Synthetical Benchmarking of Service Robots: A First Effort on Domestic Mobile Platforms

Min Cheng[1(✉)], Xiaoping Chen[1], Keke Tang[1], Feng Wu[1], Andras Kupcsik[3], Luca Iocchi[2], Yingfeng Chen[1], and David Hsu[3]

[1] University of Science and Technology of China, Hefei, China
ustccm@mail.ustc.edu.cn
[2] Sapienza University of Rome, Rome, Italy
[3] National University of Singapore, Singapore, Singapore

Abstract. Most of existing benchmarking tools for service robots are basically qualitative, in which a robot's performance on a task is evaluated based on completion/incompletion of actions contained in the task. In the effort reported in this paper, we tried to implement a synthetical benchmarking system on domestic mobile platforms. Synthetical benchmarking consists of both qualitative and quantitative aspects, such as task completion, accuracy of task completions and efficiency of task completions, about performance of a robot. The system includes a set of algorithms for collecting, recording and analyzing measurement data from a MoCap system. It was used as the evaluator in a competition called the BSR challenge, in which 10 teams participated, at RoboCup 2015. The paper presents our motivations behind synthetical benchmarking, the design considerations on the synthetical benchmarking system, the realization of the competition as a comparative study on performance evaluation of domestic mobile platforms, and an analysis of the teams' performance.

1 Introduction

Benchmarking robotic systems is challenging [1–3]. Robotic competitions are believed to be a feasible way to overcome the difficulty by appealing research groups to take their experimental results to be compared under the same test conditions [4]. Lots of well recognized competitions are held every year around the world. The focus of AAAI [5–7] and IJCAI [8] Robot Competitions is putted on benchmarking AI and robotic technology with relevance to real-life applications and changes yearly. DARPA Robotics Challenge [9] aims to develop semi-autonomous ground robots that can do complex tasks in dangerous, degraded, human-engineered environments. RoboCup[1], an initiative to promote research in AI, robotics, and related fields, currently is the largest robotics competition, with a number of leagues such as RoboCup Soccer, RoboCup Rescue, RoboCup@Work, RoboCup@Home [10–12]. RoboCup@Home aims to drive

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

L. Almeida et al. (Eds.): RoboCup 2015, LNAI 9513, pp. 377–388, 2015.
DOI: 10.1007/978-3-319-29339-4_32

research on domestic robotics towards robust techniques and useful applications and to stimulate teams to compare their approaches on a set of common tests, and has resulted in improvement of capabilities of domestic service robots (DSRs) such as mobile manipulation [13], human-robot interaction, object recognition. RoCKIn [14,15], a project of FP7, broadens the scope of RoboCup@Home and RoboCup@Work in terms of scientific validity by being organized as a scientific benchmarking competition.

Most of existing competitions focus on qualitative evaluation on the performance of a robot, do not provide quantitative evaluation on what degree of performance a robot achieves. The objective of this effort is to advance and extend benchmarking competition by introducing quantitative evaluation. We share the same objective with RoCKIn, while taking a different approach. RoCKIn is a top-down endeavor by starting from a global framework for its long-term goals. Our effort is bottom-up in the sense that we started our endeavor from a much smaller case study—synthetical benchmarking of domestic mobile plarforms (DMPs).

We describe our motivations of introducing synthetical benchmarking in Sect. 2. A set of prescribed features for benchmarking DSRs are given in Sect. 3. Based on these features, the BSR challenge was organized at RoboCup 2015. The implementation of the BSR challenge is presented in Sect. 4. We provide an analysis on performance of participating teams to the BSR challenge in Sect. 5. A brief discussion and future work are given in Sect. 6. We draw conclusions in Sect. 7.

2 Why Synthetical Benchmarking

In this paper, by synthetical benchmarking we mean benchmarking that includes both qualitative and quantitative benchmarking. A qualitative benchmarking evaluates robot performance based on completion/incompletion of the actions contained in a task, where only two outcomes, i.e., completion or incompletion of each of these actions, are considered. Then some statistics on the qualitative outcomes of the actions may be made as an evaluation of the task. As an example, consider a task consisting of only one action pick up a can. In current competition of @Home league, one can only observe whether the action is completed or not by a robot, as an evaluation of the robot's performance on this task.

A quantitative benchmarking provides quantitative evaluation of robot performance on tasks. For example, when a robot completes a task/action, accuracy (such as errors) of the task/action completion can be acquired in quantitative evaluation. For instance, consider action *move to a waypoint*. In quantitative benchmarking, one can acquire quantitative measurement, the errors, of the robot's moving performance on the task. Without this quantitative evaluation, it is very hard to acquire any objective and accurate evaluation on the moving performance.

There are strong reasons why synthetical benchmarking of service robots is needed by introducing quantitative evaluation. First, quantitative benchmarking can generates finer evaluation than qualitative benchmarking can. Suppose either

Robot-1 or Robot-2 complete a task (say, pick up a can) with 80 % success. Then one cannot distinguish between the two robots performance on the task. However, there may be significant difference between accuracies of the two robots completions of the task. In some scenarios (as for the task of move to a waypoint) and applications, accuracy is a necessary factor in performance evaluation of robots.

Second, inclusion of both quantitative and qualitative aspects in performance evaluation of service robots supports better trade-offs among these aspects. Tasks in the @Home and @Work competitions are complicated and thus should be evaluated based on trade-offs among multiple performance factors. Generally, a comprehensive evaluation of such a task should include the following performance factors: completion of the task, accuracy of completions, and efficiency of completions (which can be measured simply with the time a robot spends for its completion of a task). A more reasonable overall evaluation should reflect some trade-offs among these factors, with accuracy being included in.

Third, accuracy data enable new solutions to some of costly work in development of service robots. For many functionalities of a service robot, even if an algorithm is correct, there are still a lot of parameters in the algorithm need to be tuned. Currently, manual tuning is the only solution, which costs a lot of time and is very low efficient. However, auto-tuning based on Machine Learning technology becomes possible if a sufficient amount of relevant accuracy data can be acquired. In this case, benchmarking supports research and development of service robots in a more direct and efficient way.

Based on these considerations, we have launched this long-term effort on synthetical benchmarking of service robots. At the first phase of this effort, we have done the following work. First, we have implemented a (semi-)automatic real-time evaluation system (ARES) for benchmarking DMP performance. The system includes a set of algorithms for collecting, recording and analyzing measurement data from a MoCap system [16,17], OptiTrack[2]. Second, we organized the Demo Challenge on Service Robots, a competition at RoboCup 2015. 25 teams applied and 11 of them were qualified for the competition. 10 qualified teams actually participated in the competition, in which our ARES was used for evaluating performance of the competing robots. Third, we organized a workshop on the same subject during the competition. About 100 participants from more than 10 countries attended the workshop.

The MoCap system (showed in Fig. 1) we used is an optical detection system with passive markers [18], which uses several fixed high speed cameras around the measurement area to triangulate a precise marker position. A set of markers (showed in Fig. 2) are attached on a robot, so that the robots behaviors can be captured by the MoCap system and then evaluated by our ARES real-time.

[2] http://www.optitrack.com/.

Fig. 1. The MoCap system

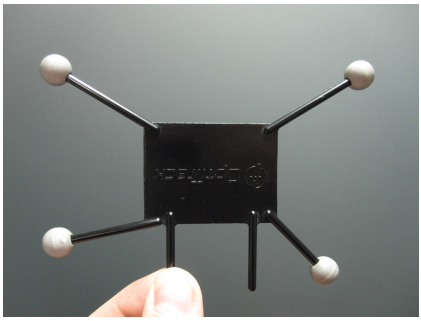


Fig. 2. The marker

3 Benchmarking on DMPs

In order to make benchmarks of DMP specific, an initial set of key features (hardware properties and functionalities) was derived from an analysis of DMPs and from experiences and observations of the common working scenes of DSRs. These features are evaluation criteria for the performance of DMPs. Furthermore, these features not only help design the benchmarks and the score system for the competition, but also allow for a later analysis of a team's performance. These features are divided into two groups: *hardware properties* and *functionalities*.

Hardware Properties. Taking into account DSRs' working environments and application demands, we propose *hardware properties* that must be implemented in each DMP in order to perform properly in the tests. To achieve these hardware properties, many technical details should be considered appropriately during mechanical designing and component selection progress. An appropriate trade-off is also needed between cost and the DMP's performance. The proposed hardware properties are characterized below.

Cost Limitation. Unlike pure theoretical research, one of goals of robotics research is to improve human life by bringing robust robotic technology to industry to create robotic applications. However, there is big gap between robotics research results and robotic products. Frequently robotics research pay more attention to verifying hypotheses and increasing knowledge, paying little attention to the cost and marketability of research outcomes. Thus, we insist that cost should be a important benchmarking condition.

Motor Feedback. Motor feedback captures the rotation angle of each wheel per control cycle, which is the source data to compute the odometry that is usually used as the input data in localization module. Besides, in the case of robot precise relative pose adjustment (e.g. mobile manipulation), odometry is the basis for adjusting robot pose, since global poses, generated by global localization techniques, are generally not as precise as odometry.

Payload Capacity. DMPs are expected to be extensible and customizable. Additional accessories, e.g. robotic arms and manipulators are expected to be integrated with DMPs to implement specific functionalities. According to the weight of common accessories, we believe that the payload capacity of DMPs should not be less than 20 kg.

Traversable Ability. Despite the fact that the floors of every-day environments are even, minor unevenness such as carpets, transitions in floor covering between different areas, and minor gaps (e.g. gaps between the floor and the elevator) are inevitable and also reflected in the RoboCup@Home competition. DMPs should be designed to adapt to these environment diversities.

Functionalities. The overall robotic system performance depends on the performance of integrated functional modules, which can be described as functional abilities or *functionalities*. As to DMPs, localization, navigation, and obstacle avoidance are the main functionalities.

Localization. The ability to estimate the real poses of a robot in the working environment.

Navigation. The task of path-planning and safely navigating to a specific target position in the environment.

Obstacle Avoidance. The ability to avoid collisions during a robot travel in the environment. Robots should be able to avoid not only static obstacles but also dynamic ones.

4 Implementation of the BSR Challenge

A robot qualified for the BSR challenge is expected to have a basic mobile platform (i.e., a robot base) and extended sensors such as camera, Laser Range Finder (LRF). The hardware cost of the basic mobile platform (including the costs of materials and components) or the market retail price (not discounted or second-hand price) should be less than 1,600 USD (about 10,000 RMB). The hardware cost or the market retail price (not discounted or second-hand price) of extended sensors should be less than 50 % of the basis mobile platform.

In order to enable the BSR challenge a synthetical benchmarking, we introduced a MoCap system to measure and, at the same time, record the movement of a DMP, with high accuracy in real time. The recorded data can not only enable quantitative analysis of the performance of DMPs in the competition, but also help make the DMP performance reproducible, which are taken as being utmost important to scientific experiments. After competitions, teams have free access to the record data.

The BSR challenge was organized in three tests and a presentation session. The mentioned key features are evaluated either as functional abilities, or as an

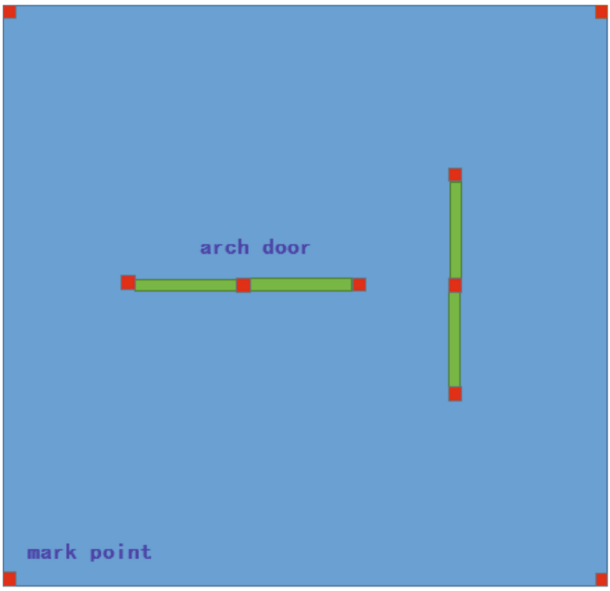


Fig. 3. The competition area (Color figure online)

integrated test. These tests and the score system are designed carefully ensuring each feature be contained in a test and be reflected in the final score. In the presentation session, each team was required to report its technical approach and share their experience with other teams.

4.1 Competition Area Layout

DMPs are tested in an indoor competition area (about 7 m × 7 m) where part of the ground may be uneven (within 3 cm of ups and downs) and there may be some obstacles on the floor. Obstacles include, but are not limited to: hollow obstacles (such as arches), furniture, small common objects, or even moving persons. Large obstacles such as arch and furniture are part of the field.

Figure 3 illustrates the setup of the competition area, where there are two sets of double arches (the greed blocks). The 10 landmarks (the red points in Fig. 3) are given as shown in Fig. 3. Among these landmarks, six are located on the arches and the other four are located in the corners. The coordinates of the landmarks in the MoCap system are provided for the participants to map the local coordinates of their robot to the coordinates of the MoCap system.

The double arch is shown in Fig. 4. The door width is 100 cm. There is a slider for each door, the height of which is adjusted randomly by the referee before a test in competition. A robot must decide autonomously whether it is able to go through a door according to its own height and the height of the slider on the door. In this case, the robot has to provide the capability of perception of 3D environment and reaction to dynamic environment. Besides, there is a plastic

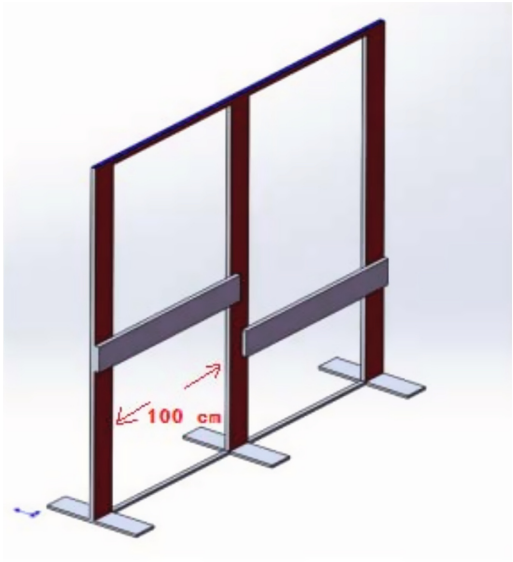
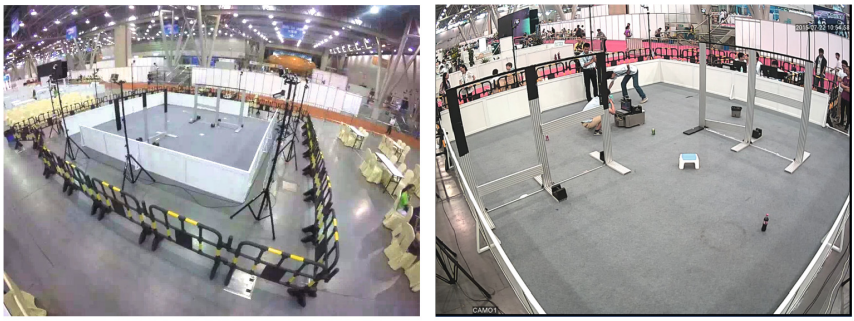


Fig. 4. The double arch.

bar (1.5 cm heights) at the bottom of each door. Robots go through a door may take a risk of being blocked by the bar, which is a trick to test their traversable ability.



(a) (b)

Fig. 5. Competition area

The MoCap system and four HD video cameras were installed for the competition, covering the whole competition area (showed in Fig. 5), by which robots' movement data and videos were recorded, in real time, from beginning to end.

4.2 Stage I Test

In stage I, robots are allowed to use only odometry as sensor. The robots are required to do two separate actions (moving in a straight line and turning at a given spot) under each payload condition: empty, 10 kg, 20 kg (showed in Fig. 1). Based on the feedback from the odometry of the robot and the measurement data collected by the MoCap system, the accuracy of the robot's movement for performing the tasks is computed. Each team is encouraged to try an extra payload once, which must exceed the maximum routine at least 20 kg, by given a bonus score. According to the movement errors measured by the MoCap system, each robot performance can be evaluated by being compared to the minimal error among all the teams under the same payload condition (Fig. 6).

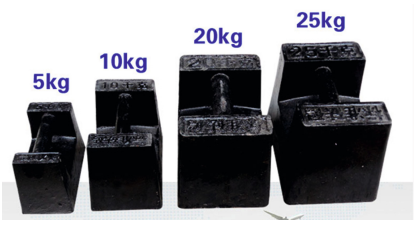


Fig. 6. Loads

Fig. 7. Obstacles

A final score for each team is computed by normalizing the scores of their performance under different load conditions with different score weights (showed in Table 1).

Table 1. The score weight under different load conditions.

Load	0 kg	10 kg	20 kg	≥40 kg
Score weight	0.2	0.3	0.5	0.2 (bonus)

4.3 Stage II Test

In stage II, a robot is allowed to use sensors besides the odometry to build a global map of the field before test. The map will also be used for evaluating the robot's performance. In the competition area, the robot is required to reach 7 way-points in the correct order (specified by the referee) under each payload condition: empty, 10 kg, 20 kg. The robot trajectory is recorded, and the distance between each way-point and robot stop point is measured by the MoCap system automatically. Before each team test, obstacles (showed in Fig. 7) in the field

and sliders on the arches are rearranged. A team gets punished when the robot colliding with obstacles or facilities in the field, and rewarded when the robot successfully passing through an arch.

The task in the final test is similar to that of stage II, but is more difficult, by adding more obstacles into the field and decreasing the maximum acceptable distance error.

5 Analysis of Team Performance

There were 10 qualified teams (partly showed in Fig. 8) participated in the BSR challenge. All the teams completed stage I test. Moreover, 7 teams could bear the payload of 40 kg. Table 2 shows the average and minimal motion errors. According to Table 2, the average motion errors (both distance and direction error) increases with the weight of the payload, which indicates that the payload has effect on the motion accuracy. Additionally, from the Table 2, we can see that some teams could achieve quite small motion errors under different payload conditions.

Fig. 8. Participating teams and robots

Table 2. Average and the minimal motion error in stage I

Load		0 kg	10 kg	20 kg
Distances error (mm per 1 m)	Average	3.7	4.96	9.88
	Minimal	0.5	0.25	0.25
	Best team	0.5	0.32	0.25
Direction error (degree per a round)	Average	7.18	7.97	8.31
	Minimal	0.35	0.47	0.42
	Best team	0.35	0.47	0.42

Since tests in stage II and final involved the localization and navigation abilities, perception sensors had great influence on the robot performance. Being limited by the cost restriction (1, 600 USD), robots can't be equipped with high price LRFs (e.g. Sick LMS100, HOKUYO URG-04LX, etc.). RPLIDARs[3], a kind of low-cost 360 degree LRF, and kinects were commonly used among teams. According to the sensor configurations, teams can be divided into three categories: teams with a low-cost LRF, teams with a kinect, and teams with a low-cost LRF and a kinect. As tasks in stage II and the final test were the same, their results are combined and analyzed according to the sensor configurations.

Table 3 presents the statistical results of stage II and the final test. From this table, it is evident that teams with a low-cost LRF got smaller motion errors and fewer collisions than teams only equipped with a kinect. This is because that the kinect provide depth data only in a limited distance interval (typically from 0.3 m to 5 m), however, the low-cost LRF can offer better observations in 2d point cloud. The Passing Door column of Table 3 shows that only teams equipped with both a low-cost LRF and a kinect could successfully make passing door actions.

Table 3. Statistical result in stage II and the final test

Sensor configuration	Number of teams	Average error (m)	Minimale error (m)	Collision (number of times)	Passing door (number of times)
Low-cost LRF	4	0.26	0.08	4	0
Kinect	3	0.47	0.29	7	0
Low-cost LRF Kinect	3	0.24	0.08	3	4

6 Discussion and Future Work

Our goal is to establish a set of synthetical benchmarks for DMPs, as a matter of fact, the BSR challenge had some limitations. Although, we proposed a set of key features of DMPs, these features can't cover every aspect of service robot benchmarks. In the future, we are going to broaden the scope of key features of DMPs, allowing more features (e.g. moving velocity, battery capacity) being evaluated. Moreover, the BSR challenge only evaluated the motion accuracy of a robot. More aspects such as time consumption will be included in the benchmarking scope, impelling teams to make trade-offs in these performance factors.

A comprehensive service robot benchmarking system contains three different levels: feature/ability benchmarking, subsystem benchmarking and system benchmarking. As an integrated system, the overall performance of a service robot not only depends on the performance of each single feature/ability, but also depends on the integration of single feathers/abilities and subsystems. But, only the feature/ability benchmarking was involved in the BSR challenge. More effort will be devoted to subsystem and overall system benchmarking.

[3] http://www.robopeak.com/.

The BSR challenge was a combination of benchmarking test and competition. However, ranking-oriented property of competition is a significant disadvantage for benchmarking. Attracted by the ranking, teams may develop solutions that converge to "local optimum" performance, by exploiting the vulnerability of rules. In the future, efforts need to be made both on organization and rules changes to overcome this drawback.

7 Conclusion

Robotic competitions play an important role in benchmarking robot systems, and hence provide a basis for this effort. However, most of existing benchmarking tools are qualitative, while in many cases quantitative evaluation is needed. Synthetical benchmarking consists of both qualitative and quantitative aspects, such as task completion, accuracy of task completions and efficiency of task completions, about performance of a robot. This paper presents our idea of introducing synthetical benchmarking into evaluation of service robots and a first realization of our synthetical benchmarking system on domestic mobile platforms. The system includes a set of algorithms for collecting, recording and analyzing measurement data from a MoCap system. We used the system as the evaluator in the BSR challenge, in which 10 teams participated. The competition was organized mainly as a comparative study on performance evaluation of domestic mobile platforms. An analysis of teams' performance is also given in the paper. Observations and future directions are made from the analysis.

Acknowledgments. A special thank is given to Intel China for its sponsorship of the BSR challenge and workshop. The authors from USTC are supported by the Natural Science Foundation of China under grants 60745002 and 61175057, as well as the USTC Key Direction Project. All authors are thankful for contributions from all collaborators who are not an author of the paper and all participants in the event.

References

1. del Pobil, A.P.: Why do we need benchmarks in robotics research? In: Proceedings of the Workshop on Benchmarks in Robotics Research, IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)
2. Nardi, L., Bodin, B., Zia, M.Z., Mawer, J., Nisbet, A., Kelly, P.H., Davison, A.J., Luján, M., O'Boyle, M.F., Riley, G., et al.: Introducing slambench, a performance and accuracy benchmarking methodology for slam. arXiv preprint arXiv:1410.2167 (2014)
3. Fontana, G., Matteucci, M., Sorrenti, D.G.: Rawseeds: building a benchmarking toolkit for autonomous robotics. In: Amigoni, F., Schiaffonati, V. (eds.) Methods and Experimental Techniques in Computer Engineering, pp. 55–68. Springer, New York (2014)
4. Behnke, S.: Robot competitions-ideal benchmarks for robotics research. In: Proceedings of IROS-2006 Workshop on Benchmarks in Robotics Research (2006)

5. Schultz, A.C.: The 2000 AAAI mobile robot competition and exhibition. AI Mag. **22**(1), 67 (2001)
6. Maxwell, B.A., Smart, W., Jacoff, A., Casper, J., Weiss, B., Scholtz, J., Yanco, H., Micire, M., Stroupe, A., Stormont, D., et al.: 2003 AAAI robot competition and exhibition. AI Mag. **25**(2), 68 (2004)
7. Balch, T., Yanco, H.: Ten years of the AAAI mobile robot competition and exhibition. AI Mag. **23**(1), 13 (2002)
8. Firby, R.J., Prokopowicz, P.N., Swain, M.J., Kahn, R.E., Franklin, D.: Programming CHIP for the IJCAI-95 robot competition. AI Mag. **17**(1), 71 (1996)
9. Pratt, G., Manzo, J.: The DARPA robotics challenge [competitions]. IEEE Rob. Autom. Mag. **20**(2), 10–12 (2013)
10. Wisspeintner, T., Van Der Zant, T., Iocchi, L., Schiffer, S.: Robocup@ home: scientific competition and benchmarking for domestic service robots. Interact. Stud. **10**(3), 392–426 (2009)
11. Wisspeintner, T., van der Zan, T., Iocchi, L., Schiffer, S.: RoboCup@Home: results in benchmarking domestic service robots. In: Baltes, J., Lagoudakis, M.G., Naruse, T., Ghidary, S.S. (eds.) RoboCup 2009. LNCS, vol. 5949, pp. 390–401. Springer, Heidelberg (2010)
12. Holz, D., Iocchi, L., van der Zant, T.: Benchmarking intelligent service robots through scientific competitions: the Robocup@ home approach. In: AAAI Spring Symposium: Designing Intelligent Robots (2013)
13. Stuckler, J., Holz, D., Behnke, S.: Demonstrating everyday manipulation skills in Robocup@ home. IEEE Rob. Autom. Mag. **19**(2), 34–42 (2012)
14. Amigoni, F., Bonarini, A., Fontana, G., Matteucci, M., Schiaffonati, V.: Benchmarking through competitions. In: European Robotics Forum-Workshop on Robot Competitions: Benchmarking, Technology Transfer, and Education (2013)
15. Ahmad, A., Awaad, I., Amigoni, F., Berghofer, J., Bischoff, R., Bonarini, A., Dwiputra, R., Fontana, G., Hegger, F., Hochgeschwender, N., et al.: Specification of general features of scenarios and robots for benchmarking through competitions. RoCKIn Deliverable D 1 (2013)
16. Corazza, S., Muendermann, L., Chaudhari, A., Demattio, T., Cobelli, C., Andriacchi, T.P.: A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. Ann. Biomed. Eng. **34**(6), 1019–1029 (2006)
17. Kurihara, K., Hoshino, S., Yamane, K., Nakamura, Y.: Optical motion capture system with pan-tilt camera tracking and realtime data processing. In: ICRA, pp. 1241–1248 (2002)
18. Field, M., Stirling, D., Naghdy, F., Pan, Z.: Motion capture in robotics review. In: 2009 IEEE International Conference on Control and Automation. ICCA 2009, pp. 1697–1702, December 2009