A Human-Robot Competition: Towards Evaluating Robots' Reasoning Abilities for HRI

Amit Kumar Pandey[1(✉)], Lavindra de Silva[2], and Rachid Alami[3]

[1] SoftBank Robotics, Innovation Department, Paris, France
akpandey@aldebaran.com
[2] Institute for Advanced Manufacturing, University of Nottingham, Nottingham, UK
lavindra.desilva@nottingham.ac.uk
[3] LAAS-CNRS, University of Toulouse, Toulouse, France
rachid.alami@laas.fr

Abstract. For effective Human-Robot Interaction (HRI), a robot should be human and human-environment aware. Perspective taking, effort analysis and affordance analysis are some of the core components in such human-centered reasoning. This paper is concerned with the need for benchmarking scenarios to assess the resultant intelligence, when such reasoning blocks function together. Despite the various competitions involving robots, there is a lack of approaches considering the human in their scenarios and in the reasoning processes, especially those targeting HRI. We present a game that is centered upon a human-robot competition, and motivate how our scenario, and the idea of a robot and a human competing, can serve as a benchmark test for both human-aware reasoning as well as inter-robot social intelligence. Based on subjective feedback from participants, we also provide some pointers and ingredients for evaluation matrices.

1 Introduction

Research in child development and human behavioral psychology clearly indicates that perspective taking, i.e., reasoning from others' perspectives, is one of the key components for social interaction and social intelligence. Perspective taking starts in children from as early as 12–15 months, in the form of understanding the occlusion of others' line-of-sight, and that an adult might be seeing something that the child is not able to see, due to it being hidden behind some barrier; this applies to both places and objects (e.g. [1]). Studies on reachability analysis (e.g. [2]) have suggested that from the age of 3, children are able to perceive which places are reachable to them and to others as they start to develop allocentrism—spatial decentration and perspective taking. In robotics, perspective taking has been used for learning from ambiguous demonstration [3], grounding ambiguous references [4], sharing attention [5], etc.

An object's affordance, specifically, its action possibilities (Gibson [6]) is another crucial aspect for shaping our day-to-day interaction with the environment and with others. Affordance is also a central organizing construct for

A. Agah et al. (Eds.): ICSR 2016, LNAI 9979, pp. 138–147, 2016.
DOI: 10.1007/978-3-319-47437-3_14

action differentiation and selection [7]. In robotics, the notion of affordance has been used in domains involving tool use [8], for checking traversability [9], for learning action selection [10], etc.

The Turing Test is a well-known test for evaluating the intelligence of a machine relative to that of a human. In the standard interpretation, the test involves a machine and human competing with each other over a conversation with another human. The design of our game was inspired by the Turing Test in the sense that a robot and human competes with each other to infer and describe environmental changes, which is judged by another human(s). However, some aspects of the Turing Test were not suitable for the kind of intelligence that we wanted to test, because the standard Turing Test *(i)* encourages making mistakes in order to look more natural and human like, and *(ii)* is more focused on carrying out a conversation. We do not advocate the robot making such mistakes, but instead identify and "penalize" them. Furthermore, because we focus on human-aware reasoning, we have based the scenario around physical changes in the environment. Another characteristic of our setting is that the ground truth is always available. Hence, one can derive two sets of evaluation criteria: *(a)* a "comparative" one, based on the competing human's and robot's reasoning abilities, and *(b)* an "absolute" one with respect to the ground truth.

Many related competitions exist in the literature, such as *Robocup* [11], the *DARPA Robotics Challenge (DRC)* [12], the *European Land-Robot Trial (ELROB)* [13], and the *HUMABOT robot competition* [14]. However, in all these applications, a robot must be created for a specific mission and target scenario, but potentially ones with no humans involved; therefore, these applications have no direct link to HRI. From the HRI literature, the *AAAI Challenge* [15] is relevant in that it proposes scenarios in which a robot attends and delivers a conference talk. Likewise, a variant of *Robocup*, called *Robocup@Home* [16], also seems relevant, as does the *RoCKIn* competition, which focuses on service robots in a real home environment. Like our proposal, the latter two competitions are also aimed at benchmarking robot systems: a set of benchmark tests is used to evaluate the robots' abilities and performance in realistic home environments. Thus, there is a clear need for HRI-oriented robot competitions, evaluation, and specialized benchmark tests. The competition that we present in this paper is another step in this direction: it provides a means by which the robot's "human-centered intelligence" (or "social intelligence") could be evaluated. More specifically, we present a competition scenario and methodology for its use, as well as an analysis of data gathered from the competition, which we believe will serve toward developing benchmark tests for evaluating a robot's combined intelligence based on perspective taking, reachability, and affordance analysis abilities. Our scenario is fully implemented on a PR2 robot, and it has been demonstrated live at an EU event, as well as to numerous visitors. Our preliminary work in this direction was presented in [17]. The current paper extends our earlier work with a detailed description of the methodology and framework; a subjective analysis of user feedback; and with pointers toward evaluation criteria, an evaluation matrix, and potential quantitative measures.

2 Competition Scenario

The scenario that we propose involves observing, analyzing, grounding, and explaining environmental changes. The setting for the scenario is a "living room" of a realistic apartment, with typical furniture (that was never moved) such as sofas, tables, and shelves, and movable objects including books, cans, and boxes.

Figure 1 summarizes the steps in the game, which are as follows. Two (human) volunteers h_1 and h_2 are asked to take a seat in the living room, and a third one h_3 is asked to stand next to the robot. Following this, the robot and h_3 inspect the living room from where they stand. Then, h_3 and the robot turn away from the scene, and h_1 and h_2 are asked to independently and/or cooperatively make manual changes to the state of the room (while the robot and h_3 are looking away). Finally, h_3 and the robot are asked to re-inspect the room, and identify any changes that might have been made. The competition concludes with a manual comparison of the responses of h_3 and the robot, with each other as well as with what h_1 and h_2 actually did—the ground truth—and by asking h_3 for his/her (subjective) assessment of the robot's "intelligence". Finally, a winner(s) between the robot and h_3 is determined (at this stage just for fun).

During the game, the robot and h_3 compete to answer the following questions: *What* has changed physically? *How* might those changes affect the agents' abilities to see and reach objects? *Who* might have done those changes and with *which* (possibly joint) actions? *Where* might any missing objects be?

A key requirement in the game, from both the robot and h_3, is the ability to *ground* changes in the environment. We define this process as follows. Given a couple $\langle s_0, s_1 \rangle$, which is respectively the initial and final states of the environment, find a suitable triple $\langle \Delta, E, A \rangle$, where Δ represents the physical changes in s_1 compared to s_0; E represents the effect of changes in Δ for h_1 and h_2; and A represents the probable sequence of (possibly cooperative) actions that were executed by h_1 and h_2 in order to bring about the change Δ. In this sense, the grounding process, among other things, needs to reason about perspectives, reachabilities, affordances, and action possibilities.

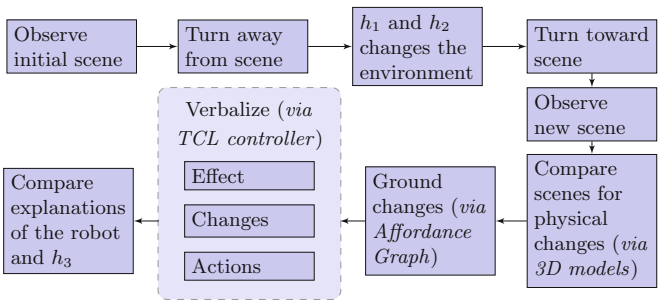


Fig. 1. Framework of the proposed Human-Robot competition between the robot and a human h_3, who observe the initial scene, and then ground the changes.

3 Instantiation

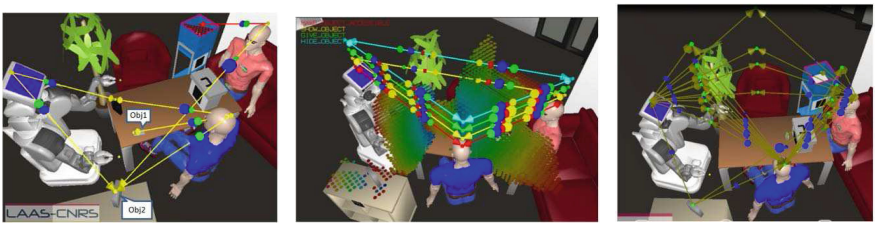
This section outlines the scientific and technical foundations on which the above reasoning capabilities were instantiated for the robot.

Scientific Foundations. We have integrated concepts from existing robotics frameworks that address issues related to perspective taking, affordance analysis, and effort analysis [18], where "effort" here is an abstraction of body-movement based effort levels for reaching objects and places, inspired by the taxonomy of reach affordances in human behavioral psychology [19]. The overarching notion that we exploit from these works is the *Affordance Graph*, which merges various kinds of human-aware reasoning into a single, unified graph. Figure 2c shows an affordance graph instantiated with respect to a specific environmental state. The graph enables the robot to reason about action possibilities among agents, and among objects distributed in the environment.

More specifically, the affordance graph is the aggregation of a *Manipulability Graph* and a *Taskability Graph*. A manipulability graph combines perspective taking, i.e., analyzing visibility and reachability of agents, with affordances between agents and objects, for the purpose of grounding symbolic notions such as grasping, picking, and placing objects to their corresponding geometric entities. Figure 2a shows an example of such a graph. The size of a sphere associated with an edge in the graph depicts the effort level required to see (Green sphere) and reach (Blue sphere) the object or place. Similarly, a *Taskability Graph*, shown in Fig. 2b, encodes relations between agents and other agents, in order to ground symbolic notions such as giving, showing, hiding, taking, and making an object accessible to their corresponding geometric entities.

Using an affordance graph as the underlying reasoning tool is appealing in that it could be analyzed using standard graph search algorithms, though in principle any other reasoning mechanism could be used. Multiple instances of the graph, such as one before and one after an environmental change, are constructed by the robot in order to infer and ground changes in the environment.

Technical Foundation. Our instantiation of the competition scenario is implemented within the LAAS robotics architecture [20]: an interconnected set of



(a) Manipulability Graph (b) Taskability Graph (c) Affordance Graph

Fig. 2. Different kinds of graphs representing affordances in the environment [18].

diverse components responsible for distinct functionalities within the system. The implementation uses Move3D [21] to represent the robot's version of the real world in 3D, which is used as input for geometric reasoning. The robot updates its 3D world state in real-time via various sensors; for example, a tag-based stereovision system is used for object identification and localization, and a Kinect (Microsoft) sensor for localizing and tracking humans.

Execution control is achieved via Tcl programs, which are grouped into three distinct sets of capabilities. The first performs tasks related to keeping the geometric 3D model of the world up to date. The second set of capabilities basically requests the geometric component to create an affordance graph for the current world state, and to compare it against the graph corresponding to the previously observed state (if any). The third set is responsible for natural human-robot interaction, i.e., for making speech more intelligible by synthesizing complete sentences out of the output generated by the geometric reasoning component, and for looking at relevant objects, places, and humans while speaking.

One example of the output produced by the geometric reasoning component is the set of couples $\{(object, gt), (action, mv)\}$, where the first element in each couple identifies whether the second element is an action or an object. This particular set is mapped to the sentence "*The grey-tape has been moved*", where the symbols *grey-tape* and *move* are obtained essentially via two user-supplied mapping functions f^{obj} and f^{act}, which respectively map object and action symbols used within the geometric component into the corresponding symbols used within the ("symbolic") execution controller; thus, $grey\text{-}tape = f^{obj}(gt)$ and $moved = f^{act}(mv)$. Similarly, plans (sequences of ground actions) found by the geometric reasoning component, such as $pick(h_2, gt) \cdot give(h_2, h_1, gt) \cdot place(h_2, p)$, where h_1 and h_2 are the volunteers and p is a new position, are mapped into sentences such as "*As for the grey tape, the second human picked it up and gave it to the first human, who then placed it at its current position*".

4 Observations Reported by the Competitors

In this section, we describe one run of a competition between the robot and a human, and in particular, the observations reported by the human competitor and the robot. The competition was carried out 12 times in total, while it was demonstrated live at an EU event, and to numerous visitors. Initially, all participants were briefed about the game, with information including the kind of data that was expected from the competitor. Later, the competitor was also guided by the cameraman, with questions such as "where was the object before it was moved?", "do you think that it was likely the object was moved jointly?", "who do you think moved the object?", and "where might the object be now" (if an object was reported missing), in order to gather as much relevant observations from the competitor as possible.

Figure 3 illustrates one run of the competition. Figure 3a shows the initial state of the environment that was examined by the robot and the human competitor before they were asked to look away. Figures 3(d) to (i) show one facet of

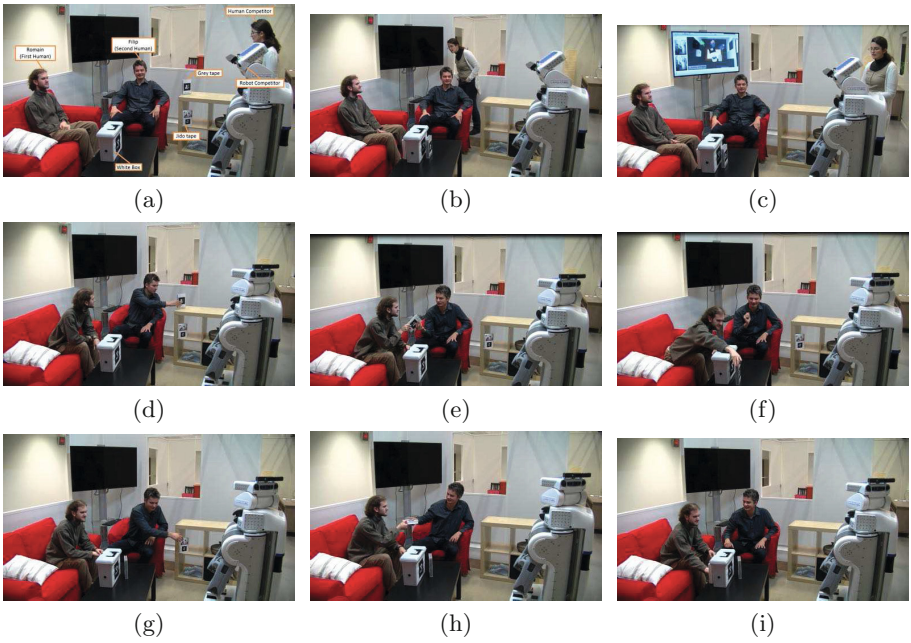


Fig. 3. Images (a) to (c) illustrate one run of the competition, and (d) to (i) show the sequence of changes that were made.

the ground truth—the sequence of steps that were discussed and jointly carried out by the other two humans in order to change the state of the environment. Figure 3b shows the competitor moving around to look behind a white box, where she suspects that an object is hidden. Figure 3c shows the robot and the human competitors describing their beliefs about what might have changed and their impact on the two sitting humans, in terms of what objects are now visible and reachable to them (or no longer so). The large flat panel display in this figure shows the current 3D environmental model maintained by the robot.

The key points that were made by the human competitor are listed below (after editing for clarity), along with our auxiliary comments inside parenthesis.

- *Initially, the grey tape was here, but it has now moved there.* (This was uttered while pointing to the correct initial and final locations of the object.)
- *I cannot see the Jido tape from where I am. It used to be there.* (This was uttered while pointing to the correct initial location of the object.)
- *The white box has not moved.*
- (She then moved to look behind the white box on the table and saw the missing tape, which she correctly suspected to be hidden there.)
- *The grey tape is no longer visible to Romain, and it is now visible to Filip; I also think that it is reachable to the robot.*
- *The Jido tape is neither visible to the robot nor to me, but I am not sure whether it is visible to Filip. The tape is both visible and reachable to Romain.*

– *I think that Filip took both tapes from here and handed them over to Romain, who then placed them on the table and rotated them.* (This was uttered while correctly pointing to the initial locations of the two objects.)

Next, we list the key points that were made by the robot (after editing for clarity) regarding the environmental changes that might have occurred, together with our auxiliary comments inside parenthesis.

– *The Jido tape has moved, and I cannot see it anymore.* (The robot also correctly guessed where the tape is hidden via mechanisms used by our framework [18], as outlined in Sect. 3.)
– *The grey tape has moved.*
– *Regarding the grey tape: the first human (Romain) will now find it more difficult to see it (compared to before).*
– *the first human can reach it now (although it was unreachable to him before).*
– *the second human (Filip) can now see it more easily (i.e., with less effort than before).*
– *the second human cannot reach it anymore (although it was reachable to him before).*
– *the robot can now see it more easily.*
– *the tape was picked up by the second human, given to the first human, and then placed by the first human.* (This was deduced using the geometric reasoner, which assumes that humans will try to balance the overall effort required for a joint task, whenever the amount of individual effort needed amounts to standing up from the seated position [18].)
– *Regarding the Jido tape: it was picked up by the second human and it was then placed.* (Like the human competitor, this was deduced based on the previous and current positions of the Jido tape.)

5 Subjective Analysis

We performed 12 runs of the competition, each of which took approximately 10 min. We noticed interesting similarities in the analyses performed by the robot and the human competitors h_3, e.g., the descriptions about how the objects might have been moved, and where an object, which was visible to them earlier, might now be hidden. There were also runs in which both competitors guessed incorrectly, or missed out on certain observations.

We also asked competitors for their (subjective) opinions about the robot's reasoning capabilities, with questions such as "*how was the robot's performance?*" and "*how was the robot as a competitor?*". Some of their key remarks were: *(1)* "*The robot showed* ***good interaction*** *with its environment and intelligence in its responses and behavior*"; *(2)* "*I was* ***better*** *than the robot*"; *(3)* "*It (the robot) performed* ***better***"; *(4)* "*The robot must have* ***cheated*** *through that reflection in the glass window*"; *(5)* "*It* ***guessed*** *(correctly)* ***most of the time***"; *(6)* "*We were* ***equally*** *good*"; *(7)* "*Oh, I* ***missed*** *that—what the robot said was correct*"; and *(8)* "*I think it has a* ***good*** *memory*". While still preliminary, these

comments show potential for the use of such competitions for evaluating a robot's reasoning abilities in HRI, particularly because competitors (from the public) tended to compare the robot's reasoning abilities with their own.

Data from multiple runs of the competition can be used to establish criteria for evaluating a robot's reasoning abilities, as well as derive evaluation matrices. For example, feedback from the 12 competitors suggested at least three evaluation criteria: *(i)* **adjectives** as **comparative measures**, such as "good", "better", and "equal", which competitors used to qualify the robot's reasoning abilities; *(ii)* **quantitative measures** such as the "number of times" an observation was correct, as competitors tended to use phrases such as "most of the time"; and *(iii)* **observations that were missed**, since competitors tended to use phrases such as "I missed that". In addition, we could take into account the ground truth, i.e., actual changes that were made by participants h_1 and h_2. Interestingly, our scenario always allows for the ground truth to be available.

The evaluation criteria established above can be further used to derive evaluation matrices, such as the one we derived in Fig. 4. This matrix relies on comparing observations reported by the human and the robot competitor against the ground truth, i.e., input from the participants who made the changes. The matrix qualifies the robot's intelligence relative to the human competitor as being highest (i.e., entry "$(High, High)$" in the matrix) when the human competitor guesses incorrectly and the robot guesses correctly, and as being lowest (i.e., entry "(Low, Low)") when the opposite happens. Missed observations are placed in the middle of the matrix and the agent who failed to observe the change is given the "benefit of the doubt", as such failures do not necessarily mean that the agent was incapable of making the correct deduction, nor that the agent has made an incorrect one. One can also come up with quantitative measures based on the matrix, e.g. $RI = \Sigma_{j=1}^{m}\Sigma_{i=1}^{n}((val_i^x + val_i^y))$, where $val_i^x, val_i^y \in [1, 3]$ (1 is the lowest) are the x and y axes values from Fig. 4, n is the number of environmental changes that needed to be observed in the game, and m is the number of competition runs. Hence, RI can indicate the "relative intelligence" of the robot. Below we illustrate interesting instances of some of the criteria in the evaluation matrix.

Incorrect deductions by the robot. The robot's deduction regarding how the *Jido* tape was moved was incorrect, whereas the human's was correct: *Filip* handed over the object to *Romain*, who then placed it on the table.

Incorrect deductions by the human. The human thought that the *grey* tape was not visible from the perspectives of the two seated participants, whereas in reality *Filip* was able to see it, which the robot deduced correctly.

Correct deductions by both. The robot correctly deduced the position of the missing object (*Jido tape*) as being behind the white box. This was done by analyzing the *Taskability Graph* for hiding an object, with the assumption that humans place objects on flat horizontal surfaces. The human also provided the same symbolic position description of the missing object.

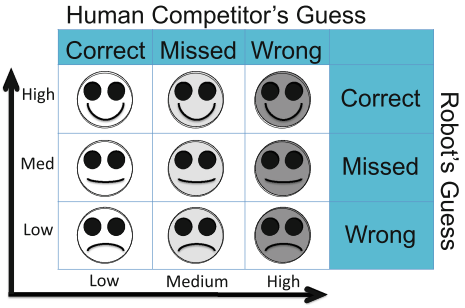


Fig. 4. A possible evaluation matrix for qualifying the robot's intelligence relative to the human competitor. Darker and happier faces denote higher intelligence.

Missed observations. Sometimes the human failed to notice when the object was moved by only a small amount, e.g. *5 cm*, whereas the robot was able to notice such small changes, because it stores a precise geometric model of the environment. On the other hand, in certain other runs, the robot failed to notice the presence of an object, because it had been placed in an orientation that prevented its tag from being detected by the robot. Consequently, the robot assumed that the object was hidden and tried to deduce its position.

6 Conclusion and Future Work

We have presented a novel human-robot competition scenario and methodology for evaluating the resulting relative intelligence of the robot (with respect to the human competitor), when certain basic building blocks of HRI reasoning, i.e., perspective taking, reachability, affordance, and effort analyses need to function together. This paper sets the stage for benchmarking and evaluation of such human-aware reasoning capabilities, which we think is now crucial. We cannot yet claim to have a conclusive set of evaluation metrics, nor do we want to impose one so early, as we believe that this has to be based on community feedback, and driven by extensive benchmarking competitions. We have only pointed out the feasibility of having such metrics, based on user studies and through the scenario presented in this paper. We have also extracted a set of criteria via the subjective evaluation and judgment of users, in order to stimulate interest in evaluation criteria for high-level intelligence.

In the near future, we aim to work in close collaboration with various competition organizers, e.g. RoboCup@Home, in order to enhance the proposed framework and the evaluation metrics, in the context of the proof of concept system presented here. It might also be interesting to develop a similar game for robot-robot competitions, to evaluate and compare the levels of human-aware reasoning capabilities of different robots, and thereby contribute to a standard competition for evaluating such capabilities, and a standard benchmark test for socially intelligent robots of the future.

References

1. Csibra, G., Volein, A.: Infants can infer the presence of hidden objects from referential gaze information. Br. J. Dev. Psychol. **26**(1), 1–11 (2008)
2. Rochat, P.: Perceived reachability for self and for others by 3 to 5-year old children and adults. J. Exp. Child Psychol. **59**, 317–333 (1995)
3. Breazeal, C., Berlin, M., Brooks, A.G., Gray, J., Thomaz, A.L.: Using perspective taking to learn from ambiguous demonstrations. Robot. Auton. Syst. **54**, 385–393 (2006)
4. Trafton, J.G., Schultz, A.C., Bugajska, M., Mintz, F.: Perspective-taking with robots: experiments and models. In: IEEE International Workshop on Robots and Human Interactive Communication (RO-MAN), pp. 580–584 (2005)
5. Marin-Urias, L., Sisbot, E., Pandey, A., Tadakuma, R., Alami, R.: Towards shared attention through geometric reasoning for human robot interaction. In: Proceedings of 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 331–336 (2009)
6. Gibson, J.J.: The theory of affordances. In: The Ecological Approach to Visual Perception, pp. 127–143. Psychology Press (1986)
7. Clark, A.: An embodied cognitive science? Trends. Cogn. Sci. **3**(9), 345–351 (1999)
8. Stoytchev, A.: Behavior-grounded representation of tool affordances. In: Proceedings of ICRA, pp. 3060–3065 (2005)
9. Ugur, E., Dogar, M.R., Cakmak, M., Sahin, E.: Curiosity-driven learning of traversability affordance on a mobile robot. In: IEEE International Conference on Development and Learning (ICDL), pp. 13–18, July 2007
10. Lopes, M., Melo, F.S., Montesano, L.: Affordance-based imitation learning in robots. In: Proceedings of IROS, pp. 1015–1021 (2007)
11. Robocup. http://www.robocup.org/
12. DARPA robotics challenge (DRC). http://www.theroboticschallenge.org/
13. European land-robot trial (elrob). http://www.elrob.org/
14. Humabot robot competition. http://www.irs.uji.es/humabot/
15. AAAI grand challenges. http://www.cs.utexas.edu/users/kuipers/AAAI-robot-challenge.html
16. Robocup-home. http://www.robocup.org/robocup-home/
17. Pandey, A.K., de Silva, L., Alami, R.: A novel concept of human-robot competition for evaluating a robot's reasoning capabilities in HRI. In: International Conference on Human-Robot Interaction (HRI), pp. 491–492 (2016)
18. Pandey, A.K., Alami, R.: Affordance graph: a framework to encodeperspective taking and effort based affordances for day-to-day human-robotinteraction. In: Proceedings of IROS, pp. 2180–2187 (2013)
19. Gardner, D.L., Mark, L.S., Ward, J.A., Edkins, H.: How do task characteristics affect the transitions between seated and standing reaches? Ecol. Psychol. **13**(4), 245–274 (2001)
20. Fleury, S., Herrb, M., Chatila, R.: Genom: a tool for the specification and the implementation of operating modules in a distributed robotarchitecture. In: Proceedings of IROS, pp. 842–848 (1997)
21. Simeon, T., Laumond, J.-P., Lamiraux, F.: Move3D: a generic platform for path planning. In: 4th International Symposium on Assembly and Task Planning, pp. 25–30 (2001)