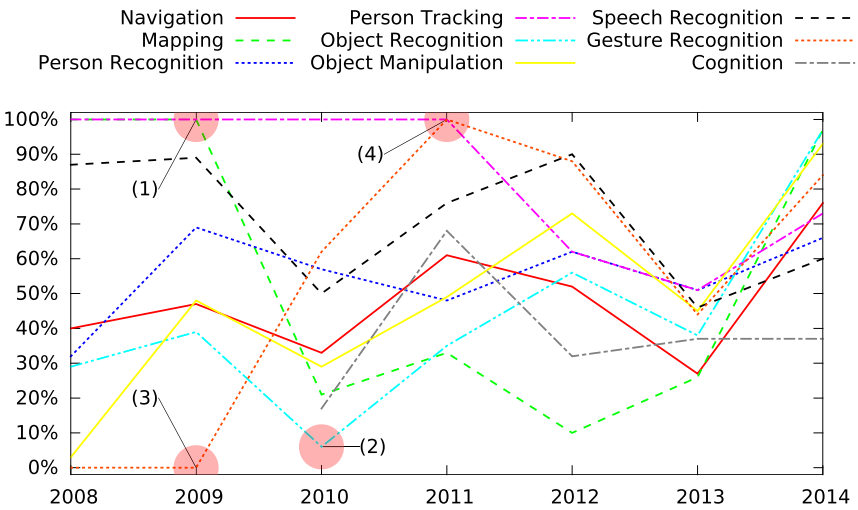


Fig. 4. Results of statistics carried out after every competition. Viewing scores achieved by the best teams (per capability) gives a good impression about what was solved well and where problems persist. Labels (1), (2), (3), and (4) refer to the corresponding numbered items described below.

Table 1 and Fig. 4 show the results since 2008, when the scoring system was changed for this analysis. The analysis of these data allows for monitoring how the Technical Committee's choices in the rule definitions affect team scores. In particular, these data clearly show that when the rules are not changed significantly, performance improves (this happened in 2009, 2011, and 2014), while significant rule changes decrease performance (2010, 2012). These results function as critical feedback for the organizers and the Technical Committee of the competition about the choices made in the rule definitions. They also enable analysis about how the competition is progressing and how to modify the rules to encourage the development of the competitions. For example, the significant increase of performance in 2014 requires a significant change in the rules for 2015 competition.

In addition to general trends, the above analysis is important to identify specific actions to be taken concerning single functionalities. In this way the organizers can better focus the rules to achieve good results in all the functionalities.

Referring to Fig. 4, the following examples of driving the competitions based on the data analysis are provided to clarify our approach.

1. A high level of performance in mapping in 2008 and 2009 was reduced by introducing a more difficult setting for it in 2010. The 'Shopping Mall' test in 2010 was performed in an actual shop of a shopping mall close to the RoboCup venue to provide a more complex scenario than the one of the 'Supermarket' test in 2009 performed in the competition arena.
2. Object recognition is a core task in all the RoboCup@Home tests. Consequently, the teams invest much time to implement reliable approaches. Whenever the average object recognition performances mature, we enlarge the set of known objects and introduce other challenges, e.g., detecting and manipulating previously unknown objects. Through 2009, the set of known objects was limited to ten and all manipulation tests were conducted in the known RoboCup@Home arena. Increasing the number of objects and introducing manipulation tests in an initially unknown environment led to a considerable performance drop in 2010, and thus the complexity was not increased for 2011. The object recognition skills of the robots quickly matured, and since 2012 both complexity and performance have steadily increased. Starting in 2013, the set of known objects also includes objects that are hard to recognize, e.g., objects being textureless or transparent objects, as well as objects of different sizes.
3. The Gesture Recognition performance in 2010 was improved by modifying some phases in the 'Follow me' test where the user had to communicate with the robot at a distance. Introducing such a requirement made it more convenient to use gestures rather than speech, since gesture recognition is more reliable than speech recognition if the person interacting with the robot is far away.
4. Person Tracking difficulty was increased in 2012 by increasing the amount of background noise in the 'Follow me' test, allowing many people within the area were the test was performed, thus increasing the difficulty for the robot to track the walker.

As a final remark on the developed analysis, we want to highlight that in the development of integrated research in AI and robotics, a good balance of all the functionalities is required. However, this balance was absent in RoboCup@Home in the first few years. As shown in Table 1 and in Figs. 4 and 5, in the first years of the competitions some functionalities were fully achieved (100%) and others completely failed (0%). Over the years, the scores of the functionalities with high performance have decreased, and the scores of the functionalities with low performance have increased, producing a trend in which all the functionalities have become more balanced. For example, in 2013 we saw average performance around 50%

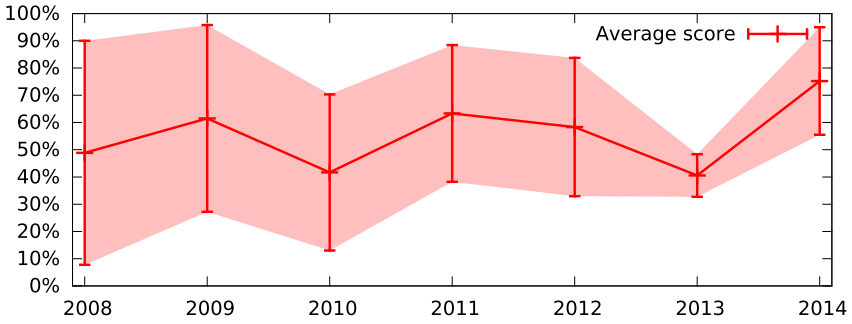


Fig. 5. Average and standard deviation of maximum normalized score of finalist teams.

with a very low standard deviation and in 2014 a higher average performance around 75% with a slightly higher standard deviation. Changing the rules is the tool with which we obtained this result. Recall how we dealt with the functionalities Mapping and Gesture Recognition in 2009 (described above), whose respective scores were 100% and 0%. In 2010 the rules increased the difficulty of Mapping and made Gesture Recognition mandatory in a test. As a consequence, the performance of the former fell to 21% and the latter increased to 60%, contributing to the overall reduction of the standard deviation of the scores in 2010.

The reduction of the standard deviation obtained over the years in RoboCup@Home (Fig. 5) demonstrates the efforts of the competition organizers and the teams in developing systems that can properly integrate many functionalities.

In summary, by monitoring the results provided by the data collected from the team scores, the competition organizers can check the progress of the results over the years, control the average scores and the standard deviation of each functionality, and make strategic decisions to change the rules to adjust the difficulty of the tasks and the functionalities based on the desired goals of the competition.

5. Overview of best technical solutions

In this section we survey the best technical solutions in the different basic functionalities that are required to accomplish the tests of the competition. Since these technical solutions are usually of general applicability, they are suitable for implementing functionalities that are also required in other applications outside the RoboCup@Home competition. The solutions described in this section are grouped by the main functionalities that have been identified and composed in the definition of the RoboCup@Home tests, as described in Section 3.2. Naturally, we focus on the top performing teams as well as on the teams that successfully demonstrated a particularly interesting approach or a very robust approach for a particular skill.

As shown in the analysis reported in the previous section and in the description of the proposed solutions below, the changes in the competition rules enabled teams to address specific solutions to the problems in different years. When addressing these problems, the teams are faced with the choice among mainly three alternatives; to keep using implemented components, to implement known methods, or to develop their own solution. This choice also reflects the specific expertise and research interests of the team members. Moreover, actual usability of the deployed components, ease of use, and integration are important features—often more important than the peak performance of a given component.

Note that many RoboCup@Home-related publications about solutions to these technical problems are published in the RoboCup Symposia and also in other top international robotic conferences and journals. For example, two special issues on Domestic Service Robotics have been organized by the RoboCup@Home league [23,24], where the most innovative research works developed by the @Home teams (but not only) were published. Indeed the development of effective service and domestic robots still requires research to develop robust and effective solutions and to properly integrate them. Just integrating standard off-the-shelf components is insufficient to form a competitive team.

In the remainder of this section, we overview several approaches used by the RoboCup@Home teams, which approaches are considered the most reliable, and which are particularly interesting for extending the state-of-the-art. Our main objective is to describe common solutions that have been well tested in the competitions and can be re-used in other application domains in the field of domestic and service robots.

5.1. Navigation, localization and mapping

Wheeled motion on horizontal surfaces, i.e., 2D Navigation, is well understood. Both the algorithms and publicly available implementations exist that allow mobile robots to build 2D maps, localize themselves on these maps, plan and follow their planned paths to reach target locations. Basic navigation abilities, such as making robots work and move in the arena, were a problem in the first years of the RoboCup@Home league. With the experience gathered over the years and the advent

of easy-to-use frameworks with out-of-the-box running components such as the Robot Operating System (ROS),[25] however, basic navigation has essentially been solved. Basically, all the teams in RoboCup@Home have shown in the past years that they can autonomously build maps and safely navigate. Most of the teams use ROS and the navigation stack of ROS. It contains the following ready-to-use implementations of the state-of-the-art algorithms: a Rao–Blackwellized particle filter (gMapping) [25] for map building, adaptive Monte Carlo localization (AMCL) [26] for localizing the robots on the built maps, A* path planning [27] and vector field histograms (VFH) [28,29] for following the planned paths and avoiding collisions. Over the past releases, the ROS navigation stack has considerably matured, becoming easy to use and well-documented. Some teams only use (or extend) particular components, e.g., gMapping and A* with an adapted AMCL and a dynamic window approach [30] for local navigation. The latter yields particularly smooth trajectories by explicitly taking into account the vehicle dynamics.

In principle, basic (2D) navigation problems is considered to be solved and in practice both the teams and the tests in the competition focus on particular problems that arise in domestic service robotics: navigation in highly dynamic environments (e.g., following a particular person through a crowd) or collision avoidance in complex scenes (requiring 3D perception).

Finally, even though metric mapping is again very well studied and many implementations are available, semantic mapping (i.e., adding semantic information to the representation of the environment) remains a non-trivial challenge. An interesting approach to semantic world modeling was demonstrated by the Tech United team [31] who presented a framework for probabilistic multiple hypotheses anchoring, especially for perceiving and tracking objects in the robot's workspace. Regarding semantic scene analysis, the ToBi team developed an Implicit Shape Model (ISM) that can learn the spatial relationships of typical object regions from a set of artificial 3D models [32]. To learn the spatial relationships of these identified regions the ISM stores descriptors for the appearance of these object regions in relation to a unique object reference point. Each detected keypoint in the training models, therefore, is matched against the generated codebook. For each matching codeword, a 3D vector is added that describes the relationship between the detected point and the object centroid. To detect object instances of the learned category in test scenes captured with a 3D camera, probabilistic Hough voting is performed. This enables the vision system to simultaneously recognize and localize object instances. Furthermore, additional grid map layers on top of the SLAM obstacle map are introduced by "Semantic Annotation Mapping" that encodes low-level visual cues calculated while the robot explores its environment. By taking into account that text might become a valuable source of semantic information for robots, the b-it-bots team developed a robust text recognition system with applications for semantic scene analysis [33].

5.2. Person recognition and tracking

Most RoboCup@Home teams use RGB-D sensors (Microsoft Kinect or Asus Xtion) and standard CCD/CMOS cameras as the main sensors to implement vision-based functionalities. Some teams complement them with stereo cameras, thermal cameras, and actuated 2D laser scanners that allow 3D perception. In terms of the vision algorithms, most teams rely mainly on functionalities provided by such libraries as OpenCV, OpenNI and ROS/PCL. OpenCV is typically used for face detection (e.g., the Viola&Jones face detector), object detection, and object recognition (e.g., SIFT/SURF based object detection, color histograms). OpenNI is used for person detection and gesture recognition, while ROS/PCL is used for managing point clouds and for 3D object recognition. Face recognition methods normally are based on simple algorithms such as eigenfaces.

In terms of human face recognition and analysis, the b-it-bots team addressed the recognition of facial expressions. Their developed system is based on the extraction of Gabor features at different orientations and scales [34]. The features are first extracted from a normalized face image and forwarded to a multi-classification stage. The number, location, orientation and scale of the Gabor filters are determined in the training phase by Adaboost. The UChile team addressed face recognition using visual and thermal images [35,36]. After comparing several methods, they concluded that the best-performing method for HRI applications, in terms of recognition rate and real-time operation, is LBP-histograms using histogram intersection or Euclidean distance as similarity measures in the visual and thermal spectra. Based on similar ideas, another work [37] proposed the robust detection of humans through thermal and visual information sources that are integrated to detect human-candidate objects, which are further processed to verify the presence of humans and their identity using face information in the thermal and visual spectra. Face detection is used to verify the presence of humans, and face recognition is used to identify them. Active vision mechanisms are employed to improve the relative pose of a candidate object when direct identification is not possible.

5.3. Object recognition

Object recognition is a critical functionality for a service robot, and many research results are available from the computer vision community. However, implementation on a mobile robot that requires real-time performance with limited on-board computational power is not always straightforward.

The recognition of objects in RoboCup@Home has become more complex over the years: from just a few different, previously known objects to a set of 25 known and 10 unknown objects randomly used in the tests. Most teams in RoboCup@Home follow the successful pipeline approach that first detects objects on horizontal surfaces as candidates and then

[25] www.ros.org.

recognizes them using a variety of features. Popular visual appearance-based feature keypoints and descriptors are SIFT [38], SURF [39], FAST [40], BRIEF [41] and RIFF [42]. A similar *feature zoo* has evolved in the 3D community that yields many different 3D feature descriptors capturing the spatial characteristics of objects. Popular examples include FPFH [43], VFH [44], and SHOT [45] as well as the Point-pair feature (PPF) that uses Hough voting [46] or RANSAC [47]. For recent comparisons of feature descriptors we refer to [42] for visual features and [48] for 3D features. Here, we focus on approaches that were demonstrated in RoboCup@Home and that yielded good performances.

Very robust object recognition, even under considerably changing lighting conditions and previously unknown environments, was demonstrated by WrightEagle@Home. They also follow a pipeline approach (detection followed by recognition) and use a combination of SIFT features [38] and LINEMOD [49].

Also noteworthy is the approach by Homer@UniKoblenz that achieved the highest score in the 2012 technical challenge on (active) object recognition. They used SURF [39] and Hough transform clustering. In order to achieve particularly high recognition reliability, they used high resolution photos acquired by a digital camera. As required by the RoboCup@Home rules on the technical challenge, the team published the source code and a paper on the approach.[26]

Another outstanding achievement in the league is the continued work on 3D object perception and tracking by NimbRo@Home. Among others, they successfully demonstrated real-time 3D object recognition and tracking for physical human–robot-interaction [50], tool use [51], and recently non-rigid object recognition and tracking for skill transfers [51].

In addition to object recognition, techniques have been developed to effective search for objects in the environment. The Tech United Eindhoven and UChile teams addressed active/indirect object searches. Tech United Eindhoven proposed a probabilistic object–object relation based approach for an active object search [52]. Its main contribution is a strategy that allows a chain of intermediate objects to be used for active object searches. This work received the best paper award at the RoboCup 2013 Symposium. The UChile team proposed a Bayesian framework for informed search using convolutions between observation likelihoods and spatial relation masks. With spatial relation masks, complex spatial relations between objects can be defined as weighted sums of basic spatial relations using co-occurrence matrices as weights.

5.4. Object manipulation

Manipulating objects is a twofold problem incorporating both the perception of objects and controlling the robot's movement to manipulate the perceived object, which includes the planning of these movements. Object manipulation tasks in RoboCup@Home have become more complex over the years: from no manipulation at all to such complicated problems as grasping objects in cluttered, unknown environments, and varying heights (ranging from a floor to high shelves). In the first years of the competition, the manipulation of objects was not mandatory in all tests; it was only present in a basic form and as extra demonstrations in the Open Challenge test. This was primarily caused by the fact that almost no team had a manipulator on its robot.

The first tests that addressed object manipulation were rather simple and did not contain any form of recognition. That is, the object to be manipulated was the only one at the designated location. Thus, the strategies applied by the successful teams were rather simple. NimbRo@Home, for example, used a rotatable 2D laser scanner in the robot's upper body to first scan vertically in order to estimate the table height and distance. Next it scanned horizontally above the table plane to detect the object and finally oriented the robot toward the detected object and grasp it with simple motion primitives. No object recognition or motion planning were involved.

In the last few years, the object manipulation tasks in the tests actually require some form of object recognition and object positioning (e.g., there may be many objects on a table at random positions and the robot has to grab the correct one). Consequently, vision and in particular RGB-D cameras have been used both to recognize the object and to determine its pose to plan effective grabbing. A typical approach adopted by many teams is based on 3D table top segmentation using an RGB-D camera mounted on the robot's head to detect and track objects at the designated grasping location in real-time [53]. This configuration also allows the dimensions of the detected objects to be estimated to perform grasp planning [54]. In this way the capability of manipulating objects of different shapes and sizes has been achieved by many teams.

The majority of the arms and the grippers used by RoboCup@Home teams are either low-cost commercial hardware or custom built hardware. Only a few teams have professional robot arms. For example, the b-it-bots team [55] uses a KUKA lightweight arm [56] on a Care-O-Bot 3 [57] platform. Many other teams use the considerably less feature-rich but also less expensive Neuronics Katana Arm.[27] Most platforms are wheeled with either one centered manipulator or with the more recent and widely accepted design of an anthropomorphic upper body with a sensor head and two arms with grippers.

The upper body design allows for human-like reach. An additional actuator lifts the whole upper body and allows for manipulation of both objects lying on the ground and on the horizontal surfaces of such different heights as side tables, dining room tables, and shelves. Another noteworthy hardware design that achieved similar variety in manipulation heights is that of Homer@uniKoblenz, who combined a simple gripper in the robot base to grasp objects on the ground and an arm on top of the base to manipulate on higher surfaces. Their robot's arm is mounted so that it can take objects out of the gripper in the base.

[26] Its source code and paper are available at wiki.ros.org/obj_rec_surf.

[27] Neuronics Katana family of robot arms: www.neuronics.ch.

For actually grasping objects, most teams simply use inverse kinematics to compute an arm configuration that aligns the gripper and the object to be manipulated. Generally, arm configurations are simply interpolated, ignoring the surroundings and possible collisions. Only a few teams use grasp and motion planning which is, however, not yet required for most of the RoboCup@Home tests (the increased complexity of object manipulation in the league is making it more necessary to efficiently plan grasps and arm motions). Some teams report using a combination of grasp/motion planning and parameterized motion primitives for particular manipulation tasks [51,58], allowing for complex grasping in complex scenes. For the control of arm and body motions, an approach has been successfully demonstrated that is compliant in task space [50]. It enables many important robot capabilities, such as opening refrigerators, cupboards and drawers or cooperatively carrying a table with a human.

The integration of object recognition and grasp planning has been addressed by combining all navigation, perception, and manipulation capabilities in a decision-making framework that seamlessly integrates human–robot interaction and planning [59]. In particular, LINEMOD [49] and SIFT [38] are used for object recognition and plain inverse kinematics for grasping.

Finally, with respect to the available algorithms and implementations, the popular ROS manipulation stack is not as stable and mature as the navigation stack. An alternative is the recently started MoveIt! project,[28] which aims to provide a complete manipulation framework that incorporates sensing, planning, and control for robot grasping. In the recent years, many teams are successfully using MoveIt! as the basis for object manipulation tasks.

5.5. Speech recognition and synthesis

Speech recognition in RoboCup@Home environments is very challenging because distant-talking Automatic Speech Recognition (ASR) is required in noisy environments. In standard situations, the distance between the user and the robot is at least 50 cm. The equivalent noise level (L_{eq}) is approximately 75 dB, and the maximum noise level is 85 dB. The noise sources specific to RoboCup@Home include crowd roars from the RoboCup Soccer league, music, and commentary through high-power speakers.

In this setting, the speech recognition module is required to have a robust noise reduction functionality. The solutions in RoboCup@Home can be roughly categorized into hardware- and software-based approaches. For hardware-based noise reduction, directional microphones were introduced in 2008, and currently most top teams adopt them. For software-based noise reduction, eR@sers' method proved to be very effective in RoboCup@Home environments [60]. This method is based on on-line noise tracking with a particle filter and switching dynamical systems [61].

A grammar-based approach works in most tasks where speech commands are limited to such names as "living room" and "John". On the other hand, large vocabulary speech recognition is required in the *General Purpose Service Robot* (GPSR) test since the commands are randomly generated by the GPSR sentence generator. The ASR software used the by top teams includes Loquendo ASR, Microsoft Speech API, iFLYTEK, and CMU Sphinx.[29]

Another important issue is to learn and recognize out-of-vocabulary words (OOVs). Until 2008, the robots were required to learn and recognize person names. This is very difficult since the participants are from many countries, and most of their names are OOVs for ASR systems. From 2009, a standard name set is defined instead, and out-of-vocabulary words does not appear in the standard tests.

For text-to-speech (TTS), many available solutions have been used by the teams. Common software include Loquendo TTS, Mac OS X Speech Synthesis API, and Festival.[30] TTS is not explicitly evaluated in RoboCup@Home. However, points are given only if the robot explains the situation well in particular tests (e.g., it declares that a specific object has been found or a command has been understood), and the rules clearly state that teams are responsible for a sound system that allows referees to understand the robot's speech. Thus a good TTS (with a clear voice in noisy environments) is important to score points.

5.6. Gesture recognition

Most teams have addressed the recognition of hand gestures over the past few years. In particular, in the b-it-bots system, first the user's face is detected and tracked with the CAMSHIFT algorithm. Then a skin color histogram is extracted, and a back projection image is calculated. For pointing with a fist, the average recognition rate is 90.28% for a distance up to 216 cm (head to pointing target). When pointing with a fingertip, the recognition application has an average recognition rate of 81.25%–87.5% for short distances [62].

In the NimbRo@Home system, the positions of the head, hand, shoulder, and elbow are first determined and processed in order to interpret gestures [63]. The perception is based on the detection of body parts in amplitude images as well as body segmentation in the 3D point clouds of the camera. Such gestures as pointing are further interpreted for their parameters, e.g., for the pointing direction.

[28] moveit.ros.org.
[29] cmusphinx.sourceforge.net.
[30] www.cstr.ed.ac.uk/projects/festival.

UChile's system detects hands and static gestures using a cascade of boosted classifiers and recognizes dynamic gestures by computing the temporal statistics of hand positions and velocities and classifying these features using a Bayes classifier. Context information is used to continuously adapt the skin model used in the detection of hand candidates, to restrict the image regions that need to be analyzed, and to reduce the number of scales that need to be considered in the hand-searching and gesture-recognition processes. On average, the system recognized static gestures in 70% of the cases, dynamic gestures in 75%, and runs at a variable speed of 5–10 frames per second [64]. Using this system, the robot was able to play rock-paper-scissors against a human in real-time without markers.

5.7. Cognition

As mentioned in Section 3.2, cognition in RoboCup@Home is defined as the ability of a robot to understand the user goals and to reason about its own skills. The cognitive abilities of robots are measured by the *General Purpose Service Robot* test, where the robots have to solve tasks generated on-line that are not completely specified beforehand. The development of on-line reasoning capabilities for robots is fundamental in this test, since it is impossible to encode the robot behavior before the test.

Moreover, all the tests require different combinations of basic functionalities. Thus, a reasoning system, which can properly combine the robot's skills based on particular user needs and environment situations is crucial for improving the modularity of the system, its overall robustness, the ease of debugging, etc. Consequently, in the RoboCup@Home's scenario cognitive abilities are actually important and explicitly embedded and measured in the competition tests.

However, measurement of this functionality and the *General Purpose Service Robot* test were introduced recently in 2010, and admittedly this capability remains incomplete among the RoboCup@Home teams. Indeed, details of the techniques used to address this problem are not generally available, with some notable exceptions described below.

The WrightEagle@Home team describes an approach, used in the *General Purpose Service Robot* test, that is based on open knowledge available as semi-structured data and on reasoning procedures provided by an Answer Set Programming (ASP) solver. The system can incrementally increase knowledge about an environment and reason about it [65,59]. A different approach based on Situation Calculus was developed by the AllemaniACs team. They proposed Readylog, an extension of Golog, that represents and reasons on continuous action execution, the probabilistic effects of actions, qualitative positional information, etc. Execution monitoring ensures proper execution of actions [66].

Since we strongly believe that cognition is an important ability for service and domestic robots, we will encourage the competition more in this direction in the near future and thus we expect more interesting results from RoboCup@Home teams.

6. Lessons learned and future plans

An important aspect of the league development is not only to evaluate the team performances and shape changes for the next year but also to evaluate the competition as a whole and subjectively measure its success in terms of the satisfaction of the participating teams, the competition organizers, and the experts who provide support, as well as awareness outside of the RoboCup community.

This section presents the lessons we have learned during the organization of the RoboCup@Home competition that we believe are important success factors as well as future plans for its development to address the remaining open issues.

6.1. Lessons learned

The lessons that we learned during seven years of RoboCup@Home can be grouped into two types: those related to the evolution of the competitions and those related to the definitions of the tests.

The most important lessons related to the organization and the evolution of the competitions are illustrated below.

Lesson 1*. *Define a score system that allows for the analysis of performance over time. This was the main contribution of this article and it has been fully described in the previous sections. As mentioned, it is a fundamental tool for driving a competition over the years and we thus believe it is crucial for other competitions as well.

Lesson 2*. *Make appealing competition for the teams. By far the most important aspect of a scientific benchmark is its acceptance in the research community. In case of a competition, this translates to the participants. A competition can only be as good as the participating teams. Thus, the RoboCup@Home competition must be challenging for good teams, allow other teams to catch up and new teams to easily enter. Since RoboCup@Home integrates many different skills and research problems, the entry level is relatively high (especially due to the integration aspect). However, we aim at providing methods and means to ease the entry for new teams by fostering the public release of developed components and developing a standard in terms of both open access software and hardware. Even though many steps have addressed the former (e.g., releasing packages under ROS and working on a special distribution for collecting all components for RoboCup@Home), the latter still remains a matter for future work (for example, defining a Standard Platform that has been proven to considerably increase the interest of other research groups in other RoboCup leagues—especially in soccer). In addition, to increase team participation and awareness

of the problems in organizing a competition, team members are warmly invited to participate to its organization as referees, members of the Organizing Committee, Technical Committee, and Executive Committee.

Other important lessons listed below are instead concerned with the definition of the tests.

Lesson 3*. *Provide a good balance between pre-defined tests and open demonstrations. We identified the importance of providing a good balance between pre-defined tests, where the desired skills need to be shown in desired robustness and efficiency, and open demonstrations, where teams can show cutting edge research results and problems that have not yet been addressed in the design of pre-defined tests. In the past, the latter has often led to the definition of new tests by uncovering previously neglected problems or situations. Open demonstrations also give teams the opportunity to focus on particular aspects that are not necessarily among the desired skills defined in RoboCup@Home.

Lesson 4*. *Define tests that require skills that address interesting research questions. Since RoboCup@Home is a scientific competition, teams must be motivated to solve important research problems. To this end, it is important to consider skills that address interesting research questions and pose state-of-the-art problems within the tests. At the same time, the tests need to be established in such a way that they are solvable in a limited amount of time and can be easily evaluated.

Lesson 5*. *Design attractive tests for non-expert audience. Since RoboCup seeks a certain standard of quality in terms of both education and entertainment, it also wants tests that non-expert can easily follow, while remaining attractive and interesting for both experts and non-experts. In the past, concentrating on this aspect has resulted not only in interesting and entertaining events for the audience but also for teams that must pay particular attention to robot appearances, their interaction and how they are perceived. For example, the majority of robots in RoboCup@Home possess the ability to intuitively present (to both experts and non-experts) what they are doing and what they are perceiving. In fact, robots that not only have a well-designed appearance but also a well-designed means of interaction are received considerably more warmly by non-expert operators and spectators. Two particular aspects must be mentioned here: 1. audio feedback in the context of speech recognition (e.g., signaling when the robot expects inputs or whether it has understood them) and 2. visual feedback and/or abstract information about robot internal status for debugging purposes (e.g., reporting what the robot is doing or what went wrong).

6.2. *Future plans for* RoboCup@Home

Establishing a well-defined and widely accepted benchmark for integrated systems and domestic service robots is an ambitious aim. However, we believe that such a competition as RoboCup@Home (even without having completely matured) is an important step towards this goal. The conducted statistics and evaluation of scores (and the subsequent amendments of the rules that define the competition) show that RoboCup@Home is on track to support integrated robotic systems that safely interact with environments and people. However, it also remains clear that our ultimate goal is far from being achieved. The changes along this path can roughly be split into the following categories:

- Competition-specific changes (adding/removing/changing tests, changing competition's format, etc.)
- Skill-specific changes (introducing new skills, increasing skills complexity, etc.)

On the competition level, we will foster the league's benchmarking characteristics, by introducing the following objectives.

Objective 1*. *Repeatability of tests. In the current implementation, many tests are only run once. A single failure can thus reduce or completely lose points. We plan to change the design of the competition and the individual tests to increase the number of repetitions for which a particular skill is being tested. This will not only yield fairer final results but also produce better statistics for both the comparative evaluation of team approaches and for driving future changes. This feature has already successfully incorporated in the RoCKIn@Home competition.

Objective 2*. *Refinement of scoring system and acquisition of robot's internal data. Again inspired by RoCKIn@Home, a refinement of the scoring system and the acquisition of additional information about a robot's status during the execution of tests (e.g., internal log data, ground truth data with external sensors, etc.) are being discussed to allow for further analysis that would better relate the functionality and task performance for every team. This would be very useful to answer questions that at this moment neither we nor the teams themselves can answer: "why did a test fail?", "which functionality has been the best/worst during this test?", "how much did a team improve over the years in a given functionality, regardless of the changes in the tests?", etc. This feature will also enable for the acquisition of benchmarking data sets that are important for off-line evaluation and comparison of different solutions.

Objective 3*. *Measuring test complexity. When designing tests, their complexity and difficulty must be quantitatively measured to properly calculate the performance differences that depend on the differences in task difficulties, as well

as to assess the competition's progress in solving more difficult tasks. Defining a method for the quantitative measure of the complexity of robotic tasks at design time is a challenging direction for future research in robotics competition and benchmarking.

Objective 4. Real-world application. In contrast to the RoboCup@Home arena in which most tests take place, the real-world is unpredictable, often crowded, and may require the robot to interact with people who are not used to seeing, hearing, or operating robots. The league currently features two tests that take place in public areas. These areas are modified by controlling direct access to the audience or simplifying the task and the environment, e.g., by avoiding problems that, from the experience of the Technical Committee, are unsolvable right now. However, we plan to perform more tests in the real world with fewer customizations and simplifications.

On the skill level, we plan to address the following challenges.

Objective 5. Endurance and long-term operation. In contrast to soccer games that have a defined duration, real-world service robot tasks are open-ended. Moreover, the tasks being tackled may take considerably longer than what is currently possible with a single battery charge. To enforce endurance, teams need to work on both hardware and software sides for reliably extending the operation time of robots simultaneously addressing the problems that arise 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 ten minutes) and solve tasks on demand. We plan to run more parts of the competition for longer periods of time, possibly with several robots being simultaneously tested. We believe that this will also make the competition more interesting for spectators since something is always happening.

Objective 6. Semantic perception and mapping. Up to now, semantic perception and mapping capabilities are not explicitly tested in RoboCup@Home. We plan to foster the development of such capabilities by introducing new tests that also address cognitive skills. A possible test might be to search for a location in an initially unknown environment, such as looking for the kitchen or the restroom in a house or restaurant.

Objective 7. Intuitive multi-modal interaction. When the league started, the only permitted way of interacting with the robot was by natural speech (i.e., spoken commands). While this was the most obvious and natural interaction way, it is not always the most convenient or successful. For example, in such really loud places as the RoboCup venue during an interesting soccer match, speech commands are doomed to fail. Instead, we are aiming at multimodal intuitive interaction by combining speech and gestures as well as buttons/displays on the robot for direct cooperation and intuitive touch pad interfaces for remote command and operation of the robots.

7. Conclusions

In this article, we presented analysis and performance measure of RoboCup@Home competition over time. We also described a set of notable techniques that have been developed and integrated in the robotic systems that have participated to the competitions.

As discussed, a careful design of the scoring system allowed us not only to rank the teams in the competition but also to analyze the general performance of the teams. This is useful to drive the competitions over the years by balancing the use of different functionalities and by progressively increasing the task difficulty.

In general, the presented approach allows some desired functionalities to be related with a set of tasks in which these functionalities are combined. When tasks evolve, the approach enables the analysis of the results over time, and these results are useful for decisions about modifying such tasks. We believe that this approach, which has successfully improved the RoboCup@Home competition, can be adopted by other scientific competitions or even in the evaluation of a research program that continues to operate for many years, when functionalities and tasks are properly associated.

Although the results obtained so far have been very satisfactory, future developments must be addressed. One direction is refining the scoring system toward a deeper analysis of the contribution of the task functionalities. This can be done by requiring the teams involved in a test to provide information about their internal states, e.g., by a logging utility. Better ways of measuring the task complexity/difficulty at design time must also be studied to improve analysis. Another direction is to increase benchmarking efforts so that scientific competitions can actually provide an effective way of benchmarking systems that integrate Artificial Intelligence and Robotics techniques. In this respect, acquiring additional data sets and better exploitation of the competition field for benchmarking purposes is planned for future competitions. Finally, we intend to increase our focus on testing the cognitive abilities of the robots. The idea of a GPSR test with robots that can reason about the world and their actions to complete tasks that are not specified beforehand will be further exploited in other tests or phases of tests.

We believe that a deep analysis of the results of scientific competitions and goal-driven management of yearly changes will improve the development of more innovative systems as well as the research quality of the solutions developed by the participating teams. Although the RoboCup@Home approach presented here may be improved in many directions, we consider it an important step towards a more scientific management and evaluation of evolving competitions.

References

[1] M. Anderson, O.C. Jenkins, S. Osentoski, Recasting robotics challenges as experiments, IEEE Robot. Autom. Mag. (2011) 10–11.
[2] F. Amigoni, A. Bonarini, G. Fontana, M. Matteucci, V. Schiaffonati, Robotic competitions as experiments: a critical view, in: Proc. of Italian Workshop on Artificial Intelligence and Robotics, AIRO, 2014.
[3] F. Amigoni, A. Bonarini, G. Fontana, M. Matteucci, V. Schiaffonati, Benchmarking through competitions, in: Proc. of 2nd Workshop on Robot Competitions: Benchmarking, Technology Transfer, and Education, 2013.
[4] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[5] D. Young, J. Ferryman, Pets metrics: on-line performance evaluation service, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, 2005, pp. 317–324.
[6] B. Juang, L.R. Rabiner, Automatic speech recognition—a brief history of the technology development, in: Encyclopedia of Language and Linguistics, Elsevier, 2005, pp. 1–24.
[7] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, M. Federico, Report on the 10th IWSLT evaluation campaign, in: Proceedings of the International Workshop on Spoken Language Translation, 2013.
[8] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[9] A.W. Black, S. Burger, A. Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, et al., Spoken dialog challenge 2010: comparison of live and control test results, in: Proceedings of the SIGDIAL 2011 Conference, 2011, pp. 2–7.
[10] K. Striegnitz, A. Denis, A. Gargett, K. Garoufi, A. Koller, M. Theune, Report on the *second* second challenge on generating instructions in virtual environments (GIVE-2.5), in: Proceedings of the 13th European Workshop on Natural Language Generation, 2011, pp. 270–279.
[11] S. King, V. Karaiskos, The Blizzard Challenge 2012, in: Proceedings of the Blizzard Challenge 2012 Workshop, 2012.
[12] T. Baarslag, K. Fujita, E.H. Gerding, K. Hindriks, T. Ito, N.R. Jennings, C. Jonker, S. Kraus, R. Lin, V. Robu, C.R. Williams, Evaluating practical negotiating agents: results and analysis of the 2011 international competition, Artif. Intell. 198 (2013) 73–103.
[13] T. Balch, H.A. Yanco, Ten years of the AAAI mobile robot competition and exhibition: looking back and to the future, AI Mag. 23 (1) (2002) 13–22.
[14] T. Wisspeintner, T. van der Zant, L. Iocchi, S. Schiffer, RoboCup@Home: scientific competition and benchmarking for domestic service robots, Interact. Stud. 10 (3) (2009) 392–426.
[15] P. Liu, Z. Li, Task complexity: a review and conceptualization framework, Int. J. Ind. Ergon. 42 (2012) 553–568.
[16] S. Zickler, T. Laue, O. Birbach, M. Wongphati, M.M. Veloso, SSL-vision: the shared vision system for the RoboCup small size league, in: RoboCup 2009: Robot Soccer World Cup XIII, 2010, pp. 425–436.
[17] R. Dodds, L. Iocchi, P. Guerrero, J. Ruiz-del Solar, Benchmarks for robotic soccer vision, in: RoboCup 2011: Robot Soccer World Cup XV, in: Lecture Notes in Computer Science, vol. 7416, 2012, pp. 427–439.
[18] R. Marchant, P. Guerrero, J. Ruiz-del-Solar, A portable ground-truth system based on a laser sensor, in: RoboCup 2011: Robot Soccer World Cup XV, 2012, pp. 234–245.
[19] A. Pennisi, D.D. Bloisi, L. Iocchi, D. Nardi, Ground truth acquisition of humanoid soccer robot behaviour, in: RoboCup 2013: Robot Soccer World Cup XVII, in: LNAI, vol. 8371, Springer, 2014, pp. 560–567.
[20] T. Niemueller, D. Ewert, S. Reuter, A. Ferrein, S. Jeschke, G. Lakemeyer, RoboCup logistics league sponsored by Festo: a competitive factory automation testbed, in: RoboCup 2013: Robot Soccer World Cup XVII, in: LNAI, vol. 8371, 2014, pp. 336–347.
[21] W. Pairo, J. Ruiz-Del-Solar, R. Verschae, M. Correa, P. Loncomilla, Person following by mobile robots: analysis of visual and range tracking methods and technologies, in: RoboCup 2013: Robot Soccer World Cup XVII, 2014, pp. 231–243.
[22] E. Bastianelli, L. Iocchi, D. Nardi, G. Castellucci, D. Croce, R. Basili, RoboCup@Home spoken corpus: using robotic competitions for gathering datasets, in: RoboCup 2014: Robot World Cup XVIII, 2015, pp. 19–30.
[23] L. Iocchi, J. Ruiz-del-Solar, T. van der Zant, Domestic service robots in the real world, J. Intell. Robot. Syst. 66 (1–2) (2012) 183–186.
[24] L. Iocchi, J. Ruiz-del-Solar, T. van der Zant, Advances in domestic service robots in the real world, J. Intell. Robot. Syst. 76 (1) (2014) 3–4.
[25] G. Grisetti, C. Stachniss, W. Burgard, Improved techniques for grid mapping with Rao-Blackwellized particle filters, IEEE Trans. Robot. 23 (1) (2007) 34–46.
[26] D. Fox, KLD-sampling: adaptive particle filters and mobile robot localization, Adv. Neural Inf. Process. Syst. (2001) 26–32.
[27] P. Hart, N. Nilson, B. Raphael, A formal basis for the heuristic determination of minimal cost paths, IEEE Trans. Syst. Sci. Cybern. 4 (2) (1968) 100–107.
[28] J. Borenstein, Y. Koren, The vector field histogram-fast obstacle avoidance for mobile robots, IEEE Trans. Robot. Autom. 7 (3) (1991) 278–288.
[29] I. Ulrich, J. Borenstein, VFH*: local obstacle avoidance with look-ahead verification, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA, San Francisco, CA, USA, 2000, pp. 2505–2511.
[30] D. Fox, W. Burgard, S. Thrun, The dynamic window approach to collision avoidance, IEEE Robot. Autom. Mag. 4 (1) (1997) 23–33.
[31] J. Elfring, S. van den Dries, M. van de Molengraft, M. Steinbuch, Semantic world modeling using probabilistic multiple hypothesis anchoring, Robot. Auton. Syst. 61 (2) (2013) 95–105.
[32] F. Siepmann, L. Ziegler, M. Kortkamp, S. Wachsmuth, Deploying a modeling framework for reusable robot behavior to enable informed strategies for domestic service robots, Robot. Auton. Syst.
[33] J.A. Álvarez Ruiz, P.-G. Plöger, G.K. Kraetzschmar, Active scene text recognition for a domestic service robot, in: RoboCup 2012: Robot Soccer World Cup XVI, 2013, pp. 249–260.
[34] G. Giorgana, P.G. Ploeger, Facial expression recognition for domestic service robots, in: RoboCup 2011: Robot Soccer World Cup XV, 2012, pp. 353–364.
[35] M. Correa, J. Ruiz-del-Solar, F. Bernuy, Face recognition for human–robot interaction applications: a comparative study, in: RoboCup 2008: Robot Soccer World Cup XII, 2009, pp. 473–484.
[36] G. Hermosilla, P. Loncomilla, J. Ruiz-del-Solar, Thermal face recognition using local interest points and descriptors for HRI applications, in: RoboCup 2010: Robot Soccer World Cup XIV, 2011, pp. 25–35.
[37] M. Correa, G. Hermosilla, R. Verschae, J. Ruiz-del-Solar, Human detection and identification by robots using thermal and visual information in domestic environments, J. Intell. Robot. Syst. 66 (1–2) (2012) 223–243.
[38] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[39] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.
[40] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2006, pp. 430–443.
[41] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: binary robust independent elementary features, in: Proceedings of the European Conference on Computer Vision, ECCV, 2010, pp. 778–792.
[42] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, B. Girod, Rotation-invariant fast features for large-scale recognition and real-time tracking, Signal Process. Image Commun. 28 (4) (2013) 334–344.
[43] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (FPFH) for 3D registration, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA, 2009, pp. 1848–1853.

[44] R. Rusu, G. Bradski, R. Thibaux, J. Hsu, Fast 3D recognition and pose using the viewpoint feature histogram, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2010, pp. 2155–2162.
[45] F. Tombari, S. Salti, L. Di Stefano, Unique signatures of histograms for local surface description, in: Proceedings of the European Conference on Computer Vision, ECCV, 2010, pp. 356–369.
[46] B. Drost, M. Ulrich, N. Navab, S. Ilic, Model globally, match locally: efficient and robust 3D object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2010, pp. 998–1005.
[47] C. Papazov, D. Burschka, An efficient RANSAC for 3D object recognition in noisy and occluded scenes, in: Proceedings of the Asian Conference on Computer Vision, ACCV, 2011, pp. 135–148.
[48] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R.B. Rusu, S. Gedikli, M. Vincze, Tutorial: Point Cloud Library: three-dimensional object recognition and 6 DOF pose estimation, IEEE Robot. Autom. Mag. 19 (3) (2012) 80–91.
[49] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, V. Lepetit, Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes, in: IEEE International Conference on Computer Vision, ICCV, 2011.
[50] J. Stückler, D. Droeschel, K. Gräve, D. Holz, J. Kläß, M. Schreiber, R. Steffens, S. Behnke, Towards robust mobility, flexible object manipulation, and intuitive multimodal interaction for domestic service robots, in: RoboCup 2011: Robot Soccer World Cup XV, in: LNCS, vol. 7416, Springer, 2012, pp. 51–62.
[51] J. Stückler, D. Droeschel, K. Gräve, D. Holz, M. Schreiber, A. Topalidou-Kyniazopoulou, M. Schwarz, S. Behnke, Increasing flexibility of mobile manipulation and intuitive human–robot interaction in robocup@home, in: RoboCup 2013: Robot Soccer World Cup XVII, Springer, 2014.
[52] J. Elfring, S. Jansen, R. van de Molengraft, M. Steinbuch, Active object search exploiting probabilistic object–object relations, in: RoboCup 2013: Robot Soccer World Cup XVII, 2014.
[53] D. Holz, S. Holzer, R.B. Rusu, S. Behnke, Real-time plane segmentation using RGB-D cameras, in: RoboCup 2011: Robot Soccer World Cup XV, in: Lecture Notes in Computer Science, vol. 7416, Springer, 2012, pp. 307–317.
[54] J. Stückler, R. Steffens, D. Holz, S. Behnke, Efficient 3D object perception and grasp planning for mobile manipulation in domestic environments, Robot. Auton. Syst. 61 (10) (2013) 1106–1115.
[55] T. Breuer, G.R.G. Macedo, R. Hartanto, N. Hochgeschwender, D. Holz, F. Hegger, Z. Jin, C. Müller, J. Paulus, M. Reckhaus, J.A.Á. Ruiz, P.-G. Plöger, G.K. Kraetzschmar, Johnny: an autonomous service robot for domestic environments, in: Special Issue on Domestic Service Robots, J. Intell. Robot. Syst. 66 (1–2) (2012) 245–272.
[56] R. Bischoff, J. Kurth, G. Schreiber, R. Koeppe, A. Albu-Schaeffer, A. Beyer, O. Eiberger, S. Haddadin, A. Stemmer, G. Grunwald, G. Hirzinger, The KUKA-DLR lightweight robot arm—a new reference platform for robotics research and manufacturing, in: Proceedings of the 41st International Symposium on Robotics (ISR) and 6th German Conference on Robotics, ROBOTIK, 2010.
[57] U. Reiser, T. Jacobs, G. Arbeiter, C. Parlitz, K. Dautenhahn, Care-O-bot®3—vision of a robot butler, in: R. Trappl (Ed.), Your Virtual Butler, in: Lecture Notes in Computer Science, vol. 7407, Springer, Berlin/Heidelberg, 2013, pp. 97–116.
[58] J. Stückler, I. Badami, D. Droeschel, K. Gräve, D. Holz, M. McElhone, M. Nieuwenhuisen, M. Schreiber, M. Schwarz, S. Behnke, NimbRo@Home: winning team of the RoboCup@Home competition 2012, in: RoboCup 2012: Robot Soccer World Cup XVI, in: LNCS, vol. 7500, Springer, 2013, pp. 94–105.
[59] X. Chen, J. Xie, J. Ji, Z. Sui, Toward open knowledge enabling for human–robot interaction, J. Hum.-Robot Interact. 1 (2) (2012) 100–117.
[60] T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, T. Toda, H. Okada, T. Omori, Learning novel objects for extended mobile manipulation, J. Intell. Robot. Syst. 66 (2012) 187–204.
[61] M. Fujimoto, S. Nakamura, Sequential non-stationary noise tracking using particle filtering with switching dynamical system, in: Proceeding of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, 2006, pp. 769–772.
[62] T. Breuer, G.R. Giorgana Macedo, R. Hartanto, N. Hochgeschwender, D. Holz, F. Hegger, Z. Jin, C. Müller, J. Paulus, M. Reckhaus, J.A. Álvarez Ruiz, P.-G. Plöger, G.K. Kraetzschmar, Johnny: an autonomous service robot for domestic environments, J. Intell. Robot. Syst. 66 (1–2) (2012) 245–272.
[63] D. Droeschel, J. Stückler, S. Behnke, Learning to interpret pointing gestures with a time-of-flight camera, in: Proceedings of the 6th International Conference on Human–Robot Interaction, HRI '11, ACM, New York, NY, USA, 2011, pp. 481–488.
[64] M. Correa, J. Ruiz-del-Solar, R. Verschae, J. Lee-Ferng, N. Castillo, Real-time hand gesture recognition for human robot interaction, in: RoboCup 2009: Robot Soccer World Cup XIII, 2010, pp. 46–57.
[65] X. Chen, J. Ji, J. Jiang, G. Jin, F. Wang, J. Xie, Developing high-level cognitive functions for service robot, in: Proc. of the 9th Int. Conf. on Autonomous Agents and Multi-agent Systems, AAMAS, 2010.
[66] S. Schiffer, A. Ferrein, G. Lakemeyer, Reasoning with qualitative positional information for domestic domains in the situation calculus, in: Special Issue on Domestic Service Robots in the Real World, J. Intell. Robot. Syst. 66 (1–2) (2012) 273–300, http://dx.doi.org/10.1007/s10846-011-9606-0.