RoboCup@Home: Scientific Competition and Benchmarking for Domestic Service Robots

Thomas Wisspeintner
Department of Mathematics and Computer Science
Freie Universität Berlin
Takustr. 9, 14195 Berlin, Germany
thomas.wisspeintner@gmail.com

Tijn van der Zant
Department of Artificial Intelligence
University of Groningen
Nijenborgh 9, 9747AG
Groningen, The Netherlands
tijn@ai.rug.nl

Luca Iocchi
Dipartimento di Informatica e Sistemistica
Sapienza University of Roma
Via Ariosto 25, Roma 00185, Italy
iocchi@dis.uniroma1.it

Stefan Schiffer
Knowledge-Based Systems Group
RWTH Aachen University
Ahornstr. 55, 52056 Aachen, Germany
schiffer@cs.rwth-aachen.de

May 19, 2009

Abstract

Being part of the RoboCup initiative, the ROBOCUP@HOME league targets the development and deployment of autonomous service and assistive robot technology being essential for future personal domestic applications. The domain of domestic service and assistive robotics implicates a wide range of possible problems. The primary reasons for this include the large amount of uncertainty in the dynamic and non-standardized environments of the real world, and the related human interaction. Furthermore, the application orientation requires a large effort towards high level integration combined with a demand for general robustness of the systems. This article details the need for interdisciplinary community effort to iteratively identify related problems, to define benchmarks, to test and, finally, to solve the problems. The concepts and the implementation of the ROBOCUP@HOME initiative as a combination of scientific exchange and competition is presented as an efficient method to accelerate and focus technological and scientific progress in the domain of domestic service robots. Finally, the progress in terms of performance increase in the benchmarks and technological advancements is evaluated and discussed.

Keywords Domestic Service Robotics, Application, Uncertainty, Benchmark, Competition, Human-Robot Interaction, ROBOCUP@HOME

1 Introduction

The general idea of personal Domestic Service Robotics (DSR) has been around for a long time, but it is a comparably young research topic. The aim of creating useful, autonomous, multipurpose personal assistant robots which can interact with humans and objects in the real world in a natural way poses a large number of unsolved problems across many scientific disciplines.

There have been many successful and impressive demonstrations of robot technology in the past. In DSR, one focus-and one of the main difficulties-is the interaction with the real world, instead of operating under constrained settings and strictly defined environmental conditions as opposed to e.g. industrial robotics. DSR systems must cope with a large amount of uncertainty. A natural home environment, for example, is not specified in size, shape, appearance, the kind of objects contained in it, lighting and acoustic conditions, the kind and number of residents, etc. Furthermore, as objects and people can move, disappear and reappear, the environment is dynamic. The system must be able to manipulate objects in various locations and from different heights, and it needs to be capable of locomotion on different terrains. When interacting with humans, the system should possess some basic (social) intelligence and should be able to distinguish different people. Last but not least, safe and robust operation of these systems in such uncertain and dynamic environments is a fundamental requirement for their future acceptance and general applicability.

The creation of such autonomous systems requires the integration of a large set of abilities and technologies. Examples include human-robot interaction (speech, gesture, person, face recognition and person tracking, among others), navigation and mapping, reasoning, planning, behavior control, object recognition, object manipulation or tracking of objects. With regard to artificial intelligence, the systems should contain adaptive but robust behavior and planning methods, social intelligence, and learning capabilities. Intuitive programming methods (instead of entering computer code) are required for a broad acceptance and usability. Appropriate procedures should, for instance, enable the robot operator to teach new behaviors and environments via voice or gesture commands. As future households will most likely contain more intelligent electronic devices capable of communicating with each other, ambient intelligence, including the use of the Internet as a common knowledge base, will certainly play a more important role.

Just very recently, progress in these research fields, as well as progress and standardization in related hardware and software development, has led to an increase in availability of required methods and components for DSR. This includes the availability of software frameworks for robot control (e.g. Carmen[1] [Montemerlo et al., 2003], Player/Stage[2] [Gerkey et al., 2003], MRPT[3], MRS[4]), simulation (e.g. USARSim[5] [Balakirsky, 2006]), and open source software libraries containing algorithms for computer vision (e.g. OpenCV[6] with diverse applications as shown in [Bradski and Pisarevsky, 2000]) or robot control (e.g. Orocos[7] [Bruyninckx, 2001]). On

---

[1]Carnegie Mellon Robot Navigation Toolkit (http://carmen.sourceforge.net/)

[2]The Player/Stage Project (http://playerstage.sourceforge.net/)

[3]The Mobile Robot Programming Toolkit (http://babel.isa.uma.es/mrpt/index.php/Main_Page)

[4]Microsoft Robotics Studio (http://msdn.microsoft.com/en-us/robotics/default.aspx)

[5]Unified System for Automation and Robot Simulation (http://sourceforge.net/projects/usarsim)

[6]The Open Computer Vision Library (http://sourceforge.net/projects/opencv/)

[7]Open Robot Control Software (http://www.orocos.org/)

the hardware side, robot construction kits (e.g. VolksBot[8] [Wisspeintner and Nowak, 2007] and base platforms (e.g. ActivRobots[9]), faster and energy efficient computation or light weight manipulation devices (e.g. Katana[10]) as well as miniature sensors (e.g LIDAR[11]) are available.

In sum, increased availability, accessibility and compatibility of these essential robot components enables research groups not only to address a small subset of the mentioned above challenges in DSR, but also to address the problem as a whole. Obviously, DSR is not solely about integrating existing solutions. But the consequent reuse of existing technology can help to save time and effort, so researchers can focus on a particular research field while maintaining a fully operable robot platform.

This is also confirmed by the presence of some rather specialized service robotic applications on the market. Such applications include floor cleaning (e.g Roomba and Scooba[12]), lawn mowing (e.g. Robomow[13]) and surveillance (e.g. Robowatch[14]). Still, these service robots do not possess the properties of a multipurpose autonomous and intelligent domestic service robot. Prominent examples of domestic and personal assistant robot research projects include Ready-Bot[15], and PR2[16]. Wakamaru [17] and PaPeRo[18] focus more on social interaction studies. Many of these projects address relevant aspects of DSR. Still, what appears to be missing is a joint, international and multidisciplinary research and development effort which also includes the aspect of application-oriented benchmarking of systems in DSR.

With this motivation, the authors initiated the ROBOCUP@HOME competitions in 2005 [van der Zant and Wisspei van der Zant and Wisspeintner, 2007]. The ROBOCUP@HOME league targets the development and deployment of autonomous service and assistive robot technology as being essential for future personal domestic applications. It is part of the international RoboCup initiative, and it is the largest annual service and home robotic competition worldwide. The ROBOCUP@HOME tournaments are organized in independent test sets, which are used to benchmark the robots' abilities and performance in a realistic non-standardized home environment. More specifically, ROBOCUP-@HOME aims to offer a combination of interdisciplinary community building, scientific exchange and competition, which iteratively defines benchmarks and performance metrics on which service robots can be evaluated and compared in a realistic, dynamic and non-standardized domestic environment.

Since the real world is not standardized, measuring the performance of non standardized robots acting in it is a difficult task. The experimental paradigm to evaluate the complex robotic systems has to use consequent scientific analysis to improve on itself. Measuring the performance of the robots requires continuous reconsideration of the methodologies used since both the robots (their capabilities) and their operation environment (and the robot's tasks) will definitely change over time. This co-evolutionary development process, the feedback and refinement procedure, is a key

---

[8]VolksBot Robot (http://www.volksbot.de/index-en.php)

[9]ActivRobots (http://www.activrobots.com)

[10]Katana robotic arm (http://www.neuronics.ch/)

[11]Hokuyo Sensor Technology (http://www.hokuyo-aut.jp/)

[12]iRobot (http://irobot.com)

[13]Robomow (http://www.friendlyrobotics.com)

[14]Robowatch (http://robowatch.com)

[15]ReadyBot (http://www.readybot.com/)

[16]PR2 (http://www.willowgarage.com/)

[17]Wakamaru (http://www.mhi.co.jp/kobe/wakamaru/english/)

[18]PaPeRo (http://www.nec.co.jp/robot/english/robotcenter_e.html)

element of the ROBOCUP@HOME league. In our case, the tools are statistical benchmarks which test certain robot abilities and the measurement of the robots' performance.

ROBOCUP@HOME also measures, in a scientific and quantifiable manner, the performance of complex systems. We firmly believe that creating and applying this experimental paradigm can greatly improve DSR developments.

This article thus addresses the problem of benchmarking DSR through scientific competitions by presenting the approach followed in the ROBOCUP@HOME initiative. The article contains several contributions:

- it presents an overview of benchmarking through competitions, describing other existing competitions and highlighting the unique features of ROBOCUP@HOME;

- it describes the underlying concept of the @Home competition and its implementation into a framework for benchmarking in DSR which aims to be a common testbed for application development;

- it provides a detailed analysis of the results from different viewpoints that are of crucial importance for assessing the actual performance of DSR and for planning future tests and other competitions.

The remainder of the article is organized as follows: The next section gives an overview of the state of the art in robotic benchmarking and DSR. Then, the concept and the implementation of the @HOME competition are presented. Section 4 will evaluate the benchmarking results of the past several years and discuss the observed increase in performance, the scientific achievements and the importance of a vital community. The article concludes with an outlook on short and mid-term goals.

2 Benchmarking Domestic Service Robotics

Benchmarking has been recognized as a fundamental activity to advance robotic technology [del Pobil, 2006, Sabanovic et al., 2006], and many activities are in progress. Some projects and special groups are working on defining standard benchmarking methodologies and data sets for many robotic problems, like Human-Robot interaction (HRI), SLAM, or navigation. Examples for such initiatives are the the EURON Benchmarking Initiative[19], the EURON guidelines on good experimental methodologies and benchmarking[20] , the international workshops on Benchmarks in Robotics Research and on Performance Evaluation and Benchmarking for Intelligent Robots and Systems, held since 2006[21], the Rawseeds project[22], which aims to create standard benchmarks especially for localization and mapping, and the RoSta project[23], which focuses on standardization and reference architectures.

Benchmarking can be distinguished in two classes: *system benchmarking*, where the robotic system is evaluated as a whole, and *component benchmarking*, where single functionality is evaluated. Component benchmarking is integral for comparing different solutions to a specific problem

---

[19]http://www.euron.org/activities/benchmarks/index

[20]http://www.heronrobots.com/EuronGEMSig/Downloads/GemSigGuidelinesBeta.pdf

[21]All these workshops are summarized in http://www.robot.uji.es/EURON/en/index.htm.

[22]http://www.rawseeds.org/

[23]http://www.robot-standards.eu/

and for identifying the best algorithms and approaches. Among the many examples, much effort has been put on mapping and SLAM (e.g. [Howard and Roy, 2003, Fontana et al., 2008]), and navigation (e.g. [Baltes, 2000, Munoz et al., 2007, Calisi et al., 2008]). While component benchmarking is useful for directly comparing different techniques of solving a specific problem, it is not sufficient for assessing the general performance of a robot with respect to a class of applications. Indeed, the best solution for a specific problem may be unfeasible or inconvenient when integrated with other components that compose a robotic application. On the other hand, system benchmarking offers an effective way to measure the performance of an entire robotic system in the accomplishment of complex tasks, as such tasks require the cooperation of various sub-systems or approaches.

In this kind of benchmarking, a standard reference environment, reference tasks and related performance metrics are to be defined. Examples of system benchmarking are given in the fields of interactive robots [Kahn et al., 2007] and of socially assistive robots [Feil-Seifer et al., 2007].

When defining standard benchmarks, two common problems arise:

- The difficulty of defining a benchmark that is commonly accepted by the community (this is due to differing viewpoints on a problem from separate research groups);

- The risk of fostering the development of specialized solutions for an abstracted, standardized setting.

To avoid these problems, scientific competitions have proven to be a very adequate method because:

- benchmarks are usually discussed and then accepted by all the participants;

- participants are usually required to solve multiple benchmarks. These benchmarks vary over the years, thus providing for a disadvantage in using solutions that are too specialized.

Moreover, competitions provide an effective means of interaction and communication among research groups because they are often associated with scientific conferences or workshops and provide participants a large audience for their research efforts. Finally, annual competitions provide regular feedback on performance increases and allow for establishing medium-term projects.

Among the many robotic competitions, the AAAI Mobile Robot Competitions were one of the first, being established in 1992 [Balch and Yanco, 2002]. RoboCup (founded in 1997) [Kitano et al., 1997] currently has the largest number of participants (e.g. 440 teams with more than 2,600 participants from 35 countries in 2006). The DARPA Grand Challenge is probably the most recognized in terms of public and media attention and the one that is most directly application-oriented.

Furthermore, educational contests, such as EUROBOT[24] or RoboCup Junior[25], are organized with the main goal of presenting robotics to young students. Thus, they deal with simpler tasks and robotic platforms.

All of these competitions have obtained very relevant results, which are analyzed in the following:

AAAI AAAI Mobile Robot Competitions are held in conjunction with the AAAI and (sometimes) IJCAI Conferences on Artificial Intelligence. Thus, it offers great visibility within the AI scientific community. Many important scientific and technological achievements demonstrated

---

[24]http://www.eurobot.org/

[25]http://rcj.sci.brooklyn.cuny.edu/

during these competitions have been reported [Balch and Yanco, 2002]. Although these competitions offer a relevant suite for benchmarking AI and robotics technology with relevance for real-life applications, their focus and benchmarks change heavily on a yearly basis. This change of focus makes it difficult to approach the problems in a continuous and iterative way and to build up a community with a long-term goal.

RoboCup Soccer The ultimate goal of the RoboCup project is: "By the year 2050, develop a team of fully autonomous humanoid robots that can play and win against the human world champion soccer team." Moreover, as opposed to AAAI competitions, RoboCup events put the main focus on the competition and offer the possibility to discuss scientific achievements in a small and more focused RoboCup Symposium. RoboCup has proven to provide an efficient means of interaction and communication among research groups. It combines scientific research, competition, benchmarking and reality checks on various concepts. Performance is measured on a yearly basis. However, having a specific focus on soccer also presents some limitations. The main limitation is an over-specialization of solutions due to more or less fixed environmental conditions and rules. For example, in the middle-size league, where the design of the robot is a major issue, all the teams rapidly converged towards the same hardware architecture (catadioptric cameras and omnidirectional driving robots) which was highly optimized on the provided scenario. Although this causes an immediate improvement of the average performance in the competition, it contains the danger of running into a single suboptimal solution which can not be applied to a real-world setting.

Robot Rescue Another example is given by Search and Rescue Robotics. Rescue competitions started in 2000 within the AAAI Mobile Robot Competition [Meeden et al., 2000] and since 2001 within the RoboCup Rescue initiative [Kitano and Tadokoro, 2001]. RoboCup Rescue competitions have defined standard rescue arenas and tasks for benchmarking robotic search and rescue missions and for measuring an increase in performance of the rescue robotic technology in a standardized abstracted environment. The concepts of the rescue robot initiative with respect to benchmarking are similar to those proposed in @Home. Within the Rescue competitions, common metrics for HRI have been defined [Steinfeld et al., 2006] and effective evaluation of HRI techniques have been carried out [Yanco et al., 2004, Drury et al., 2005], with a specific focus on the interfaces used by operators to interact remotely with the rescue robots. Nevertheless, this indirect kind of HRI via an operator station involving semi-autonomy and remote control is different to what is required in most DSR tasks, where the focus is a direct and more natural interaction and full autonomy. Still, one can think of certain DSR applications where such kind interaction is desired, e.g. to monitor and communicate with nursing cases remotely.

DARPA Grand Challenge The DARPA Grand Challenges[26] were the most prominent robotic competitions to date. GPS navigation together with multimodal sensor data fusion were commonly used to face the uncertainties and dynamics of real-world application scenarios. In the Urban Challenge, even real traffic rules were applied, but the complexity was limited by simplifying the cognition tasks. Contextual information was entered into a predefined map, the route network definition file (RNDF), which consisted of waypoints with GPS coordinates, connection types,

---

[26]http://www.darpa.mil/grandchallenge/index.asp

traffic signs, number of lanes, width of lanes, etc. Participating in these challenges required a lot of effort, as the joint work of different research groups and industries with complementary competencies was a critical factor. At this time, it is uncertain if this initiative will be continued in the future and to what capacity.

Mobile Manipulation Challenge   A recent initiative is the mobile manipulation challenge[27] , which is being held as part of the IJCAI 2009 robotics exhibition and workshop. In its first edition the emphasis is "on demonstration, rather than comparative experimentation". The aim is to bring together researchers and commercial groups demonstrating complete platforms that are able to perform mobile manipulation tasks.

Service Robotic Competitions   Initiatives that are directly related to Domestic Service Robotics mainly aim at a single specific task. For example, the AHRC Vacuum Contest[28] and the 2002 IROS Cleaning Contest[29] [Prassler, Erwin and Hägele, Martin and Siegwart, Roland, 2006] are focused only on floor cleaning, while ROBOEXOTICA[30] concentrates only on robots preparing and serving cocktails. A more general initiative is given by the ICRA HRI Challenge[31], and it is motivated by the fact that "the effectiveness of a robot engaging in HRI must be evaluated by human users who got the chance to interact with the robot for a sufficiently long period of time." However, it is still in preliminary stages because evaluation criteria for benchmarking the performance have not been defined.

Although these initiatives are very relevant to the field, it is evident that there is no major international annual competition in the field of Domestic Service Robotics that can be considered a continuous integrated system benchmarking activity. A full discussion of the above and all other benchmarking initiatives would surely go beyond the scope of this paper. What is important though, is that most other benchmarks either use a standardized environment or only test a subset of individual abilities.

ROBOCUP@HOME is an effort to compare and evaluate integrated, application oriented, systems by means of a competition. The focus on application in combination with the aim of creating multipurpose robots requires integration and testing of many abilities. The aim on testing HRI in natural non-standardized environments makes it hard to maintain precisely predefined conditions while evaluating. In @Home, each test assesses a set of abilities and each ability is rewarded with a predefined amount of points. The focus on benchmarking the entire system is also guaranteed by the sum of scores obtained in several tests. Finally, the advantage of keeping test conditions constant for all systems, is a high grade of comparability between the systems. On the other hand, the danger of over-fitting to abstracted test conditions rises with this approach.

---

[27]http://www.mobile-manipulation-challenge.net/

[28]http://www.botlanta.org/

[29]http://robotika.cz/competitions/cleaning2002/en

[30]http://www.roboexotica.org/en/mainentry.htm

[31]http://lasa.epfl.ch/icra08/hric.php

3 The @HOME initiative

With the insights and experiences gained from participating in and from observing and organizing various robotic competitions, we started with the conceptual design of a new competition for DSR in 2005.

3.1 Concept

The following considerations and criteria act as the basis of a common agreement for the ROBO-CUP@HOME initiative.

Uncertainty   To reflect the uncertainty immanent in every real-world setting, the rules should not specify or limit any more qualities of a task than (absolutely) necessary. This complies with the requirement of providing a lean set of rules. Moreover, it encourages robust solutions that remain functional over a wide number of particular situations under as many circumstances as possible. This way, object positions or environmental characteristics, such as lighting conditions, are not specified and the scenario setup is changed frequently.

Extendable framework for benchmarking   With the aim of benchmarking myriad robot capabilities for DSR with many of them yet to be developed, the framework for the competition needs to allow for constant evolution and modular enhancement of itself. The framework will consist of an initial set of independent tests all benchmarking an individual set of relevant capabilities in DSR. Over time, when an increase in performance in the individual tests is observed, these tests are either enhanced by making the tasks more difficult or the tasks are merged together to form a more integrated, and therefore more realistic, application scenario.

Autonomy   Robots in the @HOME league are required to be fully autonomous and mobile. That is, robots must complete tests without being controlled remotely. To lower the demands for on-board computers, external computation is permitted as long as nobody interacts with the external computer during the test. To foster demonstration and use of new approaches, external devices the robot can interact with (external cameras, sensors, etc.) are allowed in certain tests. It may appear that instructions given by a human acting with the robot are a kind of tele-operation, but the execution of a given high-level task such as "Bring me object A" incorporates autonomy in terms of task decomposition, decision making, perception, task planning, and task execution. We enforce this autonomy by the uncertainty inherent in the environment and strict time limits for the setup of the robots.

Natural interaction   In order to inhibit control of a robot by keyboard commands, interaction with a robot must be *natural* in all tests. This means that the interaction is either done via natural language or gestures (no keyboard control). Other modes of interaction like the use of touch screens or advanced remote controls can and should be demonstrated in the form of technical challenges (see *Open Challenge*, *Demo Challenge* and *Finals* in Section 3.3), where these restrictions are not applied. Then, corresponding solutions are to be integrated and allowed in future competitions. Moreover, haptic interaction (touching the robot) should further foster development of

intuitive modes of control and interaction (instead of using a standard computer keyboard) and consider future use by the target audience: the general public, laymen in robotics, or elderly and disabled people.

Benchmarking in uncertain conditions   The home environment in which the benchmarks take place is not standardized to represent a realistic setting. It contains common natural objects and varies over the years. Examples of previous competition environments are given in Figure 1. The degree of uncertainty contained in the benchmarks is high as the environment is hardly specified in size, shape, contained objects, kind of walls or floor, lighting and sound conditions, etc. and changes from competition to competition. Especially the interaction with randomly selected people adds to this uncertainty. The non-standardized conditions under which Human-Robot-Interaction is tested include different persons of different size, different nationalities, different accents and voices, different gender, different clothes, different sound conditions, background noise, and demand for distinguishing between active persons and the many spectators. Still, it is of importance to maintain a certain range of difficulty for all participants while boundary conditions may slightly change, so that the performance can be compared. This requires a common agreement and careful definition on the level of uncertainty. To solve the discrepancy between desired uncertainty and comparability of performance measures, statistical benchmarking is used as a method. As the competition consists of multiple tests and these tests consist of multiple tasks which are evaluated, we collect more than fifty data points per finalist on about a dozen abilities which allows comparison of the teams performances. Every team faces the same variability in the environment conditions and using multiple samples per team per ability allows for a statistical performance analysis.

Fostering a wide range of approaches and solutions   The rules should be kept as unrestrictive as possible, and the benchmarks are to be defined in such way that the solutions for the given problems are not implicit. This approach requires a high level of common sense and agreement from the teams and the community, as trivial and undesired solutions to certain problems can not be completely avoided, e.g., having a robot approaching an object with an open-loop control instead of using sensor feedback to adjust to position changes. Also, participants should have the choice to select certain benchmarks according to their background, skills, and their robot's capabilities. Besides having predefined benchmarks the teams can select from, the competition will also offer the possibility of demonstrating new abilities and scientific results or applications not yet covered in the tests. These new aspects can later be used to enhance the benchmarks in the future.

Multidisciplinary community   Putting few restrictions on the robots participating and providing the freedom to select benchmarks and approaches should motivate teams from different research backgrounds to participate in and to contribute to a growing community, one which fosters the exchange of multidisciplinary scientific and technological knowledge. Furthermore, the development of a common vision about the goals, as well as common sense and fair play in the competition, are required. Feedback from the teams is further needed to iteratively develop the competition.

Generating public awareness   The competition should also generate interest from a non-technical, public audience by demonstrating usefulness in daily life, future applications and social relevance.

This way public awareness for DSR should be increased, and links to the industry should be established.

Figure 1: The @HOME scenario in 2006 (top left), 2007 (top right), and 2008 (bottom)

3.2 Defining key features

Before starting with the implementation of the benchmarks, an initial set of robot key features (abilities and properties) was derived from an analysis of the state of the art in DSR and our experiences and observations from other robotic competitions. These features help to design the benchmarks and the score system for the competition. Furthermore, these features allow for a later analysis of the teams' performances and help to develop and later enhance the competition in a structured way. As the competition and its benchmarks are expected to evolve over time, the key features and their weights in the competition are also expected to be adapted. The key features are divided in two groups: *functional abilities* and *system properties*.

3.2.1 Functional abilities

Functional abilities include specific functionality that must be implemented on the robot in order to perform decently in the tests. Each test requires a certain subset of these abilities, as they are also directly represented in the score system. Teams must thus decide which of these abilities to implement and what degree of performance to achieve, depending on their background and the kind of tests they intend to participate in. *Functional abilities* currently are:

- *Navigation*, the task of path planning and safely navigating to a specific target position in the environment, avoiding (dynamic) obstacles
- *Mapping*, the task of autonomously building a representation of a partially known or unknown environment on-line
- *Person recognition*, the task of detecting and recognizing a person
- *Person tracking*, the task of tracking the position of a person over time
- *Object recognition*, the task detecting and recognizing (known or unknown) objects in the environment
- *Object manipulation*, the task of grasping or moving an object
- *Speech recognition*, the task of recognizing and interpreting spoken user commands (speaker dependent and speaker independent)
- *Gesture recognition*, the task of recognizing and interpreting human gestures.

3.2.2   System properties

System properties include demands on the entire robotic system that we consider of general importance for any domestic service robot. They can be described as "soft skills" which must be implemented for effective system integration and successful participation in the @HOME competition. Initial system abilities are:

- *Ease of use* - Laymen should be able to operate the system intuitively and within little amount of time.
- *Fast calibration and setup* - Simple and efficient setup and calibration procedures for the system.
- *Natural and multimodal interaction* - Using natural modes of communication and interaction such as, e.g. using natural language, gestures or intuitive input devices like touch screens.
- *Appeal and ergonomics* - General appearance, quality of movement, speech, articulation or HRI.
- *Adaptivity / General intelligence* - Dealing with uncertainty, problem solving, online learning, planning, reasoning.
- *Robustness* - System stability and fault tolerance.
- *General applicability* - Solving a multitude of different realistic tasks.

Although some of these properties cannot be benchmarked as directly as the *functional abilities*, they are considered as integral and an implicit part of the competition and tests.

3.3   Implementation of benchmarks

In the following, we are going to elaborate on how we implemented the ROBOCUP@HOME competition as a set of benchmark tests for service robots in domestic environments. This implementation is based on the concepts mentioned in the previous section.

The competition is organized in a multi-stage system. All qualified teams (currently up to 24) participate in the first stage called *Stage I*. It consists of a set of benchmarks with a focus on

testing basic tasks and checking for a small set of key features with a limited amount of uncertainty involved. Then, the ten best teams advance to the second stage called *Stage II* where the benchmarks are more demanding, more realistic and involve more uncertainty and a higher level of system integration. In the *Finals*, the performance of the five best teams is evaluated by a jury. A combination of the jury score and the previous score from *Stage I* and *Stage II* determines the ranking.

The tests themselves comprise realistic and useful tasks for a domestic service robot. Each test evaluates certain key features. A tabular overview of the functional abilities required in each test can be found in Table 1. An overview of benchmarks where certain system properties are tested is given in Section 4.1.2. The implementation of the competition described in the following reflects the situation of the competition in 2008.

3.3.1 Score system

Two types of tests exist: regular tests, which are specified in terms of the task to solve and the scoring, and open tests, where teams can either freely choose what to show (the Open Challenge and the Finals), or a topic is given according to which teams can do a demonstration (Demo Challenge). Since the scoring in the open tests is based on an evaluation by a jury, it is partially subjective. However, for every open test there is a list of criteria, which the jury bases its decision on. The criteria will be discussed in the test descriptions below. The scoring in the regular tests mainly reflects the key features mentioned earlier.

To keep the entry level for the competition reasonably low, while still aiming for high top level performance, a so-called *partial score system* was introduced in 2008. With this, a team receives a part of the total score for showing a part of the task's specification. Each of the partial scores is connected to one or more of the functional abilities and/or system properties. This does not only allow for assessing the fulfillment of these features individually, but it is also an incentive for teams to participate in a test even if they know that they cannot solve it completely.

Referees are provided by opponent teams. It is their duty to provide the same difficulty level for all teams in a certain test according to common sense and fair play. This involves, for example, the selection of random people, the definition of paths or the selection and placements of objects. As this is a critical procedure for the entire competition, it is closely monitored by the Technical Committee of the @Home league. Notice that referees in @Home only ensures the proper execution of a task according to the rules, while they do not evaluate the performance nor assign any subjective score.

For each test in *Stage I* a maximum of 1000 points, and in *Stage II* 2000 points, can be scored.

3.3.2 Stage I Tests

The overall theme of tests in *Stage I* is to benchmark essential abilities and properties that any robot in @HOME should exhibit. During the setup days before the competition, a set of ten randomly chosen and previously unknown objects is provided to the teams. A subset of these objects is then used for certain tests.

Introduce   In the *Introduce* test, a robot has to autonomously enter the scenario and move to a position in front of the audience. There, the robot has to introduce itself and the team using speech,

gestures, slides or multimedia. Afterwards, it must leave the scenario on its own. The performance of the robot is evaluated against certain criteria such as smoothness and flow in both movement and presentation, as well as the general appearance of the robot. This test calls for interaction abilities such as speech synthesis, articulation and expression of moods. Figure 5 (right) shows an @HOME robot expressing an emotion. Since there are no regulations on what and how to present, teams are free to show what they think is a useful, attractive and potentially innovative means for a robot to convey information. Team leaders of the competing teams evaluate the performances. This ensures that the teams get to know each other, their background and their robots already at the very beginning of the competition, thus fostering exchange of knowledge in the community. The variety in focus and interest among the different teams also ensures diverse feedback in the evaluation of the performing team.

Fast Follow　A robot's task in the *Fast Follow* test is to follow a person from one entrance of the environment to the other. Two teams compete against each other in this test starting from opposite ends. They need to pass by one another on a common path through the scenario. The most important capabilities evaluated in this test are detection and tracking of humans and safe navigation in a dynamic environment, a task which naturally includes obstacle avoidance. By letting the teams' paths cross one another the robots further need to discriminate their human leader from the opponent's one. Partial scores are awarded for passing a check point at around half of the track, passing the opponent, completing the track, being the fastest team, and not touching any object.

Fetch & Carry　In the *Fetch & Carry* test, the robot has to find and retrieve a certain object which it then needs to return to the human instructor. The robot is instructed to get the object using natural language. Teams are allowed to give the robot a hint on the item's location. Thus, speech recognition and natural language processing are essential to succeed in this test. Human-robot interaction by means of joint activities in common physical space is emphasized, since for the robot to understand the hint, it needs to be capable of interpreting the given spatial description. Partial scores are awarded for understanding the command, finding the object, manipulation of the object, successful delivery, and autonomously leaving the scenario in time.

Who is Who?　The main theme in the *Who is Who?* test is the detection and recognition of people. The robot has to find three persons (two of them unknown to the robot) spread out around an area near to the entrance of the scenario. The robot then needs to find these people, introducing itself to every person found. Each person has to be either identified (if known already) or learned (if unknown). Beside sufficient navigation capabilities this test checks for capabilities in person detection and recognition (mostly face recognition). It also calls for skills in engaging and conducting interaction with humans such as speaker independent speech recognition. Furthermore, basic conversation capabilities are required in order to instruct the unknown person on its behavior during the recognition process. Figure 2 on the left shows an example from the 2007 competition.

Partial scores are awarded for detecting people, discriminating between known and unknown people, learning and recognizing previously unknown people, and autonomously leaving the scenario on time.

Figure 2: Team UT Austin doing Who is Who (left), team eR@sers teaching in an object (right)

Competitive Lost & Found   For the *Competitive Lost & Found* test, two teams compete against each other directly. The assignment is to find and identify as many out of three objects as possible. The referees pick the objects randomly from the set of objects and distribute them randomly throughout the scenario (in a way that the robot actually has a chance of finding the object) just before the test starts. The major abilities tested here are object detection and recognition. An additional focus is put on reliable and fast navigation, since the fastest of the two teams finding all objects receives an extra bonus. Having two robots compete in the same scenario simultaneously accentuates the need for safe navigation in a dynamic environment. Partial scores are awarded for finding an object, identifying an object, leaving the scenario, and being the fastest team.

Open Challenge   In order to allow all participating teams to freely demonstrate their scientific achievements and their unique robot features or capabilities the *Open Challenge* concludes *Stage I*. Here, no restrictions on the kind of performance, kind of devices, or kind of interaction are applied. The Open Challenge is meant as a means of iteratively enhancing the @Home competition by integrating relevant and innovative aspects demonstrated by the teams in future tests. This test consists of a presentation given by the team (an example is depicted in Figure 3) and a demonstration of their robot. The performance is then evaluated by all other teams according to a list of predefined criteria. These criteria are as follows:

Presentation   The quality of the presentation is evaluated as an indication of attractiveness, and to foster scientific exchange between the teams.

Social relevance / Usefulness for daily life   Because RoboCup@Home is about socially relevant robotic applications, this aspect is also evaluated.

Human-Robot interaction   The quality of human-robot interaction in the demonstration is evaluated according to the judging teams' own focus and interests. This cross-evaluation should thus reflect the diverse community and give broad feedback for the performing team.

Autonomy   The amount of autonomy shown by the robot during the demonstration. This is to avoid open loop behavior.

Difficulty and success   The level of difficulty and success of the robot performance is evaluated.

Appeal/Relevance for @Home   How well does the demonstration fit the scope of the @Home

initiative? Could elements of the demonstration be integrated in future @HOME competitions?

Scientific value / Jury questions What is the scientific value of the approach, and how well did the team respond to jury questions?

Figure 3: Presentations from team Pumas(left) and team PAL(right) during the *Open Challenge*

3.3.3 Stage II Tests

In contrast to the challenges in *Stage I*, *Stage II* comprises tests that are more complex, involve more uncertainty, and which check for the integration of several features in a more realistic, application-like setting.

PartyBot The *PartyBot* test is an elaborated version of the *Who is Who?* test from *Stage I* where the robot's task is to find, recognize, and/or remember multiple unknown people randomly distributed throughout the entire environment (standing and sitting) and to tell them apart later on when serving a drink. Besides a focus on interaction capabilities, especially when having to get to know previously unknown persons, navigation, object detection, and manipulation are necessary to pick up and deliver a cup to a particular person. Two robots about to grasp a cup and a bottle are shown in Figure 4. Partial scores are awarded for detecting and navigating to the persons, navigation to the cup, grasping and carrying the cup, handing it over, and autonomously leaving the scenario.

Supermarket To mimic the possible future application of assisted shopping the *Supermarket* test was introduced. The robot needs to retrieve certain household objects from a shelf for a

Figure 4: AllemaniACs (left) and B-IT Bots (right) robots grasping and delivering a drink

person randomly chosen from the audience that does not know how to operate the robot. This demands the robot has the ability to explain its own modes of operation and to report on the robot's internal models to a layman. Furthermore, it requires speaker-independent speech recognition, and it enforces the ease of use proclaimed as a system property, since the operation by laymen raises uncertainty both in input and reaction. The lean specification allows for multimodal input such as gestures and speech. The interactive character further adds a demand for joint work-space concepts to be employed. Handling objects requires both object detection and recognition as well as manipulation abilities. Partial scores are awarded for understanding which object to get, finding the object, retrieving the object, object manipulation on different heights, delivery of the object, and multimodal input (i.e., using speech and gestures).

Walk & Talk   The task of the *Walk & Talk* test is to introduce a robot to a new environment and make it remember a set of places. A human leader guides the robot through the scenario that was completely rearranged beforehand (and is therefore unknown to the robot) and has to teach specific locations only using natural language. The robot then has to prove that it has correctly learned those locations by having to navigate to certain places in random order after a speech command is given. To accomplish this test a robot does not only need to recognize and track a human, but it also has to model and map the so far unknown environment to be able to navigate in it later on. Further, human-robot interaction capabilities such as speech recognition and gesture recognition are an indispensable means to meet the demands posed by the above setting. Partial scores are awarded for following the person to the locations, autonomously navigating to the locations learned previously, navigating back to the start position, and autonomously leaving the scenario.

Cleaning Up In the *Cleaning Up* test, a robot needs to collect a set of five unknown objects (i.e., not from the set of known objects) dispersed throughout the scenario. Objects can be anything that can be expected to lie around in a household. Restrictions are put on the size so that objects are not to small to be overlooked and not too big so they can still be handled by a robot by pushing or grasping. To solve this task, the robot has to find potential objects first (effective search and object detection). Then the robot is expected to test the assumption of having found an object by trying to manipulate it. After having figured out how to handle the objects, the goal is to move them to a predefined area in the scenario. Partial scores are awarded for correct detection of objects, having no false positive in the detection, delivery of objects to a designated area, and autonomously leaving the scenario.
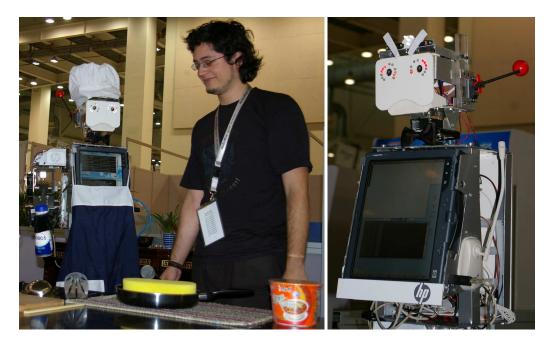

Figure 5: HomeBreakers robot Bender assisting cooking (left) and showing emotions (right)

Demo Challenge The *Demo Challenge* is an open demonstration similar to the *Open Challenge*, as no restrictions on the kind of interaction or the kind of external devices are applied. In contrast to the *Open Challenge* the topic of the demonstration is pre-defined and varies from year to year. It is meant to foster development in a certain area or on a particular theme with a strong relation to real applications and daily-life situations. It should provide a showcase of the current state of the art in home robotics and inspire both the community and the public. In 2008 the theme was "cooking", i.e., the robot should assist a human in preparing a meal. The task was not formulated in any concrete specification. Possible means to assist were, for example, fetching a recipe from the Internet and retrieving ingredients necessary for the same. Figure 5 (left) shows a robot participating in the 2008 demo challenge. Evaluation was done by a jury consisting of the organizers of the @Home competition. Evaluation criteria in 2008 were: assisting and interacting with the human, ambient intelligence and object manipulation.

3.3.4 Finals

The competition concludes with the *Finals*, where as in the *Open Challenge*, each team can demonstrate what they think is an important feature or capability of their robot. The idea, however, is to present a coherent story-like performance which is evaluated by an external jury according to a list of predefined criteria. Because teams that have reached the Finals have already proven to fulfill a variety of abilities, the criteria of the evaluation are slightly different from those in the *Open Challenge*.

Scientific contribution / Contribution to the community Amount, relevance and quality of the team's contribution to the @Home community

Relevance for RoboCup@Home/ Usefulness for daily life of the demonstration

Usability / Human-robot interaction and multimodality Ease of use, quality of HRI and multimodality during the demonstration

Originality and presentation Originality of the demonstration, quality of the presentation

Difficulty and success of the demonstration

Previous performance during *Stage I* and *Stage II* (determined by previous score)

4 Evaluation and discussion

Two important objectives for an annual scientific competition are to provide a common benchmark to many teams, which allows for the measurement of performance advances over time, and to develop relevant scientific solutions and results. In this section we describe and discuss the results obtained by the RoboCup@Home teams both in terms of performance in the tests and in terms of scientific achievements.

As for a team's performance, it is important to note that the score system of RoboCup-@Home relates the desired abilities of the robots with the scores of the competition. In contrast to other competitions (e.g., RoboCup soccer), where the score hides many factors, the @Home score provides an actual way of measuring the performance of teams in terms of such abilities. This score consequently enables an analysis of performance in order to update the rules and drive technological and scientific progress.

In the remainder of this section, first, we will present an analysis of 2008 team performance based on the relationship between key features and test scores; second, we will discuss the evolution of the league over time; then, we will highlight the teams' main scientific contributions related to @Home; and finally, we will discuss results from the @Home community.

4.1 Representation of key features in the benchmarks

In the following the representation of key features, i.e. the functional abilities as well as the system properties, in the benchmarks and in the competition score are shown.

4.1.1 Functional abilities

Table 1 relates the *functional abilities* defined in Section 3.2 with the tests described above. It quantifies the maximum score distribution per test with respect to the contained functional abilities. For ease of notation, The following abbreviations are used. Tests include Fast Follow (FF),

Fetch & Carry (FC), Who is Who (WW), Lost & Found (LF), PartyBot (PB), Supermarket (SM), Walk & Talk (WT), and Cleaning Up (CL). The abilities are Navigation (Nav), Mapping (Map), Person Recognition (PRec), Person Tracking (PTrk), Object Recognition (ORec), Object Manipulation (OMan), Speech Recognition (SRec), and Gesture Recognition (GRec). Note that for the *Introduce* test, the *Open Challenge*, the *Demo Challenge*, and the *Finals*, values are not indicated because teams can freely choose their performance and the focus on certain abilities themselves. This way, we expect new abilities to be demonstrated, which can be used to enhance the competition in the future.

Since the competition involves mobile robots, navigation is currently the most dominant ability represented in the score. Object manipulation and recognition also play an important role since service robots are useful if they can effectively manipulate objects in the environment. Person recognition, tracking, and speech/gesture recognition are needed to implement effective human-robot interaction behaviors. As gesture recognition was introduced as a new (and optional) ability in 2008, its weight in the total score is still comparably low. Finally, mapping plays a more limited role; such an ability is used in the Walk & Talk test, where the environment is completely remodeled during the test, so the robot enters in an unknown environment, while for other tests only minor modifications of the environment are made right before the tests. Thus, pre-computed maps (either built off-line by the robot or manually drawn) can be used.

This table is important in order to define the weight of each ability in a test and in order to distribute the abilities among the tests. Furthermore, one can analyze the performance of the teams and the difficulty of the tests after a competition. This allows for an iterative and constant development of the benchmarks.

Test	Nav	Map	PRec	PTrk	ORec	OMan	SRec	GRec	Total
FF	550	0	0	450	0	0	0	0	1000
FC	375	0	0	0	150	400	75	0	1000
WW	350	0	550	0	0	0	100	0	1000
LF	550	0	0	0	450	0	0	0	1000
PB	1000	0	700	0	0	300	0	0	2000
SM	0	0	0	0	400	1000	200	400	2000
WT	918	416	0	250	0	0	416	0	2000
CL	1000	0	0	0	550	450	0	0	2000
Tot	**4743**	**416**	**1250**	**700**	**1550**	**2150**	**791**	**400**	**16000**

Table 1: Distribution of test scores related to functional abilities

4.1.2 System properties

Similar relationships between system properties and the tests exist. As previously mentioned, this relationship cannot be quantified in scores as easily, as the system properties are of more implicit meaning for the tests. However, on the basis of the objective of the tests, the importance of each of the system properties can be estimated. Table 2 relates tests with system properties by denoting a 'very important' relation with '++', an important relation with '+', and a minor relation with '-'. Note that these symbols are used only to indicate the importance of system properties in a test, rather than defining the score of a tests.

Test	EUse	FCal	NInt	App	Adap	Rob	GAppl
IN	-	+	-	++	-	-	-
FF	-	+	-	-	-	+	+
FC	+	+	+	-	+	+	+
WW	+	+	++	-	+	+	+
LF	-	+	+	-	+	+	+
OC	-	+	+	+	+	-	+
PB	+	+	++	-	+	+	++
SM	++	+	++	-	++	+	++
WT	+	+	++	-	+	+	++
CL	-	+	-	-	++	+	++
Dem	+	+	++	+	+	-	++
Fin	+	+	+	++	+	-	++

Table 2: Importance of system properties in each test

Ability	Available score	Achieved score [max]	Achieved score [avg]
Navigation	4743 (40%)	1892 (40 %)	1178 (25%)
Object Manipulation	2150 (18%)	75 (3%)	15 (1%)
Object Recognition	1550 (13%)	450 (29%)	125 (8%)
Person Recognition	1250 (10%)	400 (32%)	190 (15%)
Speech Recognition	791 (7%)	692 (87%)	293 (37%)
Person Tracking	700 (6%)	700 (100%)	570 (81%)
Mapping	416 (3%)	416 (100%)	183 (44%)
Gesture Recognition	400 (3%)	0 (0%)	0 (0%)
Total	**12000 (100%)**	**4909 (41%)**	**2554 (21%)**

Table 3: Available and achieved score for the desired abilities

System properties are further represented in the general rules, in overall requirements, and in special properties in certain tests. By using laymen to operate the robots in the Supermarket test, the Who is Who test, and the PartyBot test, *Ease of Use* (EUse) is enforced. The restrictions on setup time and procedures demands for *Fast Calibration and Setup* (FCal). *Natural Interaction* (NInt) and *Multimodal input* is rewarded in the Supermarket test.

Appeal and Ergonomics (App) are part of the evaluation criteria in the Introduce test, the Open Challenge, and the Finals. *Adaptivity* (Adap) is especially important in the Cleaning Up test. The limited number of specifications in the tests and the environment, and the fact that people who interact with the robot are chosen randomly in many tests, demands *Robustness* (Rob).

Finally, a team can only reach the *Finals* if its robot performs well in many tests with different tasks to solve. This incorporates the aspect of *General Applicability* (GAppl).

4.2 Analysis of 2008 team performance

In the following, we will analyze the performance of the teams in these abilities during ROBO-CUP@HOME 2008 competition.

Table 3 presents the scores actually gained by the teams during the competition and the per-

centage with respect to the total score available, related to each of the desired abilities. The third column shows the result obtained by the best team, while the fourth one is the average of the results of the five finalist teams. This table allows for many considerations, such as: 1) which abilities have been most successfully implemented by the teams? 2) how difficult are the tests with respect to such abilities? 3) which tests and abilities need to be changed in order to guide development into desired directions?

From the table it is evident that teams obtained good results in navigation, speech recognition, mapping and person tracking. Notice that the reason for a low percentage score in navigation is not related to inabilities of the teams, but to the fact that part of the navigation score was available only after some other task was achieved. Speech recognition worked quite well, especially considering that the competition environment is much more challenging than a typical service or domestic application due to a large number of people and a lot of background noise. The achievements in mapping and person tracking may be explained instead by the limited difficulty of the corresponding tasks in the tests.

On the other hand, in some tasks, teams were not very successful. Object manipulation is difficult, especially when an object is not known in advance and calibration time is limited or null. Because a large proportion of the score was given for manipulation, many teams attempted it, but only a few were successful. A similar analysis holds for object and person recognition and reported slightly better results with the same difficulties arising from operating under natural environment conditions (i.e., lighting), with limited or null calibration time. Finally, gesture recognition was not implemented by teams, probably due to the small number of points available.

Table 4 summarizes the number of teams participating in each test and those which received a non-zero score. This table helps to evaluate team preferences and difficulty of the tests. Note that teams were not required to perform all the tests. Therefore, some of the zero scores in the table derive from a team's choice not to participate in a test.

An evaluation of system properties is more complicated since they are difficult to quantify precisely. Our current approach is to test for system properties through general requirements and to enforce the combination of functional abilities.

Test	Participating Teams	Teams with non-zero score
Introduce	12	12
Fast Follow	12	12
Fetch & Carry	9	5
Who's Who	8	4
Comp. Lost & Found	8	2
Open Challenge	13	13
Party Bot	5	2
Supermarket	3	3
Walk & Talk	10	10
Robot Chef	4	4
Cleaning	3	1

Table 4: Number of teams participating and gaining score for each test.

An analysis of these results is very helpful for the future development of the @HOME competition. It gives direct, quantitative feedback on the performance of the teams with respect to key abilities and tasks. This allows us to identify abilities and respective tests which need to be modified, and to adjust the weights of certain abilities with respect to the total score. Possible modifications involve:

- Increasing the difficulty if the average performance is already very high

- Merging abilities into high-level skills, more realistic tasks

- Maintaining or even decreasing difficulty if the observed performance is not satisfying

- Introducing new abilities and tests

As the integration of abilities will play an increasingly important role for future general purpose home robots, this aspect should especially be considered in future competitions.

4.3 League progress

The results obtained so far by the @HOME initiative can be measured on several levels:

- increased number of participating teams and of community members,

- increasing performance in the tests,

- increase of public awareness (media, press, Internet),

- increasing number and quality of scientific contributions (see next section).

For some of these measures, a quantitative analysis over the years is presented in the following.

Since 2006, a total of 25 teams distributed worldwide (12 from Asia, 8 from Europe, 4 from America, 1 from Australia), have participated in the three years of the ROBOCUP@HOME world championship. Furthermore, national competitions have been established in China, Mexico, Germany, Iran and Japan. These events are useful not only to test team developments and rules, but also to possibly select teams that will participate to the world championship.

Table 5 describes the number of participating teams in the annual world championship. The second column shows the number of teams that pre-registered and delivered the necessary qualification material, such as videos and a team description paper. The third column shows the number of teams that qualified after a review from the Organizing Committee, and the fourth column shows the number of teams that actually participated in the competitions. Finally, the fifth column shows the number of new teams (i.e., teams that did not participate in the previous years). The last line refers to the 2009 competition, for which 26 teams from 14 countries preregistered so far.

Year	Pre-registration	Qualification	Participation	New teams
2006	20	17	12 (440; 2.72%)	12
2007	16	13	11 (321; 3.42%)	5
2008	18	17	14 (373; 3.75%)	8
2009	26	23	-	-

Table 5: Number of participating teams

The number of and the increase in participating teams must be also related to the general participation across all leagues. (The number of total teams and percentage of @HOME teams are

given in parenthesis in the fourth column). Regardless of the drop in the total number of teams throughout all leagues in 2007 (in the US) and 2008 (in China), mainly due to high travel and shipping costs, as well as the difficulties in custom and visa affairs, the increase of percentage of @HOME teams is a clear indication of the growth of the league. Moreover, the number of pre-registrations for 2009 (in Austria) is very promising.

Furthermore, being part of the RoboCup community allows teams to exchange ideas and solutions, to plan long-term projects and to participate in the competition for several years. Indeed, it is interesting to see that some teams adapted their robots designed and built for other RoboCup Leagues to compete in @HOME, and that one team in 2006 and 2007 used the same robot in both the soccer Four-Legged and the @HOME leagues. One team in 2008 even used the same robot in both the Rescue and @HOME leagues. Moreover, many teams participated multiple years. Three teams have participated in all three years of ROBOCUP@HOME, (and they also plan to participate in 2009), and 6 teams have participated in two of the competitions.

Another important parameter to assess the results of the competition is the increase in performance. Obviously, it is difficult to determine such measure quantitatively. The main reason is that the constant evolution of the competition and the iterative modification of both the rules and the partial scores do not allow for a direct comparison.

However, it is possible to identify certain situations which indicate the success of the initiative in terms of general performance increase. Table 6 gives some examples for this increase over the last three years. The first row contains the percentage of unsuccessful tests, i.e., tests where no score was achieved at all, dropping from 83% in 2006 to 41% in 2008. The second row shows the increase in the total number of tests per competition. The third row indicates the average number of tests that teams participated in successfully (i.e., with a non-zero score). The enormous increase from from 1.0 tests in 2006 to 4.9 in 2008 is a strong indication of an average increase in robot abilities and in overall system integration.

Measure	2006	2007	2008
Percentage of 0-score performance	83%	64%	41%
Total number of tests	66	76	86
Avg. number of succ. tests p. team	1.0	2.5	4.9

Table 6: Measures indicating general increase of performance

4.4 Scientific achievements

In addition to numerical analyses of test performances, relevant scientific achievements have been obtained by teams participating in the competition. ROBOCUP@HOME provides a proper setting for developing and testing integrated solutions for mobile service robots. As a result, robot hardware and software architectures evolve over time.

This effort is demonstrated in scientific papers and in the teams' reports (Team Description Papers), which contain technical and scientific details on the hardware/software architectures and the implemented approaches and functionality. In particular, due to the nature of the @HOME competition, in these architectures special focuses are put on Human-Robot-Interaction (e.g. [Savage et al., 2008]), on personal assistive robots (e.g. [Ruiz-del-Solar, 2007]) and on high level programming for domestic service robots (e.g. [Schiffer et al., 2006]).

Scientific advancements can be also identified in specific functionality. Speech recognition evolved from difficult interaction with headsets and portable laptops (2006-2007) to speaker-independent speech recognition with effective noise cancellation using on-board microphones (2008) [Doostdar et al., 2008]. Face recognition has been made robust in the presence of spectators standing around the edges of the scenario [Correa et al., 2008, Knox et al., 2008] and tuned for real-time use [Belle et al., 2008] (Figure 2 left). Object recognition in @HOME requires a more general approach than the color-based recognition used in the soccer leagues, and it offers a challenging testbed. Techniques using different feature extractors and different matching procedures have been tested (e.g. [Loncomilla and Ruiz-del-Solar, 2007]), reaching a level in which the robot can reliably remember an object shown by a user (by holding it in front of the robot) and then recognize it among several others (2008, Figure 2 right)). Gesture detection and recognition has also been studied in order to communicate with the robot, and uses an effective approach based on active learning [Francke et al., 2007]. Finally, object manipulation has evolved from gathering a newspaper from the floor (2006), to grasping cups from a table (2007), to grasping different objects at various heights (2008) (Figure 4). A list scientific publications from ROBOCUP@HOME teams can be found in the league Wiki [32].

A measure of the scientific contributions is also given by the five papers (out of 56) related to ROBOCUP@HOME presented to the International RoboCup Symposium 2008, including one that received the best student paper award [Doostdar et al., 2008]. In comparison with all the RoboCup leagues and sub-leagues, @HOME ranked third out of ten with respect to the number of papers presented at the RoboCup Symposium (together with Soccer Middle-Size and Soccer Simulation).

4.5 Community

ROBOCUP@HOME does not only involve the aspect of competition, but it has also a strong focus on building a community exchanging knowledge and technology. This community plays a substantial role, because of the following reasons:

- The specifications of the tests and of the scenario are kept to a minimum to meet the aim of realistic tasks and the involvement of a defined amount of uncertainty. Therefore, the interpretation of the rules and a common vision on the goals to achieve must rely on common sense.

- The constant evolution and enhancement of the competition is mainly based on the input and feedback from the community towards new concepts and procedures.

- The large, real-world problem space in which the league is operating calls for interdisciplinary exchange of know-how, as problems can hardly be solved by a single group alone. This fosters the integration of existing components in combination with new specific approaches. The exchange, use and combination of standardized and modular system components from inside and outside the community is expected to accelerate technological progress.

- Establishing contact and exchange between science and industry should accelerate product and application development in DSR.

---

[32]List of @Home publications: http://robocup.rwth-aachen.de/athomewiki/index.php/Publications

ROBOCUP@HOME makes use of standard Internet tools to exchange technical knowledge and organize information. The web site[33] is dedicated to the initiative containing both the current information about the next competition, as well as historical data. The mailing list[34] is used for general communication to and within the community, including organization information, rule discussions, technical help, calls for scientific contributions, etc. In addition, a Wiki[35] for the @HOME initiative has been created with the goal of becoming a standard knowledge pool for international domestic service robotics research and development. The Wiki acts as as a platform for technological and scientific knowledge transfer on hardware, software, methods and abilities among the teams, and as a helpful starting point for new teams.

The community is growing fast. The mailing list currently has 250 subscribers (January 2009), and the number and kind of subscriptions indicate that the mailing list is not only used by the teams but also by various people from research institutions, other communities, universities, media and companies.

As of January 2009, the @HOME Wiki received about 22,000 page views and more than 300 page edits since it was set up at the end of 2007. The most popular pages are the software page (1,840 views) and the hardware page (1,676 views), which strongly indicates that knowledge is actually being exchanged in the community.

Finally, attention to the ROBOCUP@HOME activities in the media and press has increased, thanks to the many worldwide and regional events in which the competition has taken place. Various videos[36] and images[37] of past *RoboCup@Home* events are available online.

5   Conclusion and Outlook

This article presented the ROBOCUP@HOME initiative as a community effort to develop and benchmark domestic service robots through scientific competitions. To do so, we employ so-called *system benchmarking* that evaluates a robot's performance in a realistic, complex and dynamic environment. The general setting is designed to exhibit a high degree of uncertainty that the robots have to deal with.

The rules of the competition aim to implement the benchmark by means of general rules and a set of specific tests. Evaluation is conducted along a set of key features. These features, divided into *functional abilities* and *system properties* need to be met in order to be successful in the competition. The modular and open character of the competition's framework allows for an iterative adaptation of features and tests according to the observed and measured benchmark performances.

Special focus is put on establishing a community to foster interdisciplinary exchange of knowledge and technology. Furthermore, this community is essential to create common vision and understanding for the problems and goals of the @HOME initiative, and to give feedback for the iterative development of the competition.

---

[33]http://www.robocupathome.org/

[34]robocupathome@iais.fraunhofer.de, https://lists.iais.fraunhofer.de/sympa/info/robocupathome/

[35]http://robocup.rwth-aachen.de/athomewiki/

[36]Videos of the 2008 competition (http://www.youtube.com/user/RoboCupAtHome)

[37]Images of various @Home events (http://picasaweb.google.com/RoboCupAtHome)

Starting with the first competition in 2006, the overall development of the initiative with respect to performance increases, the growing community, knowledge exchanges and public awareness has been very promising over the past three years. @HOME has become the largest international competition for domestic service robots, with currently five national competitions in China, Japan, Germany, Iran and Mexico besides the annual world championships. Competitions in South America and the US are expected to be introduced in 2009.

The future development of the @HOME competition is highly iterative, as it involves constant feedback from the community, adjustments on the focus of desired abilities and changes of the rules. In general, the tests, functional abilities and desired system properties will evolve over the years and will be combined to form more realistic high-level tasks. New tests with different focuses and higher complexity will be added in the future, depending on the results of previous years.

The discussion on how to ensure a comparable measure of performance in the benchmarks, in the presence of a high level of desired uncertainty should be intensified.

Still, short, mid and long-term goals are necessary, as they help identify and approach the problem in the large, real-world problem space in a structured way. At the moment the focus is on physical capabilities such as manipulation, human recognition and navigation. In the future, more focus will be put on artificial intelligence and mental capabilities in the context of HRI. This includes situational awareness, online learning, understanding and modeling the surrounding world, recognizing human emotions and having appropriate responses.

The increase of complexity in the competition from 2007 to 2008 was rather high. Therefore, the Technical Committee of the @HOME league agreed to make only minor modifications to the rules in 2009. Rule changes for 2009 will involve an increased focus on HRI, e.g combined use of speech and gestures, robot operation by laymen, or following previously unknown persons. Application scenarios will become more realistic, e.g. the demo challenge will involve robots serving drinks and food at a real party setting involving many people unfamiliar with the robots. Furthermore, uncertainty and dynamics in the environment are increased by changing object positions more frequently, having more people in the scenario, and leaving the scenario with the robots.

Further, an annual @HOME camp will be established. It will consist of a set of lectures and practical sessions from and for members of the community. Having a separate event exclusively for knowledge exchange in the absence of any competitive aspect will help to foster exchange of knowledge even more. Also, new research groups and communities will be addressed and invited to join and share their knowledge with the @HOME community.

Midterm goals include the search, identification, design and use of a common robot software architecture or framework to better exchange and reuse software components already developed in the community and beyond. The same holds true for hardware, where companies or groups with relevant hardware components like sensors, actuators, or even standard robot platforms will be identified and asked to join and to support the community.

Another midterm goal is gradually testing the robots in the real world, e.g. going shopping in a real supermarket or taking public transportation. Moreover, usability and appearance of the robots will be of higher importance if one wants to increase their public acceptance.

The future @HOME scenario will contain more high-level and continuous interaction with humans living together with the robot and will evolve towards more synergistic human-robot teams, as depicted in the studies presented by Burke et al. [Burke et al., 2004]. Moreover, we foresee an

increased use of ambient intelligence, with which the robots can interact. The use of the Internet as a general knowledge base, and the communication with household devices, TVs, or external video cameras are some examples.

In general, the competition will move towards a high-level integration of the identified abilities into more realistic and relevant applications. This will increase attractiveness, generate more public awareness and hopefully inspire and accelerate consumer product development for domestic service robotic applications in the near future.

References

[Balakirsky, 2006] Balakirsky, S. (2006). Usarsim: Providing a framework for multi-robot performance evaluation. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop (PerMIS'06)*, pages 98–102.

[Balch and Yanco, 2002] Balch, T. and Yanco, H. A. (2002). Ten years of the aaai mobile robot competition and exhibition: looking back and to the future. *AI Magazine*, 23(1):13–22.

[Baltes, 2000] Baltes, J. (2000). A benchmark suite for mobile robots. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*.

[Belle et al., 2008] Belle, V., Deselaers, T., and Schiffer, S. (2008). Randomized trees for real-time one-step face detection and recognition. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08)*. IEEE Computer Society.

[Bradski and Pisarevsky, 2000] Bradski, G. R. and Pisarevsky, V. (2000). Intel's computer vision library: Applications in calibration, stereo, segmentation, tracking, gesture, face and object recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 796–797. IEEE Computer Society.

[Bruyninckx, 2001] Bruyninckx, H. (2001). Open robot control software: the orocos project. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pages 2523–2528. IEEE.

[Burke et al., 2004] Burke, J., Murphy, R., Rogers, E., V.J., L., and Scholtz, J. (2004). Final report for the darpa/nsf interdisciplinary study on human-robot interaction. *IEEE Trans. on Systems, Man, and Cybernetics Part C*, pages 103–112.

[Calisi et al., 2008] Calisi, D., Iocchi, L., and Nardi, D. (2008). A unified benchmark framework for autonomous Mobile robots and Vehicles Motion Algorithms (MoVeMA benchmarks). In *RSS Workshop on Experimental Methodology and Benchmarking in Robotics Research*.

[Correa et al., 2008] Correa, M., Ruiz-del-Solar, J., and Bernuy, F. (2008). Face recognition for human-robot interaction applications: A comparative study. In *Proceedings of the International RoboCup Symposium 2008 (CD-ROM Proceedings)*, To appear in LNAI-5399. Springer.

[del Pobil, 2006] del Pobil, A. (2006). Why do We Need Benchmarks in Robotics Research? In *Proc. of the Workshop on Benchmarks in Robotics Research, IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[Doostdar et al., 2008] Doostdar, M., Schiffer, S., and Lakemeyer, G. (2008). Robust Speech Recognition for Service Robotics Applications. In *Proceedings of the International RoboCup Symposium 2008 (CD-ROM Proceedings)*, To appear in LNAI-5399. Springer.

[Drury et al., 2005] Drury, J. L., Yanco, H. A., and Scholtz, J. (2005). Using competitions to study human-robot interaction in urban search and rescue. *interactions*, 12(2):39–41.

[Feil-Seifer et al., 2007] Feil-Seifer, D., Skinner, K., and Mataric, M. J. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423–439.

[Fontana et al., 2008] Fontana, G., Matteucci, M., and Sorrenti, D. G. (2008). The RAWSEEDS proposal for representation-independent benchmarking of SLAM. In *RSS Workshop on Experimental Methodology and Benchmarking in Robotics Research*.

[Francke et al., 2007] Francke, H., Ruiz-del-Solar, J., and Verschae, R. (2007). Real-time hand gesture detection and recognition using boosted classifiers and active learning. In *Advances in Image and Video Technology, Second Pacific Rim Symposium (PSIVT 2007)*, LNCS 4872, pages 533–547. Springer.

[Gerkey et al., 2003] Gerkey, B. P., Vaughan, R. T., and Howard, A. (2003). The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the 11th International Conference on Advanced Robotics*, pages 317–323.

[Howard and Roy, 2003] Howard, A. and Roy, N. (2003). The robotics data set repository (radish).

[Kahn et al., 2007] Kahn, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., and Miller, J. (2007). What is a human? toward psychological benchmarks in the field of humanrobot interaction. *Interaction Studies*, 8(3):363–390.

[Kitano et al., 1997] Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., and Matsubara, H. (1997). RoboCup: A Challenge Problem for AI. *AI Magazine*, 18(1):73–85.

[Kitano and Tadokoro, 2001] Kitano, H. and Tadokoro, S. (2001). RoboCup Rescue: A Grand Challenge for Multiagent and Intelligent Systems. *AI Magazine*, 22(1):39–52.

[Knox et al., 2008] Knox, W. B., Lee, J., and Stone, P. (2008). Domestic interaction on a segway base. In *Proceedings of the International RoboCup Symposium 2008 (CD-ROM Proceedings)*, To appear in LNAI-5399. Springer.

[Loncomilla and Ruiz-del-Solar, 2007] Loncomilla, P. and Ruiz-del-Solar, J. (2007). Robust object recognition using wide baseline matching for robocup applications. In *RoboCup 2007: Robot Soccer World Cup XI*, LNAI 5001. Springer.

[Meeden et al., 2000] Meeden, L., Schultz, A. C., Balch, T. R., Bhargava, R., Haigh, K. Z., Bohlen, M., Stein, C., and Miller, D. P. (2000). The aaai 1999 mobile robot competitions and exhibitions. *AI Magazine*, 21(3):69–78.

[Montemerlo et al., 2003] Montemerlo, M., Roy, N., and Thrun, S. (2003). Perspectives on standardization in mobile robot programming: The carnegie mellon navigation (carmen) toolkit. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2436–2441.

[Munoz et al., 2007] Munoz, N., Valencia, J., , and Londono, N. (2007). Evaluation of navigation of an autonomous mobile robot. In *Proc. of Performance Metrics for Intelligent Systems Workshop (PerMIS)*, page 1521.

[Prassler, Erwin and Hägele, Martin and Siegwart, Roland, 2006] Prassler, Erwin and Hägele, Martin and Siegwart, Roland (2006). International Contest for Cleaning Robots: Fun Event or a First Step towards Benchmarking Service Robots.

[Ruiz-del-Solar, 2007] Ruiz-del-Solar, J. (2007). Personal robots as ubiquitous-multimedial-mobile web interfaces. In *Proc. of 5th Latin American Web Congress (LA-WEB)*, pages 120–127.

[Sabanovic et al., 2006] Sabanovic, S., Michalowski, M., and Simmons, R. (2006). Robots in the wild: observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control*, pages 596–601.

[Savage et al., 2008] Savage, J., Ayala, F., Cuellar, S., and Weitzenfeld, A. (2008). The use of scripts based on conceptual dependency primitives for the operation of service mobile robots. In *Proceedings of the International RoboCup Symposium 2008 (CD-ROM Proceedings)*, To appear in LNAI-5399. Springer.

[Schiffer et al., 2006] Schiffer, S., Ferrein, A., and Lakemeyer, G. (2006). Football is coming Home. In *Proc. of International Symposium on Practical Cognitive Agents and Robots (PCAR'06)*. University of Western Australia Press.

[Steinfeld et al., 2006] Steinfeld, A., Fong, T., Kaber, D. B., Lewis, M., Scholtz, J., Schultz, A. C., and Goodrich, M. A. (2006). Common metrics for human-robot interaction. In *Proc. of HRI*, pages 33–40.

[van der Zant and Wisspeintner, 2005] van der Zant, T. and Wisspeintner, T. (2005). RoboCup X: A Proposal for a New League Where RoboCup Goes Real World. In Bredenfeld, A., Jacoff, A., Noda, I., and Takahashi, Y., editors, *RoboCup*, volume 4020 of *Lecture Notes in Computer Science*, pages 166–172. Springer.

[van der Zant and Wisspeintner, 2007] van der Zant, T. and Wisspeintner, T. (2007). *Robotic Soccer*, chapter RoboCup@Home: Creating and Benchmarking Tomorrows Service Robot Applications, pages 521–528. I-Tech Education and Publishing.

[Wisspeintner and Nowak, 2007] Wisspeintner, T. and Nowak, W. (2007). *Robotic Soccer*, chapter VolksBot - A Construction Kit for Multi-purpose Robot Prototyping, pages 529–548. I-Tech Education and Publishing.

[Yanco et al., 2004] Yanco, H. A., Drury, J. L., and Scholtz, J. (2004). Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition. *Human-Computer Interaction*, 19:117–149.

About the authors

Thomas Wisspeintner   studied electrical- and information technology at TU Munich with specialization in mechatronics. In 2002 he wrote his diploma thesis on mobile robot manipulation at the Fraunhofer Institute IAIS (former AIS). There, from 2002 until 2008 he led the VolksBot project, designing a construction kit for application oriented mobile robot prototyping. Since 2004 he is PhD student at the Department of Mathematics and Computer Science, FU Berlin with a focus on prototyping and benchmarking of mobile service robot applications. He is executive member of the RoboCup Federation since 2007 an co-founder of the @Home league. Since 2009 he is assistant manager of the sensor development department at Micro-Epsilon Messtechnik, Germany.

Tijn van der Zant   received his M.Sc. in Artificial Intelligence at Utrecht University, the Netherlands, in 2003. He has a specialization in real time learning in robots and artificial brain architectures. After his M.Sc. he worked at the Fraunhofer Institute in Bonn, Germany. In 2005 he started pursuing his Ph.D. at the University of Groningen at the Artificial Intelligence department, which will be finished in 2009. His research focuses on connected cursive handwriting recognition, home robotics and also on applying machine learning on machine learning, on pattern recognition and on feature extraction in order to create learning artificial brain architectures. He is also an executive member of the RoboCup Federation since 2007 and the co-founder of the largest home robotics competition called RoboCup@Home.

Luca Iocchi   received his Ph.D. in Computer Engineering from University "La Sapienza" in Rome and he is Assistant Professor in the same university within the Department of Computer and System Sciences (DIS), since 2002. His research interests in the field of artificial intelligence and cognitive robotics include: reasoning about actions and planning under incomplete knowledge, cognitive robotics, multi-agent and multi-robot coordination, mobile robot localization navigation and mapping, robot learning, vision-based people and object tracking. He is member of the program and executive committees of many artificial intelligence and robotics related journals, conferences and events and he is author of more than 100 referred papers.

Stefan Schiffer   studied computer science with a minor in medicine and specialization in artificial intelligence at RWTH Aachen University. He received his diploma degree (Dipl.-Inform., comparable to M.Sc.) from RWTH Aachen University end of 2005. He is a research and teaching assistant and a Ph.D. candidate at the Knowledge-based Systems Group headed by Prof. G. Lakemeyer, Ph.D. at RWTH Aachen University since 2006. He has served as a member of the Technical Committee of the RoboCup@Home league in 2006/2007, 2007/2008, and 2008/2009. His research interests include Cognitive Robotics, Human-Machine Interaction, and Embodied Artificial Intelligence.