RoboCup@Home: Adaptive Benchmarking of Robot Bodies and Minds

Tijn van der Zant[1] and Luca Iocchi[2]

[1] Cogitive Robotics Laboratory, A.I. dept, faculty of Natural Sciences, University of Groningen
tijn@ieee.com
[2] Department of Computer and Systems Science, University of Rome "La Sapienza"

Abstract. RoboCup@Home is the largest benchmarking effort for domestic service robots. The benchmarking is in the form of a competition, with several yearly local competitions and an international one. Every year the tests become more complex, depending on the results of the previous years. In the past four years the focus has been on benchmarking physical aspects of the robots, such as manipulation, recognizing people and human-robot interaction. In 2010, for the first time, there is a test which is targeted at the mental cognitive capabilities of the robot. In order to guarantee scientific quality of the proposed solutions and effective integration in a fully working system, all the tests include different capabilities and change every year. This novel feature of RoboCup@Home benchmarking raises the question of: How can effective benchmark tests be defined and at the same time measure the progress over many years? In this paper we present the methodology applied in and results of RoboCup@Home for measuring the effectiveness of benchmarking service robots through competitions and present a new integrated test for benchmarking the cognitive abilities of a robot.

1 Introduction

RoboCup@Home[1] is the largest human-robot interaction and domestic service robot benchmarking effort. The goal of RoboCup@Home is to foster the development of a versatile domestic service robot that can operate safely in all situations that people encounter in daily life. RoboCup@Home [5,6] is part of the RoboCup Federation [3,4], which has been promoting, for many years already, the use of competitions in order to drive research towards robust techniques and useful applications and to stimulate teams to compare their approaches on a set of common testbeds. This dynamical form of benchmarking has dramatically improved the standards of robotics. Robots have become much more robust and reactive.

RoboCup@Home competition focuses on the benchmarking of domestic service robots. It consists of several tests where many functionalities are tested at the same time in an integrated way. Most of the tests integrate navigation, localization, object recognition, manipulation, speech recognition, etc. Examples

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

B. Mutlu et al. (Eds.): ICSR 2011, LNAI 7072, pp. 214–225, 2011.

of tests are: the robot has to find and bring objects scattered in the environment in unknown locations, assist with cooking or bringing a drink, follow its owner through a real world environment such as a shopping mall, recognize known persons in a confusing situation, reason about the world and assist with shopping in a real supermarket. Most tests are performed in an apartment scenario and some tests are performed in the real world. RoboCup@Home started in 2006 and has now many local events and a global one. This year (2011) 24 teams have participated in the global event. During this event more than 30 robots are tested resulting in over 175 separate benchmarks. It is the most variable benchmarking effort within the RoboCup Federation, and arguably the most versatile robotic benchmarking effort in general. This versatility is both the strength and the weakness of the competition.

Benchmarking has many advantages, such as the creation of standard references and metrics. However, there is also a possible disadvantage when benchmarking remains static over time, that is the progress of solutions towards a local optimum, without having a guarantee that the devised solution will work in general. In robotics this can be a huge problem. By changing the tests on a yearly basis, it is possible to ensure that specialized solutions looking for a local optimum will not be rewarded in the long term.

On the other hand, a dynamic benchmark that changes over time introduces the problem of relating the different generations of the particular benchmark [2]. This problem is acute regarding tests that require the robot to operate in the real world. Statistical analysis on the results is the scientific approach for this. Robots have to solve not a single task, but many tasks, where capabilities are tested several times in different settings.

The downside of the changing tests is that this co-evolutionary process does not guarantee progress into the desired direction of versatile autonomous domestic service robots. It is essential that scientific standards are applied to this type of flexible benchmarking. In RoboCup@Home statistical analysis of the outcome of the competition is used to steer the league into interesting directions.

This paper thus presents an important contribution to benchmarking human-robot interaction, domestic service robots and robot cognition. In the following the methodology used to drive the benchmarking through competitions is presented, and a new test aiming at measuring cognitive abilities of a robot is described. This is, to the best of our knowledge, the first general benchmark for high-level reasoning of cognitive robots. Also the results of the analysis performed during the past 4 years are presented, showing the effectiveness of the benchmarking activity.

A more detailed description of the RoboCup@Home competition, the tests used until 2008 and past results can be found in [8]. In this paper, in addition to updating the results with 2009 and (partially) 2010 competitions, the methodology for steering the dynamic benchmarking is desribed in more detail. Also, a novel test for the evaluation of the high-level cognitive abilities abilities of robots is introduced.

2 Defining the Road Map Through Adaptive Benchmarking

The road map of RoboCup@Home is not completely clear. The goal is clear (a versatile domestic service robot) but how to get there is not. This is due to the fact that the prediction of future development in robotics hardware and software is not easy. This uncertainty does not prevent benchmarking though, but the approach has to be adapted to the response of the research groups participating to it. In this situation, a fixed setting in benchmarking is not likely to bring good solutions, since it may be either too complex or too simple. Moreover, defining several benchmarks is very complicated and does not ensure a large participation of research groups. The ideal solution is to devise flexible and variable benchmarks that reflect the lessons that are continuously being learned.

An example we report from RoboCup@Home experience is the case of speech and gesture recognition, where one main goal is to understand an unknown (English speaking) person. Since the start of the @Home competition, it has been mandatory to only use speech or gestures to interact with the robot. At the international competition though, there is a tribune with up to several hundreds of spectators making a rather large amount of noise. In the first year the standard solution was to use wireless headsets to interact with the robot. The next year bonus points were introduced for not using a headset. Soon several teams tried to gain points using on-board microphones [1].

Since effective and robust solutions for speech recognition have been developed by almost all teams so far, gesture recognition was not considered. In 2009 we introduced a test where gesture recognition was stimulated by means of being able to get points for it explicitly. Gestures were restricted to hand and arm movements to avoid solutions where people have to perform unnatural actions. One team got the full score for this, demonstrating that it is possible. HRI will be tested in a real noisy environment: a shopping mall including actual customers. During the new 'follow me' test the user has to walk 3 meters away while the robot is standing still, and then give a command to make the robot move towards the user. The noise levels in such an environment will make this task very challenging and it is unclear to us whether speech or gesture will be more reliable, probably the right combination of both will be the best choice. With many teams attempting to find an effective and robust solution in the very challenging environments of RoboCup@Home, we will obtain some important results:

1. Statistical evidence of performance of the developed techniques is collected
2. The difficulty of the task is measured
3. Based on the results, recommendations for changing the tests in 2011 are discussed amongst peers

We firmly believe that this approach to benchmarking will quickly raise the general level of research groups developing solutions in domestic service robotics and may even fill the gap between laboratory experimental settings and real robotic applications, thus providing an important link with industries. The statistical benchmarking approach is the only viable approach for the benchmarking of real-world robots. Any form of fixed benchmarking will only create local optimal solutions that are unlikely to work in the real world. This approach could be effectively ported to other robotic application fields where the environment has a high level of uncertainty, such as the fields of rescue, space and reconnaissance robots.

3 From Physical to Mental Capabilities

Benchmarking physical capabilities, such as Self Localization And Mapping, manipulation of objects and recognizing persons, is not an easy task. The capabilities have to be tested in different environments and situations. RoboCup@Home provides these environments and situations. In the first years the competition was situated in an apartment, which is unknown beforehand and looks different at every competition. During the competition random changes are made to the apartment to simulate a place where people live and which they adapt to their likening. The setting is different for every test. It could be that a person has lost an object and the robot has to find it, that the robot has to remember and serve the correct drinks to guests or assist with shopping in a supermarket.

In 2010 new environments have been introduced. There is a test where the robot has to follow its user through a setting in a public area. During this test the user will be occluded from sight, walk away several meters and stand close to other persons in order to test the quality of the following behaviors of the robot. Another environment is used in the supermarket test. In this test the user and the robot go into a real supermarket. The robot has to assist the user by getting objects (such as a box of cornflakes) from a shelf and hand it over to the user. Also, at the check-out, the robot has to go back into the supermarket and get an item that the user 'forgot'. This test could lead to multi-user robotic applications that can assist customers with physical challenges. In the future the environments will be extended to, for example, going into town using the public transport to do some shopping or to accompany children walking from school to home.

What's missing in the test description mentioned so far is the testing of mental or high-level cognitive capabilities. Although the robots need to posses some form of intelligence for these tasks, it is important that the robots are ready for new and unforeseen situations. During the competitions there is not enough time to benchmark all possible situations which might occur in real life. The only solution is to test the cognitive capabilities of the robots, in order to assess whether it is likely that the robot can handle new situations, and use statistics for extrapolation of the results.

3.1 The First Cognitive Benchmark in Robocup@Home

In this section we present an integrated test to measure the cognitive capabilities of a robot. This test is called 'General Purpose Test' and it focuses on the following aspects.

- There is no predefined order of actions to carry out. This is to slowly get away from state machine-like behavior programming.
- There is an increased complexity in speech recognition compared to the other tests. Possible commands are less restricted in both actions and arguments. Commands can include multiple objects, e.g., "put the mug next to the cup on the kitchen table"
- The test is about how much the robot understands about the environment and aims for high-level reasoning.

The task will be executed as follows. The robot enters the arena by driving to a specified location. The referees select an action from the list of possible actions and command the robot to perform the desired task. Once the robot has successfully solved or interrupted the execution of the task, a new action is selected. The robot can carry out a maximum of three actions within 10 minutes. Actions, locations and objects are taken from previous tests. Since this is an advanced test in the competition, we can correctly assume that participating teams are familiar with such actions, locations and objects and have already solved in an effective way the basic functionalities, like navigation, SLAM, object recognition, manipulation, HRI, etc. This guarantees that the main focus of the test is on high-level cognitive capabilities.

3.2 Actions and Categories of Tasks

Regarding the following specifications of actions and complexity classes, several actions can be combined for a *compound action*, for example,

- "This is John, now follow John"
- "Go to the table and point at the yellow cup"
- "Go to the living room and find John"
- "This is a coffee mug (showing the mug), now go to the kitchen and point at the mug on the kitchen table"

Moreover, we have defined three different categories of tasks that form the test.

1. **Understanding orders:** Understand complex sentences and perform the correct actions:
   - Robot, go to the bed room, pick up the red cup next to the bed and put it on the kitchen table
   - Robot, follow the person in front of you and stop when you are in the kitchen to pick up the red cup from the table
   - Robot, go to the living room, find John (a known person), tell him your name and then come back to me

2. **Understanding itself:** The robot gets a simple command, but it does not have all the information necessary to complete the task. The robot may ask up to five questions to get relevant information, before starting with the task. Examples are the following:
   - *user: "Robot, bring me a cup"*
     robot: "Which cup?"
     user: "My cup." or "The red cup."
     robot: "Where is the cup?"
     user: "On the table."
   - Another example:
     user: "robot, put the red cup on the table!"
     The robot does not have the red cup in its hands, it needs to get it. Since searching takes a long time, it can ask:
     robot: "Excuse me, do you know where the cup is?"
     user: "It's in the bedroom."
     And the robot is allowed to pursue further:
     robot: "Could you tell me where exactly in the bed room?"
     user: "It's on the table next to the bed"
     And the robot performs the action
3. **Understanding the world:** The robot has to answer 3 questions about (events in) the world. The necessary recognition capabilities are all from previous tests. The first two questions are about understanding the world as it is, for example asking how many persons there are in the room. The third question is where the robot is being *"tricked"*. Examples of a trick are putting a box over a cup and asking the robot where the cup is, or having a person sitting at the table with the robot and a person standing and asking the robot how many people are sitting at the table. More specific examples are the following:

   (a) The robot is in front of a table with several cups in different colors on it. The human can ask and do the following:
   user: "Robot, can you tell me what's on the table?", and the robot has to answer something like
   robot: "I see three cups, one is red, one is blue and one is yellow" or
   robot: "I see a red cup, a blue cup and a yellow cup".
   A wrong answer would be that the robot says that it only sees one cup, for example.

   (b) Then the human can ask:
   user: "Robot, where is the red cup in relation to the other cups?", the correct answer would be
   robot: "The red cup is left compared to the other cups.", of course depending on the position of the cup.

   (c) A third action/question could be that the human takes away 2 of the cups, and the puts a box over the remaining (blue) one. The human can then asks:

user: "Robot, do you know where the blue cup is?", the correct answer would be
robot: "The blue cup is under (behind) a box",
a wrong answer would be "There is no blue cup". The difference is between stating what the robot sees, or testing that the robot is aware of object persistence. The robot is in front of a table with several cups in different colors on it. The human can ask and do the following:
Although this task is complex, the criterion is always the same: Did the robot show understanding, or is it only stating what it sees.

The setup is experimental since this is the first year that high-level cognition is tested. The test is loosely based on skills which are still difficult for robots in general (such as parsing a complex sentence into a set of actions or reasoning about its own capabilities) and developmental psychology, such as reasoning about counting objects and occlusion events [7].

4 Statistical Analysis of the Benchmarking

One important objective of our work is to measure the advances of performance over time of a changing benchmarking. In RoboCup@Home a two-steps methodology for benchmarking analysis was adopted (described in details in [8]) based on the definition of a set of desired functional skills, a set of tests executed by participating teams, the definition of a score system that allow to relate score points to skills in the tests, and the evaluation of the amount of score available and actually gained by the best teams for each functionality in each test. In this way, it is possoble to evaluate the average increase of performance in the given skills over years even when changing the tests, in order to make them closer to real world applications.

The functional abilities that have been considered in the competitions so far are the following.

- *Navigation*, the ability of path-planning and safely navigating to a specific target position in the environment, avoiding (dynamic) obstacles.
- *Mapping*, the ability of autonomously building a representation of a partially known or unknown environment on-line.
- *Person Recognition*, the ability of detecting and recognizing a person.
- *Person Tracking*, the ability of tracking the position of a person over time
- *Object Recognition*, the ability of detecting and recognizing (known or unknown) objects in the environment
- *Object Manipulation*, the ability of grasping or moving an object
- *Speech Recognition*, the ability of recognizing and interpreting spoken user commands (speaker dependent and speaker independent)
- *Gesture Recognition*, the ability of recognizing and interpreting human gestures
- *Cognition*, the ability of understanding and reasoning about the world, beyond current perceptions.

As already mentioned in the previous section *Cognition* is a new ability that has been introduced in 2010.

4.1 Analysis of Desired Abilities in the Tests

The first step of our methodology is to define a *functionality-score table* that is used to decide how to distribute the total score in the different desired abilities. This is generated through an iterated process that takes into account both the plans of the Executive and Technical Committees and the feedback from the teams.

Table 1. Functionality-score tables since 2008

Ability	2008	2009	2010
Navigation	40%	33%	22%
Mapping	3%	3%	9%
Person Recognition	10%	12%	12.5%
Person Tracking	6%	4%	3%
Object Recognition	13%	17%	7.5%
Object Manipulation	18%	17%	14%
Speech Recognition	7%	8%	15%
Gesture Recognition	3%	6%	3.5%
Cognition	-	-	13%

Table 1 shows the percentage of score for each functionality over time. Navigation is the most dominant ability, because the competition involves mobile robots. However, since the quality of navigation is increasing and reached very good levels, during the years the impact on the total score is progressively reduced, thus leading the teams to focus on other functionalities. Mapping plays a more limited role, since the environment does not change in a significant way during the competition. Tests are performed outside the arena (e.g., in a real supermarket) and the ability of on-line mapping will be tested in a very realistic environment. Person recognition and tracking are also fundamental abilities. While tracking is easy (thus the score is decreasing over time), recognition is more difficult (score is increasing). In this way research on recognition is stimulated more than, for example, tracking. Object recognition and manipulation also play an important role, however these abilities are in general much more difficult than those related to people. In particular, object manipulation has reached good results only in the last year (see below). Therefore the score for these two functionalities is varying according to the difficulty of the tasks and to the actual accomplishments of the teams. Speech and gesture recognition are needed to implement effective human-robot interaction behaviors. Although in many tests the use speech or gesture is left to the teams, speech is largely preferred. In 2009, we tried to stimulate gesture recognition by increasing the available score, but it did not work (see below) since teams continued to use speech. In 2010 we have devised a test in which the robot has to understand human commands from a distance of 3 meters in a noisy environment, where it is expected that speech will be very challenging and gesture may be even more

convenient. Finally, as already mentioned, we decided to assign an important percentage of the score to measure cognitive abilities, in order to focus more on general purpose and robust solutions.

4.2 Analysis of Team Performance

The second step of the methodology is to analyze the actual performance of the teams in these abilities during the competitions. Here we present the results of RoboCup@Home 2008 and 2009.

Table 2. Achieved scores for the desired abilities. 'max' means maximum score achieved and 'av' stands for average score achieved

Ability	2008 max	2008 av	2009 max	2009 av
Navigation	40%	25%	47%	40%
Mapping	100%	44%	100%	92%
Person Recognition	32%	15%	69%	37%
Person Tracking	100%	81%	100%	69%
Object Recognition	29%	8%	39%	23%
Object Manipulation	3%	1%	48%	23%
Speech Recognition	87%	37%	89%	71%
Gesture Recognition	0%	0%	0%	0%
Average	**41%**	**21%**	**61.5%**	**44.4%**

Table 2 presents the percentage of the available scores actually gained by the teams during 2008 and 2009 competitions, related to each of the desired abilities. The second and fourth columns show the best result obtained by some team, while the third and fifth ones contain the average of the results of the five finalist teams.

This table allows for many considerations, such as: 1) which abilities have been mostly 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 future development into desired directions.

From this table, two important aspects are evident. First, the general increase of performance in all the functionalities from 2008 to 2009. Second, the functionalities that have been almost completely solved, and those that are not.

In fact, teams obtained good results in navigation, mapping, person tracking and speech recognition (average above 50%, except for navigation). 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. The good results for speech recognition is very relevant since the competition environment is much more challenging than a typical service or domestic application due to a large amount of people and a lot of background noise. On the other hand, the results for mapping and person tracking are due to the fact that they were not applied in a difficult environment.

As already mentioned, this changed in 2010 since new tests in which person tracking and mapping are important will be executed in real environments (a shopping mall and a supermarket). Person recognition performance is acceptable and thus in 2010 we increased a little bit the difficulty of this functionality during the tests. For example, more unknown people will be present during the tests, passing between the robot and the leader during person following, in order to test robustness of the developed methods. Object manipulation had in 2009 the highest increase of performance. Although more robust solutions have been developed by the teams, there is still some work to be done. Therefore for 2010 we did not increase the difficulties of object manipulation in the tests. Also object recognition is reaching good performance and will not become more difficult. Finally, we recognize a problem in gesture recognition which has not been implemented by teams. In fact, only one team (not within the finalists) implemented effective gesture recognition techniques. As already mentioned, the increase of available score was not sufficient to motivate teams to work in this direction. So for 2010 we have created situations where speech is likely to fail and gesture may be the only practical solution to solve the problem.

Summarizing, by analyzing the results of team performance it is possible to decide about future development of the benchmarks. Possible adjustments are:

- Increasing the difficulty if the average performance is high
- Merging of abilities into high-level skills, more realistic tasks
- Keeping 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 be especially considered in the future competitions.

Other important parameters to assess the success of a benchmark are the number of participating research groups (teams) and the general increase of performance over the years. Obviously, it is difficult to determine such measures in a quantitative way: the constant evolution of the competition with its iterative modification of the rules and of the scores do not allow a direct comparison. However, it is possible to define some metrics of general increase of performance. They are based on the capability of a team to gain score in multiple tests, thus showing the effectiveness not only in implementing the single functionalities, but also in integrating them in a working system, as well as in the realization of a flexible and modular architecture that allow for executing different tasks.

In Table 3, the number of teams participating to the international competition is shown in the first row. The league received a lot of interest since the beginning (2006), then we registered a general increase. 27 teams (out of 32 requests) have been qualified and are allowed to participate for RoboCup@Home 2010. The second row contains the percentage of successful tests, i.e., tests where some score greater than zero was achieved, showing a significant and constant increase from 17% in 2006 to 83% in 2009. The third row shows the increase in the total number of tests executed by all the teams during the competition. The

Table 3. Measures indicating general increase of performance

Measure	2006	2007	2008	2009
Number of teams	12	11	14	18
Percentage of succ. tests	17%	36%	59%	83%
Total amount of tests	66	76	86	127
Avg. succ. tests p. team	1.0	2.5	4.9	7.3

execution of over 100 tests in 2009 confirms the significance of the statistical analysis we are performing. Finally, the fourth row contains the average number of successful tests for each team. This is a very important measure, since the enormous increase from 1.0 tests in 2006 to 7.3 in 2009 is a strong indication for an average increase in robot abilities and in overall system integration. A team successfully participating in an average of 7 tests (that are quite different each other) demonstrates not only effective solutions and implementation of all the desired abilities, but also a flexible integrated system that has important features for real world applications. Notice that in this table all the teams were considered (not only the five finalists).

The results obtained by the analysis reported here clearly show that our methodology of dynamic benchmarking is producing a quick and significant progress in domestic service robotics.

5 Discussion

The benchmarking high-level robot cognition has just started. The correct paradigm has to be established. RoboCup@Home is actively researching this flexible benchmarking by means of the organization of a world-wide effort and measuring the progress over the years. Although the benchmarking of physical capabilities of the robot in dynamic and poorly structured environments is still in development, there should also be a focus on high-level cognitive tasks that the robot has to perform.

The statistical procedures developed in this competition are useful to start the discussion on the topic of high-level cognitive benchmarking of robots. The methodology can probably be improved and further discussion is needed. RoboCup@Home tries to stimulate not only the benchmarking itself, but also the meta-process about how to benchmark.

The results described in this paper provide evidence that dynamic benchmarking is a viable approach. It can probably be used in many more real-world settings with high levels of uncertainty.

Acknowledgments. The authors gratefully acknowledge the RoboCup Federation for their support over the past five years. We also acknowledge the effort of Thomas Wisspeintner in setting up the league and the help of Stefan Schiffer in making it progress.

References

1. Doostdar, M., Schiffer, S., Lakemeyer, G.: A Robust Speech Recognition System for Service-robotics Applications. In: Iocchi, L., Matsubara, H., Weitzenfeld, A., Zhou, C. (eds.) RoboCup 2008. LNCS, vol. 5399, pp. 1–12. Springer, Heidelberg (2009)
2. Feil-Seifer, D., Skinner, K., Mataric, M.J.: Benchmarks for evaluating socially assistive robotics. Interaction Studies 8(3), 423–439 (2007)
3. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E.: Robocup: The robot world cup initiative. In: AGENTS 1997: Proceedings of the First International Conference on Autonomous Agents, New York, NY, USA, pp. 340–347. ACM (1997)
4. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., Matsubara, H.: RoboCup: A Challenge Problem for AI. AI Magazine 18(1), 73–85 (1997)
5. van der Zant, T., Wisspeintner, T.: RoboCup X: A Proposal for a New League Where RoboCup Goes Real World. In: Bredenfeld, A., Jacoff, A., Noda, I., Takahashi, Y. (eds.) RoboCup 2005. LNCS, vol. 4020, pp. 166–172. Springer, Heidelberg (2006)
6. van der Zant, T., Wisspeintner, T.: RoboCup@Home: Creating and Benchmarking Tomorrows Service Robot Applications. In: Robotic Soccer, pp. 521–528. I-Tech Education and Publishing (2007)
7. Wilcox, T., Baillargeon, R.: Object individuation in infancy: The use of featural information in reasoning about occlusion events. Cognitive Psychology 37(2), 97–155 (1998)
8. Wisspeintner, T., van der Zant, T., Iocchi, L., Schiffer, S.: Robocup@home: Scientific Competition and Benchmarking for Domestic Service Robots. Interaction Studies. Special Issue on Robots in the Wild 10(3), 392–426 (2009)