Cognition-Oriented Building Blocks of Future Benchmark Scenarios for Humanoid Home Robots

Catherina R. Burghart
Institute of Process Control and Robotics
University of Karlsruhe
D-76128 Karlsruhe, Germany
burghart@ira.uka.de

Ralf Mikut
Forschungszentrum Karlsruhe GmbH
Institute of Applied Computer Science
P.O. Box 3640
D-76021 Karlsruhe, Germany
ralf.mikut@iai.fzk.de

Hartwig Holzapfel
Interactive Systems Lab.
University of Karlsruhe
D-76128 Karlsruhe, Germany
hartwig@ira.uka.de

Abstract

Comparing intelligent robots interacting with humans is a difficult task. First of all, robotic helps in everyday life can be involved in different tasks. Additionally, the grade of autonomy of the robot and the type of interaction between human and robot may differ to a great extent. It also is an intricate task to design common metrics to assess human-robot interaction. Thus normal benchmarks fail in this case. After an overview of already existing benchmarks in robotics this paper sketches affordances for benchmarks for human-robot interaction in a home environment. Many aspects of already existing benchmarks, evaluation criteria, and metrics have to be combined to asses human-robot interaction in a complex environment. The great difficulty is measuring human and robot behavior; here novel approaches taking into account perspectives from different fields (i.e. engineers, psychologists, sociologist) are needed, as a human being does not behave as anticipated by the system designer most of the time.

Motivation

The future will see robots interacting with humans in several parts of human everyday life: robots acting as tour guide or receptionist, bar keeper robots, household aids, assistive robotic devices for elderly people, and many more. These types of service robots usually are intended to interact with people who are not robot experts: intuitive handling and natural communication are important affordances for human-robot interfaces. This also implies an appropriate cognitive architecture of the robotic system in order to realize and coordinate all cognitive functions essential to intelligent systems. These functions comprise low and high-level perception, learning, memorizing, planning and problem solving, communication, and motor control. A major requirement for these types of intelligent systems is, that on the one hand they should be able to act autonomously. On the other hand they have to interact with humans, knowing a human's ways and habits.

Many international research groups focus on designing intelligent service robots. As everybody is looking for the ultimative cognitive architecture for their robot and as the implemented skills and abilities of service robots differ to a great extend, appropriate means to measure and compare the big variety of designs and implementations are needed. Just looking at building blocks, implemented cognitive functions, process flows, communication flows, and data flows, which all make up a designed architecture, does not really help to compare the abilities of different intelligent robots. This automatically leads to tasks to be accomplished by intelligent robots in order to compare and evaluate them. Two kinds of tasks are possible: a task specified as benchmark and a competition. The latter is used to determine the best intelligent robot out of a number of candidates, whereas a benchmark serves as testbed for each robot claimed to possess abilities needed to accomplish a defined benchmark. This paper concentrates on the affordances of benchmarks, specifically when human beings are involved. All benchmarks need quantitative metrics to measure the gradual success of different robots. A variety of metrics is suggested in this paper. The next section gives an overview of already existing benchmarks and competitions for robots and means to evaluate human-robot interaction. The following section stresses specific requirements and possible scenarios for intelligent robots interacting with people in a home environment.

Existing Benchmark and Competitions Scenarios

Many different benchmark and competition scenarios for robots have been developed for several years. Some recent examples with a great deal of public attention are robot soccer competitions in different leagues (Bredenfeld *et al.* 2006), test parcours for rescue robots (Jacoff, Missina, & Evans 2002), and the DARPA Grand Challenge (Thrun *et al.* 2006) for the autonomous driving of cars. Other benchmarks are oriented to special technical tasks as plug-in-hole benchmarks to handle control problems with closed kinematic chains. However, a transfer of such competition concepts and evaluation metrics to the household domain can cover only a part of the necessary evaluation scenarios.

So far, only few benchmarks for household environments have been proposed. The largest activity is the RoboCup@Home league funded in 2007. The rule framework was published in (Nardi *et al.* 2007). It is focused on autonomous service robots to assist humans in everyday life. The first competition in 2007 will take place in a un-

known home environment including a kitchen and a living room. Different concurrent teams will solve goal-oriented tasks like following and guiding a human, manipulation including open doors of rooms and of a fridge, navigation in the environment, imitating human movements with toy blocks, and localizing known but lost objects in a limited time. The necessary robot skills include the navigation in dynamic environments, the recognition of objects, humans, speech, gestures, and the manipulation of objects. The performance measurement is based on a score derived from competition rules and the evaluation by a jury.

A living room environment is also used by the authors of (Dautenhahn *et al.* 2006); a service robot performs fetch and carry tasks bringing a remote control to a person seated in an armchair. The robot approached the seated person from three different directions. The subjects had to state the direction and the distance of the robot they preferred. Questionnaires where used to record the user data. Additionally, the research team experimented with using a video set-up instead of a real life experimental robot-human interaction scenario. In case of the living room scenario, the video based method worked just as well (Woods *et al.* 2006). Besides preferred approaching directions of a robot Walters (Walters *et al.* 2007) also examined distances at which subjects still feel comfortable when interacting with a robot. By using floor scales and defining zones the comfortable distances could be measured in the experiments and taken from video records of the same experiments after the subject had given a stop signal. Human-Robot comfortable distances certainly can be used as a measurable criterion in human-robot interaction. These distances also play an important role during different encounters between human and robots in the human world.

A slightly modified real-world kitchen environment is used of the Karlsruhe Collaborative Research Center for Humanoid Robots (Dillmann, Becher, & Steinhaus 2004). The modifications are comparable to kitchens for handicapped persons, e.g. according to the height of cupboards. The robot performs tasks like navigation, learning real-world objects (e.g. cereal boxes), speech communication with people, and grasping objects (Asfour *et al.* 2006).

Other works cover a wider field of human robot interactions. An example is the DARPA/NSF interdisciplinary study on human-robot interaction (Burke *et al.* 2004) discussing scenarios of human-robot teams for military applications including a taxonomy for different interaction types.

Totally different is a robot interacting with autistic children. The social robot Infanoid (Kozima *et al.* 2004) was evaluated in tests with a set of 14 normally developed children and one child suffering from autism. In all experiments, the robot had the same social skills like eye-contact and joint attention with pointing toward an object. All tests were observed and analyzed: no metrics were used, solely the conditions were constant. By observation three different phases of interaction could be identified in all experimental runs.

In order to assess human-robot interaction the authors of (Dautenhahn & Werry 2002) propose an analysis of low-level criteria so-called micro-behaviors. They also experimented with autistic children interacting with a social robot. Comparative studies using a non-interactive toy similar in appearance to a robot and a socially interactive robot were performed. Besides basic actions, the robot was able to play interaction games employing turn-taking as key element with the children. Video tapes were taken of the experiments with the focus on the children's behavior. The videos were coded and fourteen criteria (behaviors of the children) were identified like eye gaze, eye contact, operation and handling, movements, speech, attention, repetitions and others. Mainly eye gaze behavior was used and compared between the subjects and their behavior toward the robot and the non-interactive toy. The researchers found out that a group of the children was more interested in the robot than in the toy truck, but they also stated, that other criteria besides eye gaze have to be considered as well.

The evaluation of all benchmarks requires a set of appropriate metrics (see e.g. the overview (Steinfeld *et al.* 2006)). Such metrics can be formulated on different levels of abstraction and cover quantitative objective measures as well as subjective measures. They are designed to evaluate autonomous robots, robots interacting with humans as bystanders (typical measure: success rate) or operators of robots in teleoperation scenarios (typical measure: the number of operator inventions).

Lampe et al. devised performance measures for autonomous mobile robots navigating in a corridor clotted by random obstacles (Lampe & Chatila 2006). Instead of comparing real robots they have developed a simulation system allowing the application of several algorithms for navigation in an environment as well as modeling different robot types equipped with sensors. The number and type of sensors applied can also be varied by parameters. With the help of their simulation the authors can measure robot capacity and autonomy with respect to a given task and related environment complexity. Metrics for performance, world complexity and information quantification were established. In the first category instantaneous velocity, traveled distance, mission duration, mission success rate and power usage were measured, whereas global complexity and the vicinity of the robot are taken into account in the second category. The last metric used is the conditional entropy measuring the information contained in the robot map compared to the world map.

Metrics for human-robot interaction have been devised from an psychologist's point of view by the authors of (Kahn *et al.* 2006). They suggest six psychological benchmarks to be considered when evaluating robots interacting with people: autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. These contenders are attributed to a robot by the person interacting with it. The list of attributes is still open to further expansion.

Another approach from an engineer's perspective uses coding of real-time social events (Nabe *et al.* 2006). The authors analyzed the behavior of children interacting with a communicative robot. Information about the children's way to use the robot was found out with the help of behavior-level codes, which described the children's adjustment to the setting, not their reaction. The analysis showed that children use the robot as a social mediator to initiate and drive syn-

chronized social events. These criteria were recorded during a field trial of a communicative humanoid robot interacting with children in a primary school for a period of two months (Kanda *et al.* 2004).

In contrast to evaluation metrics and benchmarks on human-robot interaction, works and benchmarks on human-machine interaction do exist: there are various world-spanning evaluations for perceptual components especially based on speech recognition and computer vision. Current evaluation tasks are the rich transcription (RT) workshops (RT 2007) for meeting recognition in two tasks, the conference room and the lecture room, which evaluates speech processing and includes speech to text recognition, speaker diarization (who spoke when) and speech activity detection (RT-06); CLEAR (Stiefelhagen and Garofolo 2007) workshops for multimodal detection and classification of events and activities, which include person tracking (vision-only, audio-only and multimodal), face tracking, vehicle tracking, person identification (vision-only, audio-only and multimodal), head pose estimation and acoustic event detection and classification.

These evaluation benchmarks have in common that pre-recorded data is used on which each participant can train and adapt its recognition component and evaluate on held-out evaluation data, and by using the same metrics and the same data, components from different sites can be compared against each other. A similar approach is difficult for dialog management. Here, the problem exists that pre-recorded data is not suitable for full dialog systems, where different decisions of the dialog strategy no longer match a pre-recorded dialog. Also, dialog systems are hard to compare since most systems are domain-specific. Despite these problems, approaches exist to build a unified framework for the evaluation of dialog systems and create comparable scores with the PARADISE framework (Walker *et al.* 1997) for spoken dialog systems. When adding multimodal capabilities to the system evaluation gets more complex and requires adapted evaluation schemata (Beringer *et al.* 2002). The biggest spoken dialog systems evaluation including different sites is the Communication evaluations 2000 and 2001. There, a common task, travel information, was selected, and several systems were evaluated by test subjects.

So far many ideas, methodologies, metrics, and measurement criteria do exist in order to assess human-robot interaction, but none of them consider complex interaction scenarios between persons and robots in a home environment in total.

Benchmarks for useful activities for Humanoid Robots in home environment and human-robot interaction

A main criterion for a successful benchmark scenario for humanoid robots in a home environment is to test the cognitive abilities of the robot in a fast changing dynamic environment. The benchmark should integrate many existing ideas from scenarios like RoboCup@home. In the following section, we try to identify different complexity levels for technical problems and cognition-related problems to improve the ability of the robot to gain acceptance as a personal home robot.

Levels of complexity

As in RoboCup@home and similar works, the scenario includes navigation, manipulation, exploration, and social interaction with humans and (optionally) other robots in a home environment. Here, different levels of complexity should be concerned:

- In Level 0, the robot shows only the ability to perform behavior where all tasks are learned by teaching or programming. The task structure is completely known and uncertainty about the environment is rather small (e.g. unknown positions of known objects in a small range). The objects usually have robot-friendly colors and shapes to simplify recognition tasks. A human is not directly involved. Such a level is the state-of-the-art for industrial robots.
- In Level 1, the degree of uncertainty is moderately higher. The robot is able to communicate with one known human operator by understanding a small set of known commands in natural language. After the detection of the command, associated skills each with a complexity similar to Level 0 are performed. Nevertheless, this level requires a more intelligent robot behavior (e.g. error detection and handling). Most actual research demonstrators for service robots are mainly situated on this level of complexity with some additional skills from the next level.
- For Level 2, a higher complexity is required to make the step to useful activities in real-world environments. This step forces many technical, communication, and cognition challenges. Actually, these different directions can not be strictly separated due to many overlapping skills:
  - Many researchers of the robot community especially focus on different technical challenges. In the following, only few examples from a much larger set will be given: For robot vision as key technology for navigation and manipulation, such challenges include e.g. the robust recognition of the environment in different lightning conditions, algorithms to recognize partially overlapping objects, subject and object detection in fast changing scenes, sensor fusion algorithms, objects in mirrors, transparent objects etc. Navigation needs better solutions for simultaneous localization and mapping (SLAM) to handle partially unknown environments (no explicit maps), sensor fusion, and path planning. Manipulation requires e.g. handling of deformable objects, difficult geometric environments like high racks, cupboards, or a fridge, the necessity of replacing other objects to reach the desired objects, resp. robustness aspects to handle liquids and dirt.

    The handling of such technical challenges requires numerous and sophisticated cognitive capabilities of the robot. Examples are a robust classification if an object is an obstacle or not, the performance of learned grasp skills if an expected object is missing, too many objects exist, or similar objects with a different functionality are in a region of interest (e.g. a dice box instead

of the expected coffee cup). This will remain an important field of research for many years. Unfortunately, the challenges resulting from human-robot interaction are much more complex in comparison to these technical challenges.

- Interaction between robot and human being on this level is mainly based on communication: verbal and non-verbal. Future interaction scenarios are intended for naive persons, thus the robot's active ability to converse and understand a naive person as well as its technical capabilities to overcome any acoustic diversions are of great importance. An increasing number of works focus on these challenges.
  - Active: Level 2 robots are able to communicate with natural language, which is still domain-dependent but users should be able to talk to the system naturally without learning a fixed set of commands. Therefore, the system needs to resolve ambiguous input within context, detect errors like misunderstandings or missing words, and be able to resolve these errors with clarification questions or learning dialogs to acquire missing information like unknown words, new objects or new semantic concepts. Other modalities than speech can be used, e.g. deictic or iconic gestures, and the robot should to some extend be able to observe the emotional state of the user to adapt its strategies to these conditions (empathetic agents).
  - Passive (overcoming diversions): While Level 1 allows one user and predefined commands, Level 2 allows multiple persons, who either talk to each other or directly to the robot, which requires focus-of-attention. Talking to the robot is not restricted to using a hand-held or headset microphone, rather distant speech microphones are used. This in turn imposes new challenges on speech recognition like background noise, reverberation and speech segmentation. The challenge to detect if someone talks to the robot is harder and background noise, speech from a radio or TV, as well as other persons talking to each other need to be filtered out.
- Considering semantic knowledge about persons interacting with a robot is essential for Level 2 robots. Besides adapting their speech recognition system to a specific person, they should be able to instantiate required and planned tasks with knowledge about the interacting person i.e. their preferences in various contexts or their rights to use the robot. Some researchers already dedicate their work to developing adaptation mechanisms for robots on basis of monitored human reactions or on basis of estimated models of friendship between several persons.
- Learning of unknown tasks in unknown or unstructured environments has also been a field of research for the last years. A Level 2 robot should be able to learn to perform specific motions and actions by imitating a person demonstrating these movements. Robot specific characteristics are considered when transferring human motions to robots. Ongoing research also focuses on a person acting as a tutor for a robotic system in order to teach complex actions on-line. The teaching should also involve a feedback of the robotic system to the human tutor.
- Close interaction with people also requires Level 2 robots to respect a person's feeling of comfort regarding spatial distances between robot and person. People interacting with each other usually respect a social distance when approaching or passing each other as well as when using arms and hands close to each other. Level 2 robots are required to keep these distances; some researchers have already been working on these problems for robots approaching seated persons.
- Level 2 robots should know simple social cues when interacting with people. Besides joint attention, recognizing the emotional state of the interacting person and signaling using mimics or posture this involves the robot behaving according to the unwritten rules existing in interaction between humans.
- The robot's self-awareness is a prerequisite for interacting with people, for respecting a person's level of comfort and social cues, for self-adaptation and for learning. Some researchers already take into account different aspects of self-awareness starting with internal states and success rates up to an evaluation of human behavior. The latter still cannot be done automatically by a robotic system.

- Steps to higher levels: Higher levels involve even higher complexity of interaction scenarios, environment, and required cognitive capabilities of the robot. First of all, language recognition should not be restricted to a given domain any more, as talking about or referring to topics outside an intended domain easily happens even in restricted domains i.e. home environment. A complex environment also involves different types of people and even animals coming into contact with the robotic system. The robot should then know how to react to and interact with different types of people i.e. toddlers, school children, adults and handicapped people.

  With increasing complexity meta cognition plays an important role. Based on experiences made, evaluation of behavior (both robot and human), task success rates, knowledge about people and environment, task knowledge, and situative knowledge the robot should be able to deduce essential knowledge or prospective, required actions (including self-adaptation) and plan them accordingly. Another step toward higher-level robots is their ability to create / plan new actions out of already existing patterns, to generalize action patterns and finally to transfer action patterns to further domains.

There actually is no limit of steps to higher levels: they do not only depend on technical progress but also on ethical issues and the question which capabilities we want to endow a robot with in future. Steps to higher levels have to be regarded as long-term goals whereas categories for human-robot interaction for Level 2 robots should be met in order to enable non-frustrating interactions between naive persons and robots in various contexts.

Evaluation metrics for Level 2 robots

Evaluation metrics for a Level 2 robot are much more complicated than for Level 0 or Level 1 robots. Level 2 robots must be able to adapt their own behavior to a dynamic changing environment with different communication partners and situations. They need cognitive abilities to recognize, compare, and understand situations, to detect unknown situations, to make decisions and select appropriate actions under uncertainties, and to learn from past experiences. The robot behavior strongly depends on existing situations and changes as a consequence of a continuous learning process.

A typical scenario for Level 2 robots are household robots interacting as "personal robots" for one person or a limited number of family members. It addition, they will come in contact with different guests or pets.

To test the behavior, appropriate test scenarios are necessary. To cope the uncertainties resulting from the complex environment, the tests scenarios should be designed as a series of tests on different test trials.

First of all, all interacting persons should be completely unknown for the robot at the first meeting. In addition, the subjects should not be prepared about the robot's capabilities. It is a natural situation for the first "getting in contact" with personal robots or meeting unknown guests during the whole lifetime of a robot. Such meetings should be performed for a higher number of different subjects (e.g. age, sex, profession) to catch the bandwidth of human communication behaviors. Such tests evaluate the robustness of the robot to communicate with different subjects characterized by completely different expectations about the robot's behavior. In all situations, a rough and noisy environment (e.g. radio, TV, additional people in the background) is useful. The robot operators should not be visible to avoid an influence by tips and hints to the interacting persons.

In addition, repetition trials with the same persons at different days can imitate an alternating learning process between human and robot. Here, the robots ability to exploit the knowledge about a special subject to adapt its own behavior is tested.

The known metrics discussed above can be classified in different categories:

- materials for evaluation (observation of the robot, internal robot log-files, questionnaires for interacting subjects, interviews with subjects, protocols, video tapes, different video perspectives, coding, video files from inborn robot cameras, ...),
- quantitative or qualitative metrics,
- metrics for different trials and tasks, (all trials, all trials with one subject, one trial with one subject, one task in one trial,...),
- objective metrics
  - autonomous self-evaluation on-board of the robot,
  - using external sensors and/or information processing,
- subjective measures (e.g. questionnaires)
  - by a jury,
  - by interacting persons,
- evaluation in realtime or not.

Quantitative metrics evaluate e.g. success rates for underlying skills like navigation (e.g. possible speed, deviation from the planned route, number of touched obstacles), speech recognition and generation (e.g. number of successful dialogs), manipulation (e.g. number of successful handled objects). They aggregate measures from underlying low-level motor and perceptions skills with related measures such as tracking errors to planned joint trajectories, control errors, and stability margins of basic control loops, maximal movement speeds for robot joints, the number of not-identified objects, localization accuracy of objects according distance, height, number of recognized words, delay time for answers etc. Some of these measures are useful for an self-evaluation of the robot to improve its abilities for adaptation, fault detection, and safety-related supervision.

A guiding principle to design quantitative on-board metrics is the evaluation of differences between the expected and actual behavior of robot and environment. These metrics are good measures for the ability of the robot to predict future situations based on its own experiences.

Metrics for human-robot interaction comprise objective as well as subjective criteria. Objective metrics are measurements or success rates of different kinds most of which can be partially performed automatically and need no further processing steps in between by a person. Subjective criteria vary according to the people involved. Test subjects have different notions about the success or failure of an interaction as well as the reasons for failure. Every test subject builds up a different internal mental picture of the robot's capabilities. Additionally, it is simply impossible for a system designer to predict human behavior in an robot interaction context in all facets. As soon as additional people are involved in coding human behavior, whether actions, reactions or adjustments from video data, different personal views enter the coding. Actually, the same video sequence should be coded by at least two different people.

The following measures are an example for assessing human-robot interaction, the list can still be expanded.

- objective measures
  - WER: word error rate (THE standard ASR measure),
  - CER: concept error rate (error rate to measure understanding, based on recognized concepts),
  - TER: turn error rate (number of turn that cannot be transformed to the correct semantics, including interpretation in context),
  - size of vocabulary,
  - object learning rates: detection of unknown objects, efficiency of learning dialogs, number of clarification dialogs, number of learning dialogs necessary to learn one object and re-recognition accuracy for objects,
  - success rates for introduction of new words for known and unknown objects in a dialog,
  - person learning rates: number of people successfully learned from a fixed set (vision and acoustic),
  - efficiency of learning dialogs: number of successful/unsuccessful dialogs, length of dialogs,

  - re-recognition rates of new words,
  - comparison of robot log-files for speech recognition and transcribed video data,
  - human awareness (Steinfeld *et al.* 2006) (comparing robot log-files and video data),
  - autonomous recognition of interaction failures.
- subjective
  - questionnaires / interviews, the outcome of which depends on the kind of questions, scales, and the interviewer,
  - video analysis, which requires video transcription. Due to differences in coding, several people should code the same sequences,
  - coding of behavior (Nabe *et al.* 2006), (Dautenhahn & Werry 2002),
- objective assessment based on coded subjective criteria (Burghart *et al.* 2005)
  - criteria describing the interaction context (interaction patterns, interaction rules, roles, degree of freedom),
  - criteria describing the interaction itself (intensity, congruity, convergence, synergy, efficiency),
  - criteria describing the activity of actors (transparency, roles),
  - criteria describing non-verbal actions and emotions (mimics, postures, gestures, affects),
  - specific coding of micro behaviors.

Subjective measures like questionnaires, interviews, stories, protocols taken by team members, and video data are hard to quantify and qualify. Usually, there are no measurable scales on comfort, acceptance, or other criteria, as each test subject has his or her own scale as well as his or her individual image of the interacting robot. Additionally, video data, questionnaires, and log-files have to be compared, as statements given in questionnaires can contradict the actual behavior of a test subject recorded on video. All subjective measures need additional people to either process the data i.e. transcription of video files or coding or to analyze and interpret acquired data. Here again, personal opinions of the evaluators can impair results. Once either categories have been assigned to individual test runs and turns between a person and a robot or behavior has been coded, these codes can serve as metrics, as coded categories and behaviors form characteristic patterns (Burghart *et al.* 2005; Dautenhahn & Werry 2002; Nabe *et al.* 2006). Then different test runs with the same subject as well as test runs with different subjects are much easier to assess and compare.

Overall, metrics for all trials are success rates or probabilities for complex tasks and the related performance times, whether they can be assigned definite values or patterns of coded categories and behaviors.

All listed measures above are useful to evaluate some facets of the robot behavior. It will be impossible to aggregate all measures to an overall performance measure due to their different natures, their requirements in further processing and interpretation and their partially non-automatic acquisition. An alternative will be a multi-modal evaluation leading to scores for Pareto-optimal robot behaviors with advantages for different sub-measures.

Conclusion

Evaluating intelligent robots interacting with people in a home environment proves to be a difficult task. So far, robot evaluation mainly focuses on sheer technical robot skills. As shown above, common metrics applicable to manipulation tasks, navigation of a robot and object recognition do not suffice. Social-oriented and object-oriented metrics have to be brought together in complex benchmark scenarios for human-robot interaction in a home environment. In order to assess autonomous robot capabilities, cognitive robot skills required in the human world and interaction with naive persons benchmarks increasing in complexity and integrating required robot skills of different levels, as described in this paper, have to be designed. In this article we could also show, that a big set of various metrics already do exist, which have not been combined in a benchmark scenario so far. A large set of different metrics is needed to evaluate Pareto-optimality for different subgoals of the benchmarks; an overall measure cannot be established due to the varying nature of measurements and due to the impossibility to quantify some of the applied measurements. The latter are mainly measurements describing human-robot interaction. Here, problems arise as getting data on human-robot interaction for assessment requires a large set of naive persons as test subjects, different test runs per person (first with unknown contact), and experimental conditions avoiding outside influence on test subjects. Also a big team is needed to transcribe, code, analyze and interpret recorded data of these experimental runs. Once coding of human behaviors and interaction categories is done, these codes can serve as metrics for the evaluation of human-robot interaction as well. If coding can be automated in future due to better cognitive skills of the robotic system enabling a robot to identify human behavior and adapt its own behavior to human reactions, it will be possible to easily create combined objective metrics evaluating robot performance in human-robot benchmark scenarios. Till then, subjective metrics are an important part when evaluating robot performance in human-robot interaction scenarios, which cannot be neglected.

Acknowledgements

This work has been performed within the framework of the German humanoid robotics program (SFB 588) funded by the German Research Foundation (*DFG: Deutsche Forschungsgemeinschaft*).

References

Asfour, T.; Regenstein, K.; Azad, P.; Schröder, J.; and Dillmann, R. 2006. ARMAR-III: A humanoid platform for perception-action integration. In *Proc., International Workshop on Human-Centered Robotic Systems (HCRS), Munich*. 51–56.

Bredenfeld, A.; Jacoff, A.; Noda, I.; and Takahashi, Y., eds. 2006. *RoboCup 2005: Robot Soccer World Cup IX*. Springer, Lecture Notes in Computer Science.

Burke, L.; Murphy, R.; Rogers, E.; Lumelsky, V.; and Scholtz, J. 2004. Final report for the DARPA/NSF interdisciplinary study on human-robot interaction. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews* 34(2):103–112.

Dautenhahn, K., and Werry, I. 2002. A quantitative technique for analysing robot-human interactions. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1132 – 1138.

Dillmann, R.; Becher, R.; and Steinhaus, P. 2004. ARMAR II - a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics* 1(1):143–155.

Jacoff, A.; Missina, E.; and Evans, J. 2002. Performance evaluation of autonomous mobile robots. *Industrial Robot: An International Journal* 29(3):259–267.

Lampe, A., and Chatila, R. 2006. Performance measure for the evaluation of mobile robot autonomy. In *Proc, IEEE International Conference on Robotics and Automation (ICRA'06)*. 4057–4062.

Nardi, D. et al. 2007. *RoboCup@Home: Rules and Regulations (Draft, Version 1.0, Revision 16)*. RoboCup Federation.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common metrics for human-robot interaction. In *Proc., ACM SIGCHI/SIGART Human-Robot Interaction*. ACM Press New York, NY, USA. 33–40.

Thrun, S. et al. 2006. Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics* 23(9):661–692.

Burghart, C., and Haeussling, R. 2005. Evaluation Criteria for Human Robot Interaction. In *Proceedings of the AISB'05 Conference, Workshop on Robot Companions*. Hatfield, Hertfordshire, England.

Dautenhahn, K. et al. 2006. How may I serve you? A robot companion approaching a seated person in a helping context. In *Proc. of the ACM International Conference on Human Robot Interaction HRI 06*. Salt lake City, Utah, USA.

Walters, M. L. et al. 2007. Robotic Etiquette: Results from User Studies Involving a Fetch and Carry Task. In *Proc. 2nd ACM/IEEE International Conference on Human-Robot Interaction*. Washington DC, USA,.

Woods, S. et al. 2006. Comparing Human Robot Interaction Scenarios Using Live and Video Based Methods: Towards a Novel Methodological Approach. In *Proc. of International Workshop on Advanced Motion Control* . Istanbul, Turkey.

Nabe, S. et al. 2006. Robots as social mediators: coding for engineering. In *Proc. of the International Symposium on Robot and Human Interactive Communication*. Hatfield, UK.

Kahn, P. et al. 2006. What is a Human? - Toward Psychological Benchmarks in the Field of Human-Robot Interaction. In *Proc. of the International Symposium on Robot and Human Interactive Communication*. Hatfield, UK.

Kanda, T. et al. 2004. Interactive robots as social partners and peer tutors for children: a field trial. Vol.19, No.1-2, pp. 61-84, 2004.

Kozima, H. et al. 2004. A Humanoid in Company with Children. In *Proc. of the International Conf. on Humanoid Robots*. Los Angeles, Usa.

RT'07: The Rich Transcription Workshop 2007. http://www.nist.gov/speech/tests/rt/rt2007/

Stiefelhagen R., Garofolo J. (Eds) 2007. *Multimodal Technologies for Perception of Humans*. Proceedings of the First International evaluation workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Springer Lecture Notes in Computer Science No. 4122., January 2007.

Walker M.A.; D.J. Litman; C.A. Kamm; and A. Abella 2002. *Paradise: A framework for evaluation spoken dialogue agents*. Annual Meeting of the Association of Computational Linguistics. ACL.

Beringer, N.; U. Kartal; K. Louka; F. Schiel; U. Türk 2002. *PROMISE - A Procedure for Multimodal Interactive System Evaluation*. in Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation'. Las Palmas, Gran Canaria, Spain.