Applications in HHI – Physical Cooperation

Markus Rickert and Andre Gaschler and Alois Knoll

Abstract Humans critically depend on permanent verbal and non-verbal interaction—for aligning their mental states, for synchronizing their intentions and goals, but also for performing joint tasks, such as carrying a heavy object together, manipulation of objects in a common workspace, or handing over components and building or assembling larger structures in teams. Typically, physical interaction is initiated by a short joint planning dialog and then further accompanied by a stream of verbal utterances. For obtaining a smooth interaction flow in a given situation, humans typically use all their communication modalities and senses, and this often happens even unconsciously. As we move toward the introduction of robotic co-workers that serve humans—some of them will be humanoids, others will be of a different shape—humans will expect them to be integrated into the execution of the task at hand, just as well as if a human co-worker were involved. Such a flawless replacement will only be possible if these robots provide a number of basic action primitives, for example, hand-over from human to robot and vice versa. The robots must also recognize and anticipate the intention of the human by analyzing and understanding the scene as far as necessary for jointly working on the task. Most importantly, the robotic co-worker must be able to carry on a verbal and non-verbal dialog with the human partner, in parallel with and relating to the physical interaction process. In this chapter, we give an overview of the ingredients of an integrated physical interaction scenario. This includes methods to plan activities, to produce safe and human-interpretable motion, to interact through multimodal communication, to schedule actions for a joint task, and to align and synchronize the interaction by understanding human intentions. We summarize the state of the art in physical human-humanoid interaction systems and conclude by presenting three humanoid systems as case studies.

Key words: Conversational Dialog, Human-Robot Cooperation, Humanoid System Architectures, Joint Action, Multi-Sensor Fusion

1 Introduction

For their survival, humans depend on communicating with each other (*one cannot not communicate*), but they just as strongly depend on modes of interaction, i.e., doing things together, either in direct physical contact, through the exchange of words (and expressions in other modalities) that change each other's state of mind, and by doing both in parallel. Through this interaction perception, decision-making, and production of behaviors, humans are tuned to their peers with whom they synchronize and share believes, desires, and intentions. It is highly desirable to transfer the concepts of interaction and *joint action* as cooperation metaphors when developing cooperating robots and in particular when it comes to the development of robots working together with humans.

It is easy to predict that the development of techniques for effective and efficient joint action based on multimodal communication flows between humans and artifacts will be of utmost importance for the advancement of service robotics as the whole. Once this development gains enough momentum, the requirements for an ever-increasing responsiveness of robots, of wider applicability to new scenarios, situations, and object domains, and of an easy integration of the robots into scenarios with many cooperating robots and many cooperating humans will grow quickly. There have been various attempts to design robots that directly interact with humans—for the purpose of *programming by demonstration* [8], for controlling their behavior within certain limits [46], or for *force amplification* [43]. All of these have shown that the development of truly interactive robots that combine multimodal communications with physical interaction is a very complex matter. It depends very much on progress in various fields, which is why to this date only laboratory samples of systems exist that implement individual aspects needed for smooth interaction over long time horizons.

In a typical setting, human instructions are perceived by the robot in just one modality, e.g., through a camera system. This precludes the system from constructing cross-modal associations by evaluating clues from other modalities (audition, touch, etc.). It also prevents humans from giving additional explanations in *natural* modalities, e.g., teaching robot hand movements supplemented by instructive speech statements. Partly due to mono-modality, the communication flow for enabling joint action is not in the form of a dialog between human and robot. Dialog-oriented interaction is often very useful because it is the source of additional information, but it becomes indispensable in the case of error conditions. Furthermore, the aspect of physical cooperation for supporting instructing the robot in parallel through corresponding utterance in several modalities has hardly been addressed so far. Hence, this chapter will review recent implementations of human-humanoid interaction methods. To allow a better insight into the practically realized methods, this chapter will review some of the underpinning questions of physical human-human cooperation and communication, i.e., dynamic arm handover characteristics and multimodal gesture communication. This requires careful transfer into respective technologically realized humanoid systems. Thus, a detailed review of recent projects in human-humanoid interaction will address step-by-step some of the is-

sues mentioned above. Among the projects reviewed are MORPHA [49], Robonaut [1], and the European Union funded project called JAST (Joint Action Science and Technology) [63, 5].

The *MORPHA* project [49] envisioned the use of a two-arm service robot at the turn of the millennium (Section 2.2). The vision then was that the robot communicates with its instructors via different communication channels, including physical contact, comments on its actions, and can be used in the most different settings—on the factory floor and at home. Ideally, it can also work in groups and transfer acquired knowledge and skills to its peers (see Chapter [40]).

A joint-action setting involving more than just one robot was part of the *Robonaut* project [1] (Section 2.2). It was also conceived around the year 2000. Two (real) humanoid robots worked on structures with one human in a simulated scenario, e.g., for the International Space Station (ISS). The purpose of the *Multi-Agent Truss Assembly Test* was to develop *teaming strategies* for extra-vehicular astronauts working side-by-side with highly dexterous, teleoperated robots and to study the operational trade-offs inherent in human and robot teaming in a space assembly context. Working together, the two Robonauts operated in various roles supporting the astronaut, who operates in both a leader and support role. A mixture of manipulation and teaming skills was required to complete the truss-assembly task. Truss-assembly agents must not only be capable of mating nodes and struts, they must also be able to coordinate cooperative manipulation, hand-offs and other multi-agent interactions in the pre-planned assembly sequence. Note that the Robonauts were remote-controlled by humans, i.e., they did not have any intelligence or autonomy of their own.

While there had been a number of interesting projects in robotics that concentrated on verbal communication with robots (*do what I say/mean*), a project funded by the European Union called JAST (Joint Action Science and Technology) [63, 5] for the first time adopted findings from neuroscience, cognitive science, and linguistics and fused them with ideas from robotics. It investigated joint action in autonomous systems: To develop intelligent, embodied agents that cooperate and communicate with their peers and with humans while working on a mutual task (Chapter [16]). In the end, a cognitive control architecture with perception, reasoning, motor behavior, and verbal and non-verbal communication tools was demonstrated in various systems that performed a variety of joint-action tasks (Section 3.1).

This chapter will be able to detail many of the techniques only to a given limit, however, there are other contributions within this section of the Handbook which are closely related and we will refer to these accordingly. For example, from a cognitive perspective, it is highly desirable that the robot systems contain cognitively adequate modes of interaction with humans—for dialog control in a given audio scenery (see Chapter [12] on speech), for dynamic control (see Chapter [44]), for synchronizing utterances with motor control (see Chapters [67] and [50]), and for life-long learning and plasticity. In particular the latter has not been investigated very much in the context of joint action between humans and artifacts. Hence, there are no adequate dynamical models that incorporate both the interpretation of sensor stimuli for learning—e.g., (how to learn) to extract the right clues from the learning

examples—and action triggering along with the generation of actions and action sequences with many degrees of freedom. The construction of a formal model of the underlying cognitive processes that addresses these issues is thus fundamental.

Our chapter is structured as follows: In Section 2, basic aspects of physical human-human cooperation and its direct relationship and transfer to human-humanoid cooperation will be discussed, providing explicit examples such as handover motion and timing, cooperative manipulation, followed by conversation and multimodal communication. This is then followed in Section 3 by the discussion of three exemplary projects, Clara [63, 62], Domo [19, 20], and James [23]. The chapters will detail the cognitive technologies and review the decision and planning framework for communication in these projects.

2 Basic Aspects of Physical Cooperation

In contrast to the interaction with a virtual agent, e.g., a face on a screen or a 3D simulation of a human or a robot, the interaction with a physical system presents a whole new set of issues that need to be considered. In contrast to a virtual agent, human and robot can now directly interact with each other using natural modalities, not just mouse and keyboard. This however raises many additional questions, e.g., what properties of physical cooperation are essential for a natural interaction and whether there is a noticeable benefit to the interaction when the robot displays more human-like motions over unnatural and unpredictable trajectories classically associated with robots. Conversation between humans during collaboration is very different to written dialog and is an important factor in task synchronization. The following sections give an overview on three basic topics in physical collaboration: handover between robot and human, cooperative manipulation, as well as aspects of conversation and multimodal communication.

2.1 Handover Motion and Timing

Handing over objects from one person to the next is a task performed several times a day by an average person, typically without thinking about parameters such as timing or other factors. For example, people are often taught to handover scissors with the handle toward the other person and to pick up heavy objects using a different grasp than the one we choose for fragile parts. Given a couple of iterations, human coworkers are able to better synchronize their motions leading to an increase in overall performance.

Meulenbroek et al. [56] examined the coordination between two humans in a joint-action task with the goal of transferring an object. With a focus on kinematic movement parameters, they wanted to test if these were adapted based on the observation of movements of the other test subject. In this user study, a participant was

asked to transfer a vertical cylindrical object from a position in front of him to a new position within a circle in front of the other person, with the latter then transferring it to a position in his own working area. With variations regarding the size (small and large) and weight (light and heavy) of the object, expectations included different trajectories between putting and fetching actor based on observations of the latter to correct for the wrongly perceived diameter/weight ratio. Variations to the size of the target regions (9 or 18 cm) were assumed to lead to differences in speeds adjusted to the size of the circle, with a smaller radius leading to a lower speed. During the study, the participants wore earphones and a facial mask in order to prevent communication or other external influences (e.g., task-related noise). The evaluation showed that the person doing the second transfer was less surprised by a false weight estimate than the one doing the initial motion.

Humans tend to optimize the movement trajectories of their limbs, e.g., using a minimum-jerk profile, thus resembling a bell-shaped velocity profile with a duration of about one second to grasp objects in their workspace. Robot trajectories on the other hand are optimized for maximum efficiency and low cycle times, leading to an unnatural motion (e.g., trapezoidal velocity) commonly associated with robots. User studies evaluating a subjective safety rating for the maximum Cartesian speed using typical robot motions ended up at only $0.225\,\mathrm{m\,s^{-1}}$ [42], which is significantly lower than the average speed used by humans for movements in their workspace and therefore a major issue in efficient human-robot collaboration.

In [38], the motion parameters of human-human handovers were analyzed with focus on the arm movement velocity profile and its effect on human-robot interaction performance. Two human participants sitting opposite each other at a table were given the task to hand over six small wooden cubes. The cubes were aligned in a single line and the giving subject was asked to hand them over to the receiver using one hand, while their arm movements were recorded with a tracking system. Analysis of the recorded data identified three distinctive phases in the interaction: the reaction of the receiver to the giving subject's lifting motion, the manipulation phase with the transfer of the object, and the post-handover with the receiver's placement of the cube on the table (Fig.1a). The receiver started moving to the handover position while the giving subject's hand was still in motion. The subject's velocity profiles showed a typical bell-shaped velocity profile with a mean peak velocity of $0.93\,\mathrm{m\,s^{-1}}$ for the person handing over the cube and $0.85\,\mathrm{m\,s^{-1}}$ for the receiver (Fig.1b). The overall duration for the six cubes was roughly 25 seconds. The duration of a single handover decreased over the six steps of the interaction. If reaction time decreased compared to a previous step, manipulation time often increased slightly, as anticipatory reaction had to be compensated for by a fine tuning of the handover position. The experiment was then modified to study the interaction between a human participant and a humanoid robot (for details of the setup see Section 3.1). For the robot, different upper limb movements were implemented for comparison: a trapezoidal velocity profile in joint space and a minimum-jerk profile in Cartesian space. The former controls each joint of the robot separately and thus results in curved trajectories of the end effector. The trajectory duration was adjusted to roughly 1.2 seconds for each point to point motion, approximating human speed

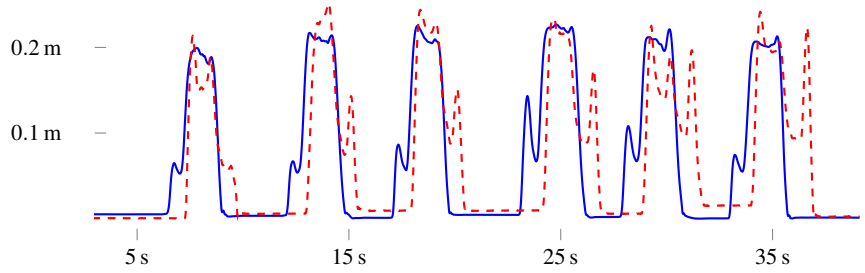

Fig. 1a Plot of tracked hand motions of giving subject (*solid, blue*) and receiver (*dashed, red*) in a handover experiment. The lines show the height of each subject's hand over the table when handing over six cubes.

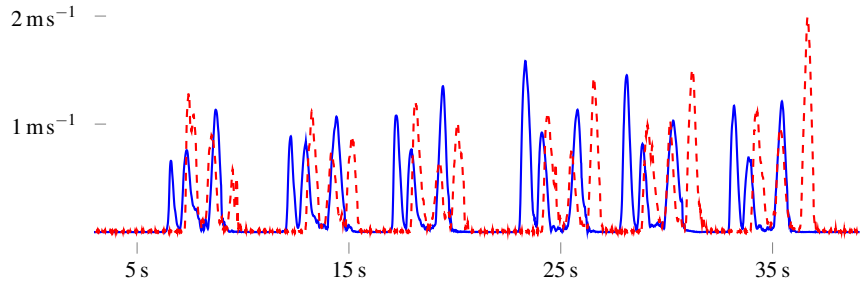

Fig. 1b Absolute hand velocities of subjects in a handover experiment with six cubes. The *solid blue* line shows the participant handing over the cubes, while the *dashed red* line shows the one receiving the cubes.

from the previous trials. Participants were given the same handover task as before, this time with the robot as partner using one of the implemented trajectories, chosen in random order. After completion of this run, the handover task was repeated with the other trajectory type. Apart from recording hand motions for both human and robot, participants were asked to fill out a questionnaire, rating subjective feeling of safety and human-like motion. The evaluation showed a reduction in reaction time for the minimum-jerk profile similar to the one in human-human handover trials, but not for the trapezoidal velocity profile. Overall duration for the reaction time was similar to the human-human performance, with only slightly higher duration for the manipulation phase despite disadvantages in the parallel gripper design. Post-handover duration increased by over 0.5 seconds due to a much slower retraction of the robot's gripper compared to the human in the role of the giving subject. In the subjective rating of safety, human participants felt safer with minimum-jerk velocity profiles and considered the chosen peak velocity of roughly $1\,\mathrm{m\,s^{-1}}$ as acceptable as they were able to predict the robot movements.

The above user study was extended in [37] to include a refined motion profile for the robot during the handover to better resemble human arm movements. The handover trial between two human participants was repeated and modified, adding

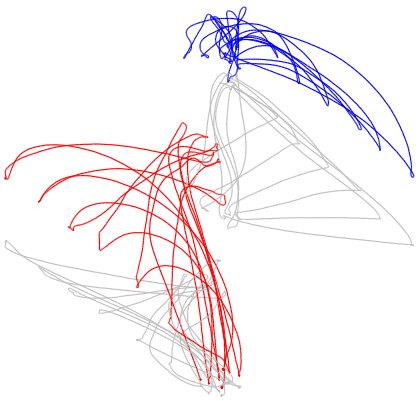

Fig. 2a 3D plot of human-human interaction in the handover experiment. The giving subject (*top, blue*) hands over six cubes to the receiver (*bottom, red*). The projection of the movements on the table surface are shown in *gray*.

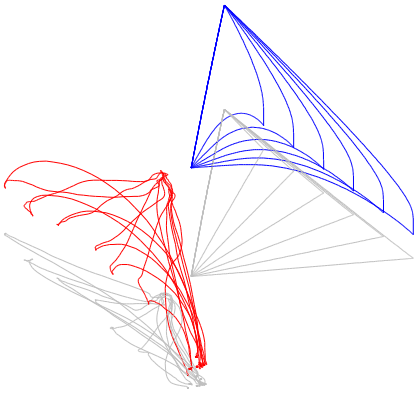

Fig. 2b Movements of robot (*top, blue*) handing over six cubes to human receiver *bottom, red* with projection onto the table surface shown in *gray*.

headphones for the participants and a randomized audio start signal for each step to prevent adaptation to a timing pattern. Fig. 2a shows a typical set of trajectories of such a handover session with six cubes. In contrast to the minimum-jerk motion profile in Cartesian space used in the previous study, it can be clearly seen that human trajectories are not straight lines in the workspace, but rather slightly curved. This resulted in the design of a new *decoupled* minimum-jerk motion profile with different velocities for the x-y plane and the z axis, better resembling those of human trajectories (Fig. 2b). Apart from capturing motions with a tracking system, participants were again asked to fill out a questionnaire focusing on the subjective feeling of safety and comfort, movement predictability, human-like appearance of motion, and abruptness of the start of the robot's motion. The trapezoidal velocity profile in joint space with its unpredictable velocity profile was rated least comfortable and least human-like, while the new decoupled profile had an even higher rating than the standard minimum-jerk profile. Due to its higher maximum velocity, the decoupled profile appeared more abrupt to the participants than the other two trajectory variants.

In the user studies described above, the robot was the individual that initiated a handover and the interaction consisted of a sequence of identical objects in short order. In a joint action setting however, e.g., between a worker and his assistant, the latter needs to be able to anticipate the correct object and the correct time for handover. Both typically depend on contextual information, such as handing over the correct tool for a screw during assembly, and observation of gestures to optimize timing. Huber et al. [36] and Glasauer et al. [32] investigated handover timing in an assembly scenario with tasks of different complexity. The participants were asked

to jointly construct a tower of cubes. Each cube had a different number of holes and could be assembled with a matching number of bolts. The time required for each individual step and the full assembly was measured and both finger, head, and torso movements as well as gaze direction was recorded. In general, assembly time varied depending on the complexity of each step as determined by the number of required bolts and could be described by a linear function. Each participant showed different timings according to his own work speed and some of them also demonstrated a learning effect after their first steps. In order to create predictions of a human's individual assembly time for a robot assistant, a probabilistic model based on a Kalman filter was implemented. It is able to predict the duration of consecutive assembly steps using a normal distribution of the assembly time and continuously updates its parameters based on the worker's current speed and the complexity of the next assembly operation. Evaluation showed that the model is able to adjust itself to the worker after the first two assembly steps. The root mean square (RMS) of the difference between the measured and the predicted assembly duration was used for evaluating the performance. While a linear model was able to achieve an accuracy of 2.50 s (RMS), the probabilistic model resulted in an overall accuracy of 2.48 s (RMS) and an accuracy of 2.06 s (RMS) without the initial two steps. The latter corresponds to an accuracy of 18.03% for the average assembly duration in the given task.

2.2 Cooperative Manipulation

Direct and intertwined physical human-robot cooperation offers many advantages over humans and robots working separately on tasks or subtasks that are interrelated. For example, at certain points during the work, the robot can provide unique skills that humans lack (e.g., high positioning accuracy, high forces), while at other points in time, as needed, humans can contribute their superior perception and other high-level cognition capabilities.

There are several modes or degrees of cooperation one can imagine between robots and humans, which obviously depend on the physical design of the robot:

1. A humanoid robot with legs, arms, and a head can perform cooperative tasks over longer physical distances, e.g., jointly carrying a door with the aid of a human. On the perception side, this would require the humanoid to sense the forces exerted by the human, yet at the same time it would have to stabilize its own body. It also requires permanently observing the human and being able to understand, or even to anticipate the human's intentions, if the interaction is to be deployed smoothly. In this mode of full physical cooperation, the human and robot are not in touch with their respective bodies, only though the object to be worked on. Such robots could eventually even replace the human counterpart in co-working scenarios [74, 35].
2. A fixed humanoid torso with a head (and its associated perception capabilities) and two arms and hands (but without legs) can cooperate almost as universally as

Fig. 3 Humanoid robot Yumi in an assembly task [45].

Fig. 4 Multiple Robonaut humanoids in a joint construction task (Copyright NASA).

a complete humanoid, especially if it has the same perception and communication capabilities as a complete humanoid. Since the robot cannot move, the class of tasks is restricted to what can be performed within the geometrical workspace of the arms [63, 19] (Section 3).

3. A stripped-down version of a humanoid torso is a configuration of two lightweight arms, as implemented, for example, in the *Yumi* robot system [47] shown in Fig. 3. This type of robot is easier to program, but for lack of communication capabilities beyond force sensing on the part of the robot, the permanent synchronization between human and robot is restricted. Therefore, the range of tasks is rather limited, and robot manufacturers have been searching for useful applications for quite some time.
4. The most basic form of physical interaction between robot and humans is a setting where one robot arm with one specialized tool is guided by a human as it works. A typical example is a learning scenario: A human guides the robot by grasping its flange or tool and moves it along the desired trajectory. Later, at run-time, the human can then correct the robot's movement for slight variations in space and time, depending on changes required by that specific work task [68, 72].

Historically, the first realistic visions developed by scientists (as opposed to playwrights or film makers) were joint assembly scenarios for construction on earth or in space. One of the most well-known examples is NASA's *Robonaut* [1, 18] shown in Fig. 4. One of its envisioned tasks was to carry out the joint assembly of large structures in space. This example is interesting because it involves two humanoids and one human astronaut. It is logical then to think about extending this to situations in which there are $n \geq 1$ humans and $m \geq 1$ robots working together. It is also interesting because it takes place in a very controlled environment: Cooperation on

Fig. 5 Simulations of physical human-robot interaction from the MORPHA project's official video: Human and robot jointly carrying a heavy load in a factory setting (Copyright GPS Stuttgart).

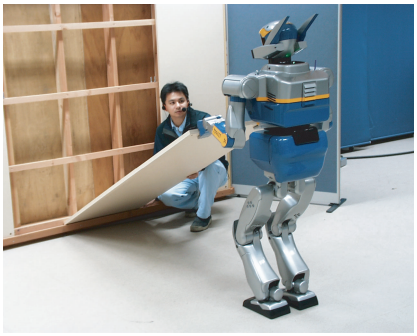

Fig. 6 HRP-2 humanoid working together with a human in the construction of a cottage [35].

earth will typically involve many more factors of uncertainty, e.g., sudden changes in lighting, falling objects, etc.

A more down-to-earth vision was developed in the German research project MORPHA, which started around the year 2000 [49]. In a simulated factory scenario (Fig. 5), the humanoid robot (a torso on a wheeled platform) is already close to replacing a human co-worker: It can help carry a heavy object, it can be instructed to weld along a certain trajectory, it can fetch objects, and it can support the human worker in many other ways. A crucial ingredient in this cooperation is the perception capability of the humanoid: Not only can it receive instructions from the human over this channel, it can also outperform the human in certain subtasks, e.g., counting screws or finding objects quickly in a warehouse.

In Japan, the direct cooperation between humanoids and humans has always been a field of active research. Numerous tasks have been defined that involve using a humanoid in place of a human. Among them are some that involve direct interaction, the most famous may be the one involving the joint assembly of a cottage with the Japanese HRP-2 robot (Fig. 6) [35]. However, carrying an object together was one of the first demonstration scenarios. In 2000, Kosuge el al. [48] demonstrated the mobile, upper-torso humanoid *MR Helper* to lift a panel cooperatively with a human. Yokoyama et al. [74] implemented a walking humanoid that can help carry a long panel and understand spoken commands to grasp, carry, or release that panel. Although it was not really emphasized by the researchers at the time, the ability to communicate accurately via the physical state and then aligning each other's *mental state* is a prerequisite for smooth, successful cooperation. Just pushing and pulling the object and sensing forces is not enough.

Another important step in achieving smooth, successful interaction was the work done at CNRS in France, also based on the Japanese HRP-2 robot. A full library for trajectory control was developed that allows the humanoid and the human to carry objects together, but also to move them along in random turning movements, just

like two humans would move a heavy cabinet that they cannot carry, but which they can lift on one side, shift that lifted side, and then continue by lifting and shifting the other side [54].

While these experiments are certainly interesting from a scientific point of view, they are less relevant for industrial practice. In these contexts, what is currently most interesting are tasks that involve one light-weight arm that works on a rather elementary task with the human co-worker [6]. These tasks are typically single-step tasks and they do not involve sophisticated perceptual skills. They can, however, be considered to be the initial steps toward a future in which more and more difficult tasks can be performed together.

The next step in this development will be reached when the physical interaction can be accompanied by a parallel dialog over the scenery, as pioneered in the early work of one of the co-authors of this chapter [46].

2.3 Conversation and Multimodal Communication

While two persons working on the same task next to each other can easily talk to each other to synchronize their actions, hoist and crane operators use hand signals to accomplish this goal in loud environments or over long distances. Even in places where speech could easily be used to coordinate actions, special circumstances such as police or military operations may prohibit this kind of communication.

The use of gestures as a medium of communication during conversation is however controversial. De Ruiter [64] considers both positions and argues that there is no actual conflict between the view that gestures are a communication device and the opposite argument: While the effectiveness of the gesture may vary, the speaker's goal is always to use it as a communication device. People may use gestures while speaking on the phone because they are used to it, even if the other person cannot see them. In a conversation, the speaker's intention may be to improve the listener's understanding of an abstract idea with—when using them without speech—cryptic motions (see also Chapter [50] for non-verbal communication).

McNeill [55] distinguishes between five different types of gestures: *Iconics* illustrate the spoken text with a gesture that refers a concrete object or event, e.g., when the speaker sweeps his own arm backwards while explaining how someone else did this. *Metaphorics* are similar, but instead depict an abstract idea such as a genre of cartoons. In this case, the speaker's hands might rise up to offer the listener something concrete in the form of an image or bounded object. *Beats* are a short and simple flick of the hand or finger movement with two movement phases, e.g., in/out or up/down, that mark the accompanying word or phrase as significant. *Cohesives* are based on repetition and can consist of iconic, metaphoric, pointing, or beat gestures. A constant series of beats during a political speech is given as an example that aims to highlight consistency. Finally, *deictics* refer to the pointing gesture used to indicate objects or abstract places during a conversation.

Clark [14] shows that interaction goes beyond simple verbal communication and that non-verbal input and output such as pointing at objects and placing of objects is just as relevant. Some verbal expressions can only be fully understood in combination with non-verbal expressions, e.g., when referring to an object "over there" or as "that one" by pointing at it. Placing oneself with items and money in cash at a counter in a store is essential to completing a shopping transaction in this context. He distinguishes between two basic techniques of indicating: With *directing-to*, a speaker directs the addressee's attention to a specific object, whereas with *placing-for*, he places an object in the addressee's attention. Among the many methods of *directing-to* are pointing with a finger, sweeping with an arm, nodding with the head, tapping with finger or foot, turning with the torso, directing with the face, or gazing with the eyes. The latter requires mutual attention to be effective and is often combined with pointing as well as face and torso direction. Voice is another common device to to indicate a speaker ("me"), a location ("here"), or a time ("now"), as are artificial devices such as laser pointers or markers. *Directing-to* is often used with composite signals, e.g., demonstrative pronouns ("this", "that", "these", "those") or adjectives, summonses ("hey, you"), emblems (e.g., goodbye wave, thumbs-up, shoulder-shrug), or iconic gestures. *Placing-for* on the other hand is about placing a specific object at a specific place with a specific action. Among these objects are persons (*self-objects*) and material things (*other-objects*). Examples for the former include standing behind a counter as a clerk or in front of a counter as a customer, whereas a waiter putting a plate with food in front of a customer is an example for the latter. The site of placement plays an important role in the interpretation of the object. The store counter of the previous example has a site for the clerk, one for the customer, and one for the transaction items. Equally important are the three phases of *placing-for*: The *initiation*, where an object is placed, the *maintenance*, where an object remains in place, and the *termination*, where an object is replaced, removed, or abandoned. One *placing-for* act will often set up a following joint action, e.g., stepping up to the counter leads to the interaction with the clerk.

The coordination of references during conversation is another important aspect in natural collaboration. Clark and Wilkes-Gibbs [15] see this as a collaborative process that involves both parties in a shared effort. It is different from written text because the speaker's time for planning and revision is limited, the listener has to follow the dialog in real-time, and the speaker can adjust his dialog based on his observation of the listener. In conversation, both sides work together to ensure a proper understanding of each reference. They continue the dialog only when they share the belief that this is the case. The listener can signal this either by simply letting the speaker continue or with "yes", "right", "I see", or nodding his head. The observations are verified in a user study, where two participants—director and matcher—were asked to arrange 12 cards of Tangram figures in the correct order. The director had to describe the order of his set of cards to the matcher so he could rearrange his set accordingly and both were allowed to talk freely. The task was repeated six times. The evaluation showed, that the number of words used by the director to describe a figure reduced with each trial as the partners became more efficient.

De Ruiter et al. [65] modified the experiment of Clark and Wilkes-Gibbs [15] to evaluate a connection between gesture and speech in the production of referring expressions. In the investigated tradeoff hypothesis, more gestures are used when the use of speech alone is more complicated, whereas in the alternative hand-in-hand hypothesis [69], more speech goes together with more gestures. In this variant of the study, both parties sit in front of a wall with the Tangram figures at slightly more than arm's length and have to identify a specific one, with and without a dividing wall between each other. Figures shapes range in increasing complexity from simple to humanoid-like to abstract and three trials were performed for each team. The evaluation distinguishes between *pointing* gestures, *obligatory iconics* with essential information not included in speech (e.g., drawing a curve in the air while saying "the one with a shape like this"), and *nonobligatory iconics* that do not add further meaning (e.g., drawing a triangle in the air while saying "the big triangle"). The study showed that no pointing gestures or obligatory iconics were used when the participants could not see each other, but the use of nonobligatory iconics was not affected. Without the wall, the number of pointing gestures was related to the number of location references in speech and more iconic gestures were used when the amount of described features increased. The results were inconsistent with the tradeoff hypothesis and evidence supporting the hand-in-hand hypothesis was found. A model where speech and gestures convey the same information is therefore suggested to better mimic human behavior.

Bard et al. [4] use a joint construction task [13] to analyze referring expressions during collaboration, i.e., the use of indefinite/definite expressions ("a/the red triangle"), deictics ("that triangle"), or personal pronouns ("it"). In the task, two players have to construct shapes from basic Tangram pieces on separate computer screens. Based on the test setting, they can communicate via speech, are able to see each other's mouse cursor, and are shown a cursor highlighting the gaze of the other participant as tracked via an eye tracker. Participants were either both given the role of task manager or one was assigned task manager and the other one acted as assistant. In order to join two pieces together, each player must be holding one of the parts and the resulting assembly is permanent. Pieces break when they overlap or if the players select the same object. Build accuracy, broken parts, and assembly time were measured. The evaluation with speech showed no difference in build quality when comparing groups with different or identical roles, but the former one showed faster performance. With or without visible mouse cursor resulted in the same accuracy, but the latter had shorter dialogs and slightly fewer broken parts. Gaze did not have any measurable effect on any task criterion. With visible mouse movements, participants used less definite expressions in exchange for deictic ones, while the use of pronouns was unaffected. Participants used more deictic expressions over indefinite ones when moving objects, regardless of mouse cursor visibility.

Based on the results of this study, Foster et al. [24] performed a task-based evaluation of referring expressions using the toy assembly scenario and the humanoid robot system described in Section 3.1. In this scenario, participants were asked to build an assembly consisting of various wooden parts such as bolts, cubes, nuts, and slats. A new context-sensitive generator for referring expression was designed,

as the previous study showed that participants did not prefer indefinite expressions when first referring to an object, but often used other categories. Two scenarios were evaluated: In the first one, the robot is the only one with knowledge of the assembly plan. In the second one, both participants are aware of the plan, but the user can optionally be given an incorrect one for a trial. The study compares the new algorithm with the standard incremental algorithm [17] for reference generation. It considers objective measures such as duration, turns, response time, task success and subjective measures in the form of a questionnaire. While the standard incremental algorithm uses a predefined domain-specific preference order to select the most relevant attributes, the context-sensitive approach uses the dialog history and position of the current object. The objective measures did not show any significant difference between the two algorithms, but the participants rated the contest-sensitive version better compared to the classical one.

Similar to the store example introduced earlier, a bar setting is an interesting scenario in that it features the interaction of a bartender with multiple customers. Loth et al. [51] recorded customer interactions in several German bars to study the use of service initiation signals with a focus on results that could be applied to a robot bartender. The recorded data was annotated and used to identify behaviors that occur with high frequency. Based on the results, the signals *looking at the bar* and *being directly at the bar* were selected for further validation. In two experiments, participants were shown video material of bar interactions and asked to respond whether in their view the customer had the intention to order. The study showed, that just one of the signals was not sufficient to trigger an order, but the participants could not distinguish between false and correct orders if both signals were present. They also first looked at the customer's position relative to the bar before checking their gaze direction. Results from this study were used in the first version of the bartender robot system shown in Section 3.3 for estimating whether a customer is currently seeking for attention. The performance of this rule-based estimator versus one based on supervised learning was evaluated in [22]. The former proved to be more stable, whereas the latter was shown to be faster at detecting initial intended user engagement. The different classifier did however not influence the users' subjective rating of the system.

Giuliani et al. [31] evaluated a robot bartender with a purely task-based approach against a socially appropriate system with an additional set of rules. A central hypothesis was that service robots need not only achieve a certain task, but also behave socially appropriate when interacting with humans. Participants were asked to fill out questionnaires with their subjective ratings before and after their interaction with the humanoid. In the interaction phase, each participant interacted with both system versions in a randomized order. The pre-test/post-test design was chosen to control for prior user expectations. In addition to subjective measures, objective measures were gathered on the task success rate, the number of repeated dialog actions because of speech recognition issues, the number of repeated questions, and the duration from a customer's appearance to when a drink was served. In the experiment, the task-based system used fewer dialog system turns than the social version, but the social version led to a shorter duration of the overall transaction

from a customer entering to when their drink was served. The regression analysis did not show a significant difference in subjective ratings between the two versions. However, there were strong correlations between objective and subjective measures. Higher numbers of system turns resulted in lower perceived intelligence ratings for the humanoid. Repeated order requests strongly reduced subjective likeability. Interestingly, longer durations of the interaction increased the likeability. In general, the user study demonstrated that the humanoid can successfully interact, but surprisingly, it did not matter whether the socially-appropriate interaction scheme was turned on.

A study with the second version of the bartender robot system of Section 3.3 examined which sensor modalities are the most informative for humans to manage an interaction in place of the humanoid bartender [52]. The underlying idea was to study human-human interaction through a limited user interface that shows speech and visual recognition results and then design the humanoid system based on the findings. In the *Ghost in the Machine* experiment, a participant (or, ghost) observes speech and visual recognition results and can trigger robot actions in real-time. In this kind of experiment, the robot interacts with confederates. The participant of the study is located in a different room, can observe recognized data (recognized speech, locations of people, body posture), and is supposed to trigger robot actions (speech output, eye gaze direction, gestures, manipulation actions) to interact with the human confederates. Contrary to *Wizard of Oz* experiments, participants cannot observe the interaction directly, but only see recognition results. Using an eye tracking device and measuring eye gaze and dwell time on the user interface, it could reliably be established on which modalities humans base their interpretation and response. The study showed, that position and posture of guests are the most important signals for initiating an interaction in the bartending domain. Humans tend to respond in the same modality as the interaction partner, like to establish eye contact for a visual handshake, and echo parts of an order to verify it (see also Chapter [67] for more information on eye gaze etc.).

3 Examples of Physical Human-Humanoid Interaction Systems

3.1 Clara

Two persons collaborating on an assembly task, e.g., a toy set, either divide the project into two parts that can each be assembled on their own, or one partner takes over the role of builder, while the other one prepares the parts for the next stage of the assembly according to the instruction manual and even highlighting individual operations. This is essentially a form of *placing-for* and *directing-to* as outlined by Clark [14]. In this joint action scenario, both persons typically work together on a tabletop with a shared workspace between them and an own personal workspace that the collaborator cannot access. Both verbal and non-verbal communication as well

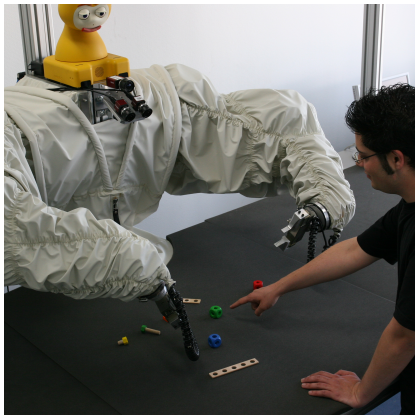

Fig. 7a Humanoid robot Clara with two industrial robot arms and an animatronic head used in the joint assembly of Baufix parts.

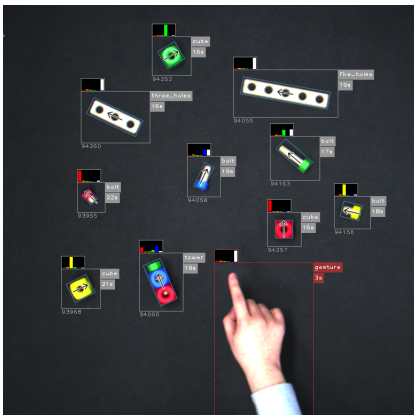

Fig. 7b View of Clara's object and gesture recognition using a camera mounted on top of the table with the shared workspace.

as a combination of both play an important role in these types of interaction. The two collaborators share an understanding of the domain, in this case an assembly process, as well as common knowledge, such as colors, shapes, or basic actions. They have to observe each other's behavior and coordinate their actions accordingly.

The humanoid robot Clara (Fig. 7a) was developed around such a joint-action scenario by Rickert et al. [63, 62] at the Technical University of Munich in the context of the JAST (Joint-Action Science and Technology) project funded by the European Union. It supports a human and robot working together in the assembly of various wooden components—slats, nuts, bolts, cubes, etc.—of a Baufix toy set. The parts can be used to create assemblies such as airplanes, motorcycles, or railway signals. As part of this construction process, the collaborators also create subcomponents that are referred to by individual names. Understanding dialog, referring expressions, and multimodal input are therefore highly important in the design of such a system. A human will often not refer to an object with a complex verbal expression, but rather just point at it and say "this one"—while the coworker has knowledge of the parts required in the next assembly step [21]. In case of errors, he will also correct his collaborator by using expressions such as "the other one". For human and robot to actually work together, the robot also needs to be able to handle the parts of such an assembly process. The most common operations used by humans in these settings include *pick up*, *screw*, *point*, *put down*, and *plug* [39]. The wooden parts used in the scenario exhibit large variations, therefore performing these actions with a robot system is quite challenging.

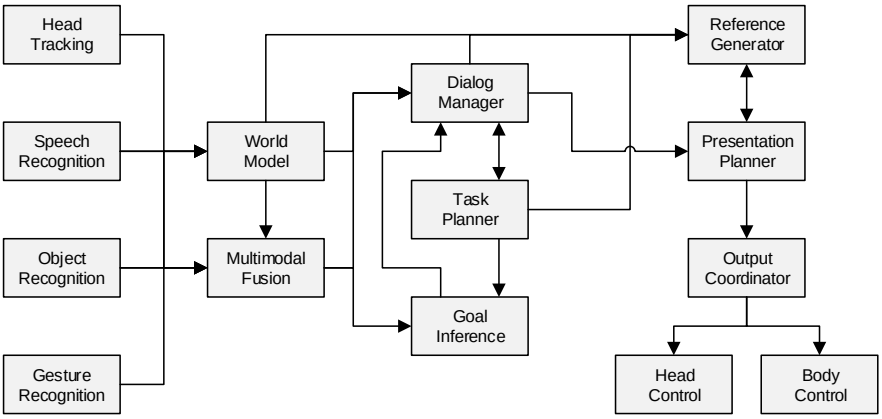

Fig. 8 System architecture of Clara used in the JAST project. It is structured into four sections from left to right: *Input*, *Interpretation*, *Reasoning*, and *Output*. The individual modules communicate over a distributed and cross-platform middleware.

3.1.1 Physical Setup and System Architecture

Clara uses two standard industrial manipulators, each featuring six degrees of freedom, mounted opposite each other on a cage-like structure of aluminium profiles around a large wooden table (Fig. 7a) [63, 62]. The arms and the base mount are covered in a protective hull commonly used in painting or welding applications to hide the mechanical structure. Each manipulator is equipped with a force/torque sensor and a pneumatic two-finger gripper. On top of the base mount, it features an *iCat* animatronic head [10] with 13 servos for body control and generation of facial expressions. It has a camera in the location of its nose, a microphone and speaker in its base, and can synchronize its lips to spoken text. During interaction, an additional headset is worn for better speech recognition accuracy. A camera mounted above the table and facing downward is used as input for the object and gesture recognition modules. Additional cameras mounted below the head and facing toward the human and the table are used for face detection and object/gesture recognition.

The system features a distributed architecture with multiple components running on a number of computers, operating systems, and programming languages. The modules communicate over a middleware that supports remote procedure calls and publish-subscribe connections [34]. It is broadly categorized into four main sections (Fig. 8): The first one handles the *Input* of various information streams, including object and gesture recognition, head tracking, as well as speech recognition and processing. The shapes and colors of the objects in the construction task are well known and template matching (Fig. 7b) is used to detect this information together with matching position and orientation [57]. Overlapping of objects is a common feature in this scenario and the templates support occlusion up to a certain amount. Assemblies consisting of multiple individual parts can be detected as well if a matching template is provided. New objects can be introduced at runtime in

combination with a provided name. As the robot manipulators can enter the camera image and lead to false recognition results, their Cartesian position can be queried and used to avoid these regions. Gesture recognition uses the same camera image to detect the presence of a human hand and type of gesture it is using, including *pointing* with an index finger, *grasping* with index finger and thumb, or *holding-out* an open palm. The module reports the type of an individual gesture together with a probability and can be extended to support new gestures with a number of training images [75]. Head tracking is based on a Contracting Curve Density (CCD) algorithm [58] and reports the location of the human's head in the camera frame. Speech recognition uses a commercial software package with a software development kit (SDK) together with a customized grammar for improved recognition results. It reports multiple hypothesis together with their individual probabilities. All input data is broadcast together with timestamps in order to enable further processing later on.

The *Interpretation* modules contain a world model and a multimodal fusion component. The world model stores information on parts in the current interaction context and their locations, i.e., the tabletop itself, one of the robot's hands, the human collaborator, or an assembly. For the tabletop, this includes the pose of the object. While the object recognition module reports the currently visible objects, the world model has a persistent view and remembers the state even if the part in question is currently obstructed. Objects can be queried according to descriptions, locations, or world coordinates. Multimodal fusion takes information from all inputs and detects correlations between verbal and non-verbal communication channels. A detected object combined with a pointing gesture that includes this object among relevant parts in its related area together with a verbal expression "take this one" and an overlapping timestamp results in a strong instruction hypothesis for following modules. Speech processing—an important part of this module—is based on Combinatory Categorial Grammar (CCG) and results in a logical expression of the verbal statement (see also Chapter [12] for information about the use of speech as communication mode in humanoids). It provides both German and English language support and is able to process task-related imperative sentences, questions, statements, confirmations, barge-in, and—to some degree—elliptical sentences. Speech and gesture input are then transformed into a compatible representation and combined with a given rule set [30].

The *Reasoning* section consists of a dialog manager, a task planner and a goal inference module. The system combines both symbolic and sub-symbolic reasoning to focus on different aspects of communication [25]. With a focus on symbolic representations, the dialog manager [26] excels in the processing of complex verbal instructions, while the dynamic field theory [7] of the goal inference module is able to anticipate future and unexpected actions (Section 3.1.2 and 3.1.3).

Reference generation, presentation planning, output coordination, as well as body and head control sum up the modules for the *Output* section. After the reasoning components come to a decision that should result in an action involving one of the output channels, e.g., speech and/or motion, the dialog manager sends a corresponding command to the presentation planner [21]. If speech is a part of this

action, it will involve the reference generator in order to create a more appropriate response based on its dialog, world, and task state (Section 3.1.2). The generated output plan is then sent to the output coordinator module for temporal and spatial synchronization of animatronic head and robot manipulators, e.g., to match gestures and speech output. The SDK of the head offers the ability to rotate and tilt the head, control eyes, eyelids, and lips for facial expressions, and can synchronize the lips to a commercial text-to-speech engine. Robot control is performed on a computer with a real-time operating system. It offers a set of skills to the coordinating module that are relevant for the assembly domain, for instance, *pick up*, *point*, or *put down* [63]. Each skill coordinates several components, e.g., a robot manipulator and a force/torque sensor, and is configured with parameters used in the individual steps of its state machine. A single step can specify a target and limits, such as a joint or Cartesian position or a force or torque value, and is executed until a combination of exit conditions is met, e.g., a time limit, a goal position or velocity, a force or torque limit, or any external input. Due to this, motions can be interrupted at any time to enable barge-in. Global limits can be used to ensure maximum joint or Cartesian positions and velocities. A step can also trigger the execution of an individual command, such as opening or closing a gripper. A *pick up* skill therefore consists of individual steps, that move to a Cartesian approach pose above the object coordinates, open the gripper, move down with a given velocity until a force threshold is reached, close the gripper, and move up to a Cartesian goal position. Motions for handover follow the results of Section 2.1 to increase joint-action performance.

3.1.2 Task-Based Human-Robot Dialog

Objects in the Baufix scenario include elements such as bolts, cubes, nuts, and slats. All cubes are identical and only differ in color. Bolts are of different sizes, identified by color, and come in two types of heads, round and hexagonal. Slats are of different sizes and are differentiated by the number of holes, ranging from three to seven. During the assembly, specific holes need to be referenced for the correct placement of parts. All objects of the current assembly interaction are stored in the world model component together with their location and state, i.e., on the table in the workspace of the user or the robot, in the hand of the user or the robot, as a single part or in an assembled state.

The task planner module represents the assembly plan in an AND/OR graph that allows it to model different assembly sequences [26]. Each vertex of the graph has a unique ID and can be given a label. The specific word depends on the language of the used grammar. Fig. 9 shows an example of such a graph for the *railway signal* assembly. Subassemblies can also be given names, in this case their English names are *snowman* and *L-shape*. The order of assembly is either based on instructions of the robot or can be chosen by the user. The dialog manager is based on the TrindiKit toolkit [71] and uses an information state update approach. States in Clara's architecture include data of the user's knowledge, the current assembly plan step, the history of assembly steps described to the user, and the interaction history.

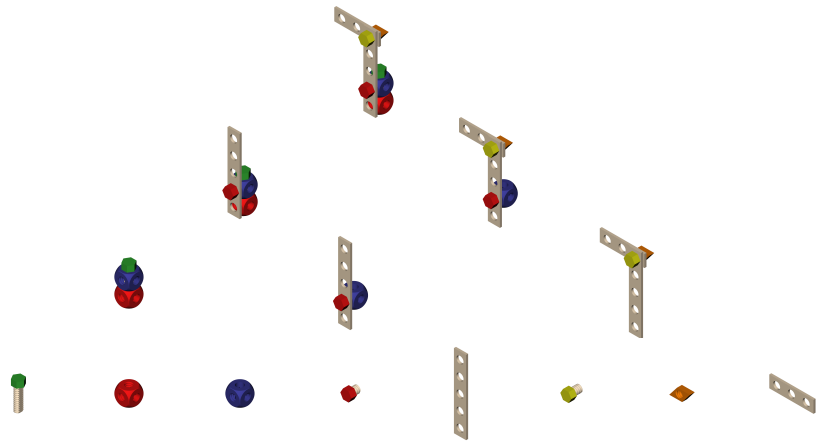

Fig. 9 Assembly plan for a *railway signal* assembly. The final object requires eight Baufix parts in total and includes the labeled subassemblies *snowman* (second row from bottom, left) and *L-shape* (second row from bottom, right). Different sequences can be chosen that result in the same final assembly.

To start an assembly with the robot as an instructor, the dialog manager requests the task planner to load the specific AND/OR graph for the selected assembly and the planner chooses a valid assembly sequence. The dialog manager then instructs the user with one of two strategies: *Depth-first* gives assembly instructions for a step and only names the object upon completion, while *top-down* does both at the same time. As the system keeps track of steps already explained to the user, it can ask if it should explain a repeated step again or if the user still remembers how it is done. This can also be used to reduce linguistic output when multiple sequences with the same result are available. The dialog manager then generates a representation based on Rhetorical Structure Theory (RST) [53] that can be sent to the presentation planner component. RST structures are used to describe relations between different parts of natural text, e.g., a condition, an elaboration, or a sequence relation. The latter describes a succession relationship between one sentence part and the next, for instance when describing that one specific step has to be performed before the next one. The user can utter various forms of verbal acknowledgment, indicate a misunderstanding, or use simple yes/no answers. The task planner is notified on each completed assembly step so it can update its internal representation. The world model is updated accordingly and the information state of the dialog manager includes the performed assembly step in the user's knowledge data [26].

Identifying specific objects is critical for the interaction in the Baufix scenario. As highlighted above, there are for instance many different versions of bolts and slats differentiated by color and shape. Rather than relying on complex wording to precisely identify the object in question, the use referring expressions play an important role in the generated dialog to make it more accessible [21]. As the robot system is able to physically interact with its environment in this scenario, it can go

beyond referring to a previously used object with verbal expressions. While picking up an object and referring to it as "this one" is more effort, a more intense and accurate reference can be achieved this way. If the object has to be used in the next assembly step and is in the robot's rather than the user's workspace, this can directly be followed up with a handover action. The system features two different implementations: The first one is based on the incremental algorithm of Dale and Reiter [17] and several of its extensions. From a given set of distractor objects, it incrementally selects attributes missing in at least one other object in the set and repeats this until only the target object remains. This greedy approach does not necessarily lead to the optimal solution, but closely resembles human decisions. The second one is context-sensitive and inspired by results from a user study [4, 24].

3.1.3 Combining Goal Inference and Natural-Language Dialog

As presented in Section 3.1.2, a dialog manager's strength lies in dealing with complex verbal input and output, while gestures and multimodal input and output can be naturally integrated as well. When it comes to the anticipation of future actions from its human collaborator, however, it can only give an estimate based on information from a known assembly plan for this scenario. Dynamic field theory on the other hand is inspired by findings from neurocognitive mechanisms [7]. In a joint action scenario, a cognitive agent can compensate temporarily missing sensor input through self-stabilized inner states and can anticipate future inputs related to a specific goal-directed behavior. Different outcomes compete with each other based on information from observed actions, contextual cues, and shared task knowledge until one emerges as the winner.

The two implementations are fundamentally different, yet share a number of identical information properties [25]. On the level of gesture, object, and action representation, a common ground can be established: the different sets of gestures handled by each system include pointing, grasping, holding-out, and unknown, together with the indicated object if applicable. An object has a classification, a pose, and an indication if the object is within the robot's workspace. The system can perform a limited set of actions, i.e., grasp-and-give, demand-and-receive, speak, and an undefined action together with a string for further data, e.g., the ID of the corresponding object. Both reasoning engines receive updates from the multimodal input channels. The inferred goals and suggested responses from the goal inference module are sent to the dialog manager, which integrates this information via appropriate update rules and generates an output for the presentation planner.

With input from the goal inference module, the dialog manager is able to react to unexpected actions from the human collaborator that do not match the current subgoals [25]. It can choose to inform the user of the correct assembly sequence, engage in a dialog to clarify the user's intention, or update its internal representation of the current task. The design can be adjusted to follow either the user's lead or the plan selected by the system and thus lead to completely different personalities.

3.2 Domo

A classical industrial robot with six degrees of freedom is designed to achieve high precision and low cycle times in structured industrial settings without human interaction. In order to handle very high payloads such as welding guns and due to limitations of mechanical design, the ratio of payload to weight is typically in the order of 1:10. The high speeds of these enormous moving masses on the one side and the lack of sensors or optimized hardware to prevent contused wounds at low speeds on the other side make this kind of design unsuitable for physical human-robot cooperation [33]. Due to the rigid mechanical design of these systems, the interaction with objects needs to be programmed very precisely in order to prevent damage to the system or the environment. The fixed customization to individual industrial use cases with specialized tools and gripper fingers limits their usability in environments with different types of objects that humans are easily able to handle.

In a completely different approach, the humanoid robot Domo shown in Fig. 10a was designed by Aaron Edsinger at the Humanoid Robotics Lab of MIT with the goal to assist humans in everyday tasks [19, 20]. It is based on the principle, that human environments are optimally suited for the human body and human behavior and it is designed to be able to handle a wide range of household items such as cups, bottles, tools, or food items without the need for significant modifications.

Domo's design is characterized by three major themes, each highlighting two or three subcategories: The first one titled *Let the Body do the Thinking* [19] focuses on a number of design strategies. With the *Human Form* as foundation, it enables the use of tools created for humans and perceives the world from a similar physical perspective as a human. Its proportions are chosen to be close to those of a human in order to be able to take advantage of an environment designed with people in mind, e.g., the use of standard human workplaces. Its *Design for Uncertainty* is centered around inherent safety through the use of compliant and force controlled actuators. Passive compliance in the body and the soft material used in its hand enable Domo to handle uncertainty in its environment, e.g., in the case of unexpected contact or when dealing with position uncertainty in a contact task. By *Taking Action*, the robot aims to use its ability to actively participate in its environment to its advantage. It can move its head and eyes to improve its recognition of objects via active perception, adjust the stiffness in its arms in unknown environments, or increase the opening of its hand when grasping in the dark.

The second theme *Cooperative Manipulation* [19] aims to enhance the robot's interaction capabilities with humans through adequate behavior patterns. Proper use of social cues [9] understandable by a co-worker can help the robot assume the role of a collaborator and both human and robot are then able to complete tasks as a team (see also Chapter [40] for more detailed information on social aspects of human-robot teaming). The field *Assistive Robotics* contributes the aspect of physical interaction as seen in elderly care or manipulation assistance with tasks such as food preparation, operation of electronic devices, and picking up or placing of objects. Due to the direct contact between robot and human, safety is a major factor in the design of Domo. *Collaborative Cues* play an important role in the interaction

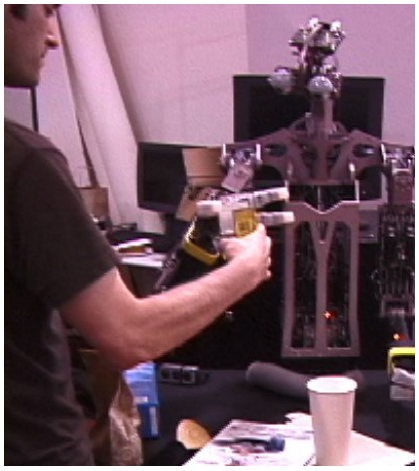

Fig. 10a The humanoid robot Domo grasping a bottle in a handover scenario [20].

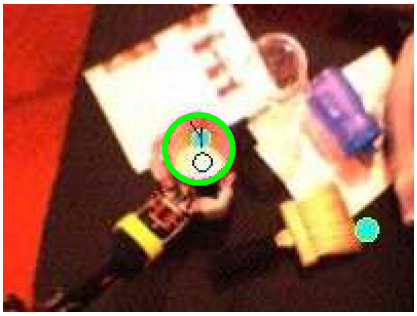

Fig. 10b Domo's view of the environment and the corresponding feature detection [20].

between two co-workers. The use of hand gestures such as pointing to an object as well as eye and head movement to direct the attention to a specific object or to highlight turn taking are all important aspects to improve collaboration.

The last theme aims to identify *Task Relevant Features* [19] that are common in objects designed for human use. Domo's goal is to be able to interact with objects that have not been specifically modeled beforehand. To accomplish this, the software must address *Perceptual Robustness*, as the typical home or office environment have very cluttered backgrounds and lighting conditions that change heavily throughout the day. *Generalization* means that principles that were successfully applied to a specific object should also work on a similar one, e.g., picking up a a coffeepot or a briefcase via its handle or opening a lid with the tip of a screwdriver or a knife.

3.2.1 Physical Setup and Safety Aspects

Domo is designed as a humanoid robot with two arms, two hands, and an actuated head in dimensions matching those of an average human [19]. The robot's kinematics feature a total of 29 degrees of freedom (DOF), with 9 DOF in the head, 6 DOF in each arm, and 4 DOF in each hand. The 22 DOF from the neck down feature a *Series Elastic Actuators* (SEA) [61] design with a compliant element between motor output and load in order to provide force sensing and passive compliance. All components are custom built in order to fulfill the design requirements mentioned in the previous section.

The arms are designed to be light-weight in order to reduce inertia, improve safety, and reduce motor requirements for higher efficiency. A cable-driven design

inspired by the Barrett WAM manipulators [70] is used to enable a move of the shoulder motors into the torso for further weight reduction. These two DOF control pitch and roll of the shoulder, while four actuators in the bicep control shoulder yaw, elbow pitch, as well as wrist roll and pitch. With a weight of 2.1 kg and a payload of 5 kg they offer a good balance between safety and usability quite opposite to the ratio offered by standard industrial manipulators.

Design goals for the hand included passive compliance and force control, especially for dealing with unknown objects and position uncertainty. In contrast to the SEA design of the arms, they use a compliant element between motor housing and chassis for a more compact design. Each hand weighs 0.51 kg and consists of three fingers with three joints. Only one joint is actively controlled while the other two are passively coupled. The one remaining DOF is used to control the spread between two of the fingers. The fingers are covered with 24 tactile sensors and a soft urethane material, the latter enabling robust grasping of a variety of objects and improving the ability to maintain a stable force controlled grasp. This also improves the overall robustness of the design when interacting with the environment.

Domo's head is a mechanical copy of the MERTZ design by Aryananda and Weber [3]. It features a neck with three DOF and an upper head with one DOF each for roll and tilt. It is equipped with two CCD cameras controlled by a single tilt DOF and two independent pan DOF. The remaining DOF is used to control the eyelids. A 3-axis gyroscope provides an absolute reference with respect to gravity. As visual attention concepts play an important aspect in the design of the robot, special emphasis was given to human-like eye movement. With the exception of the optokinetic response, Domo supports saccades with fast, ballistic eye movements of 900°/s, smooth pursuit with slow, controlled tracking movements of up to 100°/s, vergence with independent control of eye pan, and the vestibulo-ocular reflex through a head mounted gyroscope to counter-rotate eyes during head movement.

Physical safety when collaborating with a human co-worker was a major design aspect of Domo. Edsinger [19] evaluated this aspect of robot interaction with the *Head Injury Index* (HIC) [73]. The HIC summarizes the impact acceleration of the head and the duration of impact in one function, where a value of 100 is considered safe and a HIC of 1000 can be fatal. His experiments for Domo show a HIC of 167 for an impact velocity of $1\,\mathrm{m\,s^{-1}}$ when using both SEA and cable-drive compared to a HIC of 489 for a non-SEA and non-cable-drive version. A HIC of 100 can be reached by limiting the maximum velocity to $0.84\,\mathrm{m\,s^{-1}}$ for the former and $0.52\,\mathrm{m\,s^{-1}}$ for the latter version. However, as shown in [33], HIC alone is not an accurate measure of injury severity in human-robot interaction.

3.2.2 Behavior-Based System Architecture

While the traditional sense-plan-act approach requires precise models of objects for a proper interaction with them, these kinds of models are hard to define for typical household objects. Domo therefore uses a behavior-based approach [2] and refers to perceptual observations instead of internal models. A set of simple behaviors is

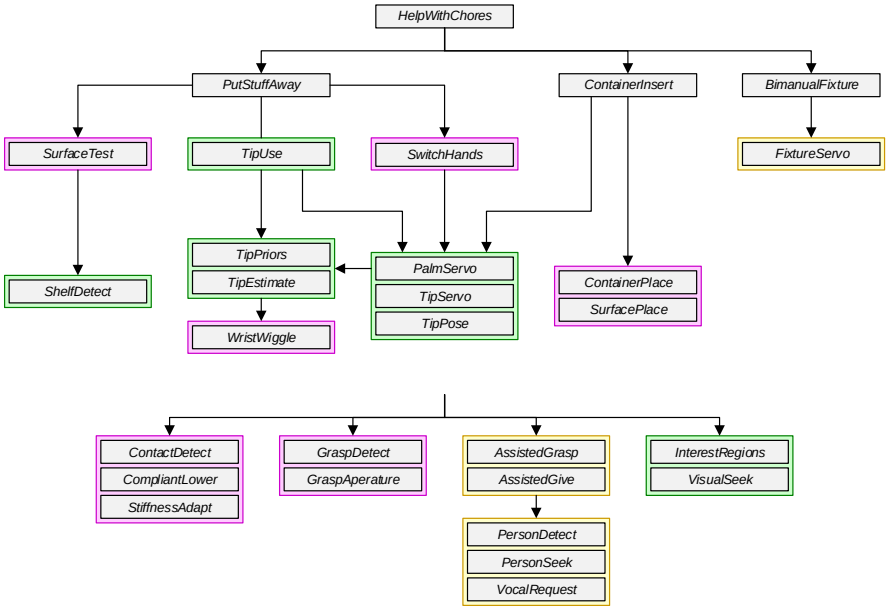

Fig. 11 Overview of Domo's architecture for the *HelpWithChores* task. It is structured into four layers, with the *Perceptual Detectors* and *Motor Primitives* on the bottom up to the *Task Skills* on the top. The colors highlight modules according to their underlying strategy, with *yellow* for *Cooperative Manipulation*, *green* for *Task Relevant Features*, and *purple* for *Let the Body do the Thinking* [19].

designed to work around an incomplete sensory representation of its environment. These modules receive access to resources such as a robot's manipulator based on their priority and the estimate of their readiness level. Higher levels of certainty lead to a higher likelihood of the behavior being given control. Domo's architecture can be structured into four basic layers (Fig. 11), which are detailed in the following paragraphs [19].

A number of *Perceptual Detectors* and *Motor Primitives* provide the basis for the layers above. The former modules include implementations for detecting contact of the manipulator with the environment, the contact surface of the palm, or the aperture of a grasp. The latter ones take care of modifying joint stiffness, directing eye gaze, or reaching toward someone.

On the layer above, *Compensatory Actions* reduce perceptual uncertainty, *Precondition Actions* deal with adjusting the robot's pose for follow-up actions, while *Task Relevant Features* estimate stable features in the environment over time and can combine these with control actions.

The coordination of these detection and control actions is done via hand-designed *Manual Skills*. Here, an algorithm is split into several stages connected in a control flow. The skill starts in a ready stage, where it waits until related perceptual preconditions are met. After this, relevant compensatory and precondition actions are

triggered to prepare the system for the following detection stage. With the detected features as input, the control stage takes over the respective hardware component and monitors the execution until the task is successful. The skill will fall back to the ready stage at any point during execution, if the system does no longer meet the readiness conditions.

At the very top of the architecture are the *Task Skills*. These coordinate the manual skills from the layer below in close interaction with a human collaborator.

3.3 James

The James humanoid bartender (Fig. 12a and 12b) is a robot system designed to study short-term and multi-party human-humanoid interaction that was developed in the project JAMES (Joint Action for Multimodal Embodied Social Systems) funded by the European Union [23]. Its application to the bartending domain motivates both task-based interaction, such as ordering and serving drinks and cleaning the bar, as well as social aspects of how to engage in conversations and how to handle requests following social rules. Experiments include multiple user studies to evaluate the acceptance and quality of the human-humanoid interaction [31], studies to compare different planning approaches to control the robot's behavior [31], and a *Ghost in the Machine* study [52], where participants control the robot in real-time.

3.3.1 Physical Setup and System Architecture

The first version of James was based on the hardware of the system described in Section 3.1. Its upper body consists of two industrial robot arms with six degrees of freedom (DOF), hidden under a plastic cover. In contrast to the parallel grippers used before, the hands are based on an enhanced design of the ones described in Section 3.2.1 and feature three force-controlled tendon-driven fingers and an opposable thumb. On top of the torso is still the *iCat* [10] animatronic head that can rotate, produce facial expressions, and output sound. The sensing was updated and the system now includes two stereo color cameras, a depth camera, and a microphone array. This setup was used in the initial development and during the first user studies. As the industrial manipulators of this version are not safe for physical human-robot interaction, an improved version was developed during the project once the rest of the enhanced hardware of Section 3.2.1 became available for purchase. This setup includes one manipulator with seven DOF and a torso with 3 DOF, both featuring *Series Elastic Actuators* [61] and cable-driven design [70]. The head uses a custom software on a tablet that is able to produce facial expressions and lip synchronization on a browser-based interface.

The architecture of the distributed system is shown in Fig. 13. From a high-level view, it generally follows the sense-plan-act pattern. Contrary to most human-humanoid interaction approaches, dialogs are not managed by a component that

Fig. 12a First version of the James robot using large parts of the hardware setup of the system shown in Section 3.1.

Fig. 12b Second version of James with an enhanced design of the hardware described in Section 3.2.1.

follows rules or a state machine, but dialog actions result from generic knowledge-level planning [59].

From sensing to action, the software components for visual and speech recognition, planning, and speech output and robot manipulation interact as follows: Visual classification observes the bar area with both stereo cameras and an infrared-light depth camera. It segments moving image areas by color and detects faces and hands. Blob tracking allows people to be re-identified after an occlusion. At a lower rate, a face detector is executed to update appearance models and achieve more robustness against changes in illumination. When a person has been recognized and tracked, their body pose is estimated by model-based fitting of primitive shapes in color and depth images [23].

For recognizing speech, the system uses the Microsoft Speech API with a simple, domain-specific vocabulary for the two languages English and German. Even though the microphone array receives a high level of background noise, for instance from other guests, the limited vocabulary allows adequate recognition rates. For parsing and understanding recognized text and generating spoken language output, a bi-directional, bilingual OpenCCG grammar is used.

A state manager filters and fuses frame-based visual recognition and sentence-based speech understanding results and passes the symbolic world state to the automated planner. The automated planner manages all interaction with customers and generates sequences of actions for communication and manipulation. To generate a plan, it performs a forward search with knowledge-level reasoning (Section 3.3.3). Changes to the world state are monitored to detect infeasible plans and trigger re-planning.

On the output side, the text-to-speech output is synchronized with facial expressions and eye movement of the animatronic head to establish eye contact with the addressed customer. When a drink is to be served, the robot motion planner finds

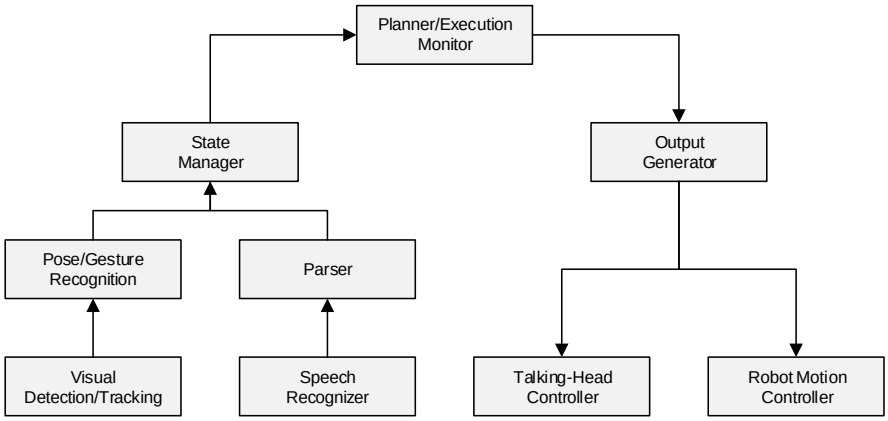

Fig. 13 Software architecture for multi-party interaction used in the humanoid James. Input from various sensors is evaluated and processed by a *State Manager*. A combined *Planner* and Execution Monitor then generates actions for the *Output Generator*, that coordinates head and body output.

reachable grasp configurations and a collision-free trajectory. Because drinks are to be placed close to a customer, trajectories need to be computed and verified at run-time. To enhance the efficiency of robot-robot and robot-environment collision checks, convex hulls of all rigid objects are precomputed, which enables checks in less than a millisecond [23].

3.3.2 Visual Intention Recognition

In a social setting like the bartending domain, the intention and action of guests follows patterns, and learning these patterns may enhance the quality of the interaction between human and humanoid. For example, most of the guests will order shortly after entering the scene. However, people express their intention not necessarily in a spoken dialog. Through empiric studies and interpretation of human-human interactions in the same domain, it is known that body posture is the most crucial signal to initiate an interaction [27]. In addition, head pose serves as signal for beginning and ending an interaction. When an interaction is ending, most customers turn their head downwards or sidewards. In sum, body orientation, body posture, head pose, and—to a little extent—position relative to a group of people are the relevant signals to show the intent to speak to or end an interaction with the humanoid [27].

Based on the human-human studies, a Hidden Markov Model (HMM) was trained to detect interaction states between a human guest and the humanoid robot from visual features. Body posture and the relative position to other people was available from a depth camera and reconstructed skeletons. Head poses were recognized on color imagery through identification of individual facial features by Haar-feature classification. Estimated positions of eyes, nose, and mouth were processed by a 3-layer artificial neural network (ANN), which was previously trained on anno-

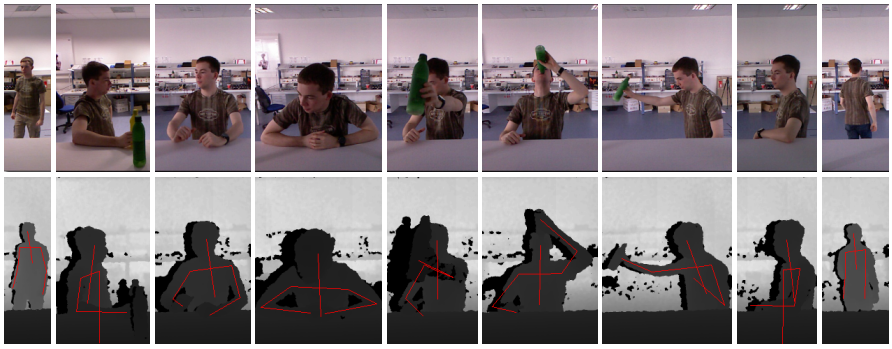

Fig. 14 Camera images (top) and corresponding skeleton poses (bottom) of Kinect sensor used for visual intention recognition. The recognized states in the bar scenario include (from left to right): entering bar setting, leaning on bar with right side, requesting attention, raising glass with left hand, drinking, raising glass with right hand, leaning on bar with left side, and leaving bar setting.

tated photographs to finally output pitch, yaw, and roll angles of recognized faces. Because the recognition problem of social interaction states is temporally coherent and depends on previous states, a Hidden Markov Model was selected for training and recognition. A fine-grained model with eight states was chosen to detect whether a person was idle, interacting with the humanoid, interacting with another person, reading the menu, entering, leaving, clinking glasses, or drinking (Fig. 14). More than 200 scenes with a total of 1720 states were enacted, recorded, and manually labeled. Of these data, two thirds were used for training. Comparing precision and recall scores on a small cross-validating data set, a suitable feature set of 19 real values was selected, including torso and hand positions, body alignment, head pose (both as a normal vector and as pitch and yaw angles), and two fuzzy features that responded to a position close to other people. Possible graph structures of the model were again evaluated on the cross-validation data set, from which a graph with three linear hidden states for each interaction state and a feature emission model with a mixture of three full covariance matrices was selected.

On the test set, the visual intention recognition system recognized 82.9% of the interaction states correctly. Compared to the simpler rule-based classifier that was used in all user studies to detect engagement with the humanoid, the HMM approach did not provide more reliable results, but could be trained for a more complex interaction model to recognize eight different interaction states and gestures instead of just two.

3.3.3 Planning Interaction and Dialog Management

The James humanoid does not manage dialogs and interaction with rules or state machines, but generates its manipulation and speech actions with automated planning [59]. In particular, the *Planning with Knowledge and Sensing* (PKS) planner

is used, which is a general-purpose planner that can reason with incomplete knowledge. Because the bartending scenario allows multiple customers and natural language understanding is imperfect, the PKS planner with its knowledge-level reasoning and its execution monitor is well suited.

In the bartending scenario, the goal of the planner is to ensure that all agents that have sought attention will be served [59]. The knowledge state of the world is provided by the visual detection, pose recognition, and speech understanding components after filtering by the state manager. The state is modeled by a set of predicates that are true, false, or unknown. Unary predicates include whether a customer seeks attention, has been greeted, has ordered, has been served, or was not understood by speech recognition. A function with a parameter *customer* models which drink a customer has requested. The planner then searches for a sequence of actions that lead to state that fulfills the goal. Actions include both speech and facial expressions (greeting a customer, asking for an order, acknowledgment, asking a customer to wait) as well as manipulation actions to serve drinks.

Contrary to other automated planners [66], PKS operates on the knowledge-level [60]. For example, an effect of the ask-for-an-order action is that the customer's wish becomes known, which is necessary to fulfill a precondition for the serve action. Therefore, question-and-answer dialog actions are not tasks themselves, but rather planned to gather knowledge that is necessary to fulfill the goal. Of course, the world state is continuously monitored and updated with sensor input by the state manager. Replanning only occurs if the preconditions of planned actions are no longer fulfilled by the world state. In the experiments, planning usually took less than 0.1 seconds.

In contrast to related works in automated dialog planning [11], the described planner can reason with incomplete knowledge, which is useful for planning questions, and is fully integrated in James' interactive humanoid system.

3.3.4 Task and Motion Planning for Interactive Manipulation

When a robot interacts with humans in a domain where objects can be moved and manipulated, the state space is already so large that simple state machines or rules will not solve the planning problem in full generality. Actions as simple as pick-and-place have intricate preconditions and effects, for instance, which objects have to be moved first to avoid collisions and to achieve a certain goal. As an example, placing an object at a random location may block a picking action for the other robot arm several steps later in a plan, something which cannot be predicted by rules or heuristics. In addition, randomized path planners can only fail inconclusively with a timeout, therefore a high-level, discrete search that calls path planning as a subroutine cannot be complete.

Rather than following greedy heuristics that cannot solve every case, a combined task and motion planner was developed to search in the full space of discrete actions and motion paths to solve generic pick-and-place tasks [29, 28]. The integrated planner can start from an arbitrary world state, such as one where random bottles

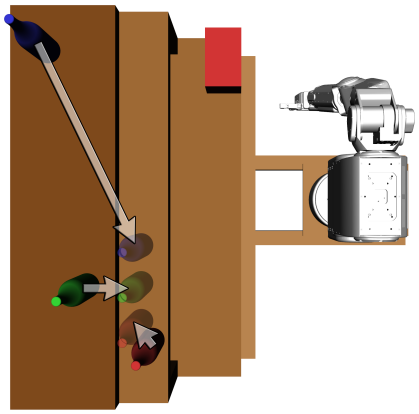

Fig. 15a Top view of a scenario where the robot has to rearrange three bottles by grasping or pushing as indicated by the arrows.

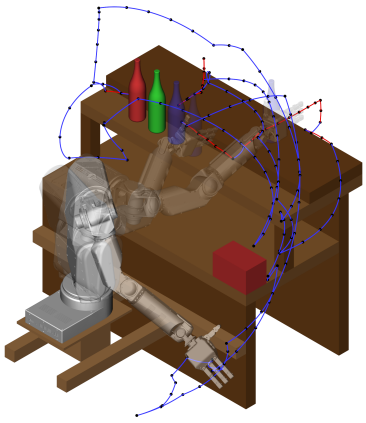

Fig. 15b Solution path with intermediate robot configurations. Edges drawn in *red* correspond to robot manipulation actions, *blue* ones to robot motion only.

have been placed on a bar table by human interaction partners. It then searches for a sequence of actions to fulfill a goal condition, such as removing all bottles from the bar table. The planner accepts a formal definition of actions, such as picking or placing an object or moving one of the two robot arms, that are defined in terms of discrete preconditions, random generators for geometric choices, geometric preconditions (such as collision avoidance), and effects. The planner then progresses the search, both by evaluating additional actions in the search graph and by trying out additional geometric choices from existing states, until a state is found that fulfills the goal.

In practical experiments with the bimanual humanoid, it was easy to identify cases that could only be solved by such generic search. Because the James robot bartender has two arms with different workspaces, cleaning the bar table requires non-trivial sequences of actions. As an example, when a customers has placed an empty bottle on the right side of the bar table, only one arm can pick it up, but it needs to put it down in the common workspace where the other arm can reach it to transfer it to the empty bottle storage region [29]. To allow concise definition of scenarios and goal conditions, predicates to model support surfaces and the inclusion relation are also available [28]. The search takes a few seconds for scenarios with four objects and two arms, with most of the time spent on collision checking. In general, evaluation on multiple scenarios has shown that integrated task and motion planning can solve pick-and-place tasks, even those that require both arms or sequences of more than ten actions.

In an alternate approach, sampling-based motion planning can be extended with Diverse Action Manipulation (DAMA) [41]. In addition to simple robot motion, the following diverse action manipulations are available: picking up an object, transferring the rigidly attached object, pushing an object with the interior, and pushing an

object with the exterior surface of the hand. These manipulation actions induce various hand poses, including grasping and both pushing poses. Fig.15a and 15b show a scenario, in which the robot has to rearrange three bottles. Considering only translations for objects, the search space for this scenario is 18-dimensional, consisting of important subspaces induced by the constraints of manipulation actions defined for this configuration space.

4 Conclusions and Future Directions

In this chapter, we have shown the potential of human-robot physical interaction, its state of development, its deficiencies, and the results of extensive real-world experimentation. We have seen that performing this highly interdisciplinary type of research—ranging from user-studies and neuroscience investigations to the theory of robot motion planning all the way to a robust systems implementation in its entirety—is absolutely vital for the progress of the field. Over the many years of performing research into direct physical robot-human interaction, we have seen that the following, non-exhaustive list of research areas can provide the basis for further exciting research on our way to practical systems:

- Advances in speech output should specifically be used in the generation of referring expressions in situated dialog.
- More expansive and complex hand-over experiments can reveal many interesting properties of human behavior and how a robot should adapt to the human partner.
- fMRI studies and neuroscience experiments can potentially reveal links between brain activity, e.g., in the mirror neurons, and joint action.
- Error recognition and error handling is an absolute must, it may be even more important than the planning of the original task.
- The collection of a multimodal data corpora of human-human task-based joint action and making them accessible to automatic processing though powerful ontologies can dramatically improve the skill sets of the robot co-worker.

Without this underpinning research, robotic co-working will never find the necessary acceptance to evolve into a mass market. However, if made accessible as a base platform for roboticists and industry developers, it can be the starting point for another robot application success story.

References

1. Ambrose, R., Aldridge, H., Askew, R., Burridge, R.R., Bluethmann, W., Diftler, M., Lovchik, C., Magruder, D., Rehnmark, F.: Robonaut: NASA's space humanoid. IEEE Intelligent Systems and their Applications **15**(4), 57–63 (2000). DOI 10.1109/5254.867913
2. Arkin, R.C.: Behavior-Based Robotics. Intelligent Robotics and Autonomous Agents. MIT Press (1998)

3. Aryananda, L., Weber, J.: MERTZ: A quest for a robust and scalable active vision humanoid head robot. In: Proceedings of the IEEE/RAS International Conference on Humanoid Robots, pp. 513–532. Santa Monica, CA, USA (2004). DOI 10.1109/ICHR.2004.1442668
4. Bard, E.G., Hill, R.L., Foster, M.E., Araia, M.: Tuning accessibility of referring expressions in situated dialogue. Language, Cognition and Neuroscience **29**(8), 928–949 (2014). DOI 10.1080/23273798.2014.895845
5. Bekkering, H., de Bruijn, E.R.A., Cuijpers, R.H., Newman-Norlund, R., van Schie, H.T., Meulenbroek, R.: Joint action: Neurocognitive mechanisms supporting human interaction. Topics in Cognitive Science **1**(2), 340–352 (2009). DOI 10.1111/j.1756-8765.2009.01023.x
6. Bender, M., Braun, M., Rally, P., Scholtz, O.: Lightweight Robots in Manual Assembly – Best to Start Simply! Examining Companies' Initial Experiences with Lightweight Robots. Fraunhofer Institute for Industrial Engineering, Stuttgart, Germany (2016)
7. Bicho, E., Louro, L., Hipólito, N., Erlhagen, W.: A dynamic field approach to goal inference and error monitoring for human-robot interaction. In: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction, Adaptive and Emergent Behaviour and Complex Systems Convention, pp. 31–37. Edinburgh, Scotland (2009)
8. Billard, A., Calinon, S., Dillmann, R., Schaal, S.: Robot programming by demonstration. In: B. Siciliano, O. Khatib (eds.) Springer Handbook of Robotics, first edn., chap. 59, pp. 1371–1394. Springer, Heidelberg, Germany (2008)
9. Breazeal, C., Brooks, A., Chilongo, D., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A.: Working collaboratively with humanoid robots. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots, pp. 253–272. Santa Monica, CA, USA (2004). DOI 10.1109/ICHR.2004.1442126
10. van Breemen, A., Yan, X., Meerbeek, B.: iCat: An animated user-interface robot with personality. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 143–144. Utrecht, Netherlands (2005). DOI 10.1145/1082473.1082823
11. Brenner, M., Kruijff-Korbayová, I.: A continual multiagent planning approach to situated dialogue. In: Proceedings of the Workshop on Semantics and Pragmatics of Dialogue. London, UK (2008)
12. Cangelosi, A., Ogata, T.: Speech and language in humanoid robots. In: Humanoid Robotics: A Reference. Springer
13. Carletta, J., Hill, R.L., Nicol, C., Taylor, T., de Ruiter, J.P., Bard, E.G.: Eyetracking for two-person tasks with manipulation of a virtual world. Behavior Research Methods **42**(1), 254–265 (2010). DOI 10.3758/BRM.42.1.254
14. Clark, H.H.: Pointing and placing. In: S. Kita (ed.) Pointing: Where Language, Culture, and Cognition Meet. Lawrence Erlbaum Associates (2003)
15. Clark, H.H., Wilkes-Gibbs, D.: Referring as a collaborative process. Cognition **22**(1), 1–39 (1986). DOI 10.1016/0010-0277(86)90010-7
16. Curioni, A., Knoblich, G., Sebanz, N.: Joint action in humans – A model for human-robot interactions? In: Humanoid Robotics: A Reference. Springer
17. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science **19**(2), 233–263 (1995). DOI 10.1207/s15516709cog1902_3
18. Diftler, M.A., Mehling, J.S., Abdallah, M.E., Radford, N.A., Bridgwater, L.B., Sanders, A.M., Askew, R.S., Linn, D.M., Yamokoski, J.D., Permenter, F.A., Hargrave, B.K., Platt, R., Savely, R.T., Ambrose, R.O.: Robonaut 2 – The first humanoid robot in space. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 2178–2183. Shanghai, China (2011). DOI 10.1109/ICRA.2011.5979830
19. Edsinger, A.: Robot manipulation in human environments. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (2007)
20. Edsinger, A., Kemp, C.C.: Two arms are better than one: A behavior based control system for assistive bimanual manipulation. In: S. Lee, I.H. Suh, M.S. Kim (eds.) Recent Progress in Robotics: Viable Robotic Service to Human, *Lecture Notes in Control and Information Sciences*, vol. 370, pp. 345–355. Springer (2007). DOI 10.1007/978-3-540-76729-9_27

21. Foster, M.E., Bard, E.G., Hill, R.L., Guhe, M., Oberlander, J., Knoll, A.: The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In: Proceedings of the ACM/IEEE International Conference on Human Robot Interaction, pp. 295–302. Amsterdam, Netherlands (2008). DOI 10.1145/1349822.1349861
22. Foster, M.E., Gaschler, A., Giuliani, M.: How can I help you? Comparing engagement classification strategies for a robot bartender. In: Proceedings of the International Conference on Multimodal Interaction, pp. 255–262. Sydney, Australia (2013). DOI 10.1145/2522848.2522879
23. Foster, M.E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., Petrick, R.P.A.: Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 3–10. Santa Monica, CA, USA (2012). DOI 10.1145/2388676.2388680
24. Foster, M.E., Giuliani, M., Isard, A.: Task-based evaluation of context-sensitive referring expressions in human-robot dialogue. Language, Cognition and Neuroscience **29**(8), 1018–1034 (2014). DOI 10.1080/01690965.2013.855802
25. Foster, M.E., Giuliani, M., Müller, T., Rickert, M., Knoll, A., Erlhagen, W., Bicho, E., Hipólito, N., Louro, L.: Combining goal inference and natural-language dialogue for human-robot joint action. In: Proceedings of the International Workshop on Combinations of Intelligent Methods and Applications, European Conference on Artificial Intelligence. Patras, Greece (2008)
26. Foster, M.E., Matheson, C.: Following assembly plans in cooperative, task-based human-robot dialogue. In: Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue. London, UK (2008)
27. Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., Knoll, A.: Social behavior recognition using body posture and head pose for human-robot interaction. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2128–2133. Vilamoura, Portugal (2012). DOI 10.1109/IROS.2012.6385460
28. Gaschler, A., Kessler, I., Petrick, R.P.A., Knoll, A.: Extending the knowledge of volumes approach to robot task planning with efficient geometric predicates. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3061–3066. Seattle, WA, USA (2015). DOI 10.1109/ICRA.2015.7139619
29. Gaschler, A., Petrick, R.P.A., Giuliani, M., Rickert, M., Knoll, A.: KVP: A knowledge of volumes approach to robot task planning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 202–208. Tokyo, Japan (2013). DOI 10.1109/IROS.2013.6696354
30. Giuliani, M., Knoll, A.: MultiML – A general purpose representation language for multimodal human utterances. In: Proceedings of the IEEE International Conference on Multimodal Interfaces, pp. 165–172. Chania, Crete, Greece (2008). DOI 10.1145/1452392.1452424
31. Giuliani, M., Petrick, R., Foster, M.E., Gaschler, A., Isard, A., Pateraki, M., Sigalas, M.: Comparing task-based and socially intelligent behaviour in a robot bartender. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 263–270. Sydney, Australia (2013). DOI 10.1145/2522848.2522869
32. Glasauer, S., Huber, M., Basili, P., Knoll, A., Brandt, T.: Interacting in time and space: Investigating human-human and human-robot joint action. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, pp. 252–257. Viareggio, Italy (2010). DOI 10.1109/ROMAN.2010.5598638
33. Haddadin, S., Albu-Schäffer, A., Hirzinger, G.: Requirements for safe robots: Measurements, analysis and new insights. The International Journal of Robotics Research **28**(11–12), 1507–1527 (2009). DOI 10.1177/0278364909343970
34. Henning, M.: A new approach to object-oriented middleware. IEEE Internet Computing **8**(1), 66–75 (2004). DOI 10.1109/MIC.2004.1260706
35. Hirukawa, H., Kanehiro, Kaneko, K., Kajita, S., Fujiwara, K., Kawai, Y., Tomita, F., Hirai, S., Tanie, K., Isozumi, T., Akachi, K., Kawasaki, T., Ota, S., Yokoyama, K., Handa, H., Fukase, Y., ichiro Maeda, J., Nakamura, Y., Tachi, S., Inoue, H.: Humanoid robotics platforms developed in HRP. Robotics and Autonomous Systems **48**(4), 165–175 (2004). DOI 10.1016/j.robot.2004.07.007

36. Huber, M., Knoll, A., Brandt, T., Glasauer, S.: When to assist? – Modelling human behaviour for hybrid assembly systems. In: Proceedings of the International Symposium and the German Conference on Robotics, pp. 165–170. Munich, Germany (2010)
37. Huber, M., Radrich, H., Wendt, C., Rickert, M., Knoll, A., Brandt, T., Glasauer, S.: Evaluation of a novel biologically inspired trajectory generator in human-robot interaction. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, pp. 639–644. Toyama, Japan (2009). DOI 10.1109/ROMAN.2009.5326233
38. Huber, M., Rickert, M., Knoll, A., Brandt, T., Glasauer, S.: Human-robot interaction in handing-over tasks. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, pp. 107–112. Munich, Germany (2008). DOI 10.1109/ROMAN.2008.4600651
39. Hulstijn, M., Meulenbroek, R., Wijers, M., de Ruiter, J.P.: A frequency analysis of joint-action primitives. Deliverable D2.3, EU FP6 IST Cognitive Systems Integrated Project JAST (FP6-003747-IP) (2005)
40. Iqbal, T., Riek, L.D.: Human robot coordination. In: Humanoid Robotics: A Reference. Springer
41. Jentzsch, S., Gaschler, A., Khatib, O., Knoll, A.: MOPL: A multi-modal path planner for generic manipulation tasks. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 6208–6214. Hamburg, Germany (2015). DOI 10.1109/IROS.2015.7354263
42. Jindai, M., Shibata, S., Yamamoto, T., Shimizu, A.: A study on robot-human system with consideration of individual preferences. JSME International Journal Series C Mechanical Systems, Machine Elements and Manufacturing **46**(3), 1075–1083 (2003). DOI 10.1299/jsmec.46.1075
43. Kazerooni, H.: Exoskeletons for human performance augmentation. In: B. Siciliano, O. Khatib (eds.) Springer Handbook of Robotics, first edn., chap. 33, pp. 773–793. Springer, Heidelberg, Germany (2008)
44. Khan, S.G., Bendoukha, S., Mahyuddin, M.N.: Dynamic control for human-humanoid interaction. In: Humanoid Robotics: A Reference. Springer
45. Kirgis, F.P., Katsos, P., Kohlmaier, M.: Collaborative robotics. In: D. Reinhardt, R. Saunders, J. Burry (eds.) Robotic Fabrication in Architecture, Art and Design, pp. 448–453. Springer (2016). DOI 10.1007/978-3-319-26378-6_36
46. Knoll, A., Hildenbrandt, B., Zhang, J.: Instructing cooperating assembly robots through situated dialogues in natural language. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 888–894. Albuquerque, NM, USA (1997). DOI 10.1109/ROBOT.1997.620146
47. Kock, S., Vittor, T., Matthias, B., Jerregard, H., Källman, M., Lundberg, I., Mellander, R., Hedelind, M.: Robot concept for scalable, flexible assembly automation: A technology study on a harmless dual-armed robot. In: Proceedings of the IEEE International Symposium on Assembly and Manufacturing. Tampere, Finland (2011). DOI 10.1109/ISAM.2011.5942358
48. Kosuge, K., Sato, M., Kazamura, N.: Mobile robot helper. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 583–588. San Francisco, CA, USA (2000). DOI 10.1109/ROBOT.2000.844116
49. Lay, K., Prassler, E., Dillmann, R., Grunwald, G., Hägele, M., Lawitzky, G., Stopp, A., von Seelen, W.: MORPHA: Communication and interaction with intelligent, anthropomorphic robot assistants. In: Tagungsband Statustage Leitprojekte Mensch-Technik-Interaktion in der Wissensgesellschaft. Saarbrücken, Germany (2001)
50. Lohan, K.S., Lehmann, H., Dondrup, C., Brooz, F., Kose, H.: Enriching the human-robot interaction loop with natural, semantic and symbolic gestures. In: Humanoid Robotics: A Reference. Springer
51. Loth, S., Huth, K., Ruiter, J.P.D.: Automatic detection of service initiation signals used in bars. Frontiers in Psychology **4**(557) (2013). DOI 10.3389/fpsyg.2013.00557
52. Loth, S., Jettka, K., Giuliani, M., de Ruiter, J.P.: Ghost-in-the-machine reveals human social signals for human-robot interaction. Frontiers in Psychology **6**(1641) (2015). DOI 10.3389/fpsyg.2015.01641

53. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text **8**(3), 243–281 (1988). DOI 10.1515/text.1.1988.8.3.243
54. Mansard, N., Stasse, O., Evrard, P., Kheddar, A.: A versatile generalized inverted kinematics implementation for collaborative working humanoid robots: The stack of tasks. In: Proceedings of the International Conference on Advanced Robotics, pp. 1–6. Munich, Germany (2009)
55. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press (1992)
56. Meulenbroek, R.G.J., Bosga, J., Hulstijn, M., Miedl, S.: Joint-action coordination in transferring objects. Experimental Brain Research **180**(2), 333–343 (2007). DOI 10.1007/s00221-007-0861-z
57. Müller, T., Ziaie, P., Knoll, A.: A wait-free realtime system for optimal distribution of vision tasks on multicore architectures. In: Proceedings of the International Conference on Informatics in Control, Automation and Robotics, pp. 301–306. Funchal, Portugal (2008)
58. Panin, G., Ladikos, A., Knoll, A.: An efficient and robust real-time contour tracking system. In: Proceedings of the IEEE International Conference on Computer Vision Systems, pp. 44–51. New York, NY, USA (2006). DOI 10.1109/ICVS.2006.13
59. Petrick, R., Foster, M.E.: Planning for social interaction in a robot bartender domain. In: Proceedings of the International Conference on Automated Planning and Scheduling, pp. 389–397. Rome, Italy (2013)
60. Petrick, R.P.A., Bacchus, F.: Extending the knowledge-based approach to planning with incomplete information and sensing. In: Proceedings of the International Conference on Automated Planning and Scheduling, pp. 2–11. Whistler, BC, Canada (2004)
61. Pratt, G.A., Williamson, M.M.: Series elastic actuators. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 399–406. Pittsburgh, PA, USA (1995). DOI 10.1109/IROS.1995.525827
62. Rickert, M.: Efficient motion planning for intuitive task execution in modular manipulation systems. Dissertation, Technical University of Munich, Munich, Germany (2011)
63. Rickert, M., Foster, M.E., Giuliani, M., By, T., Panin, G., Knoll, A.: Integrating language, vision and action for human robot dialog systems. In: Proceedings of the International Conference on Universal Access in Human-Computer Interaction, HCI International, *Lecture Notes in Computer Science*, vol. 4555, pp. 987–995. Springer, Beijing, China (2007). DOI 10.1007/978-3-540-73281-5_108
64. de Ruiter, J.P.: The production of gesture and speech. In: D. McNeill (ed.) Language and Gesture, pp. 284–311. Cambridge University Press (2000). DOI 10.1017/CBO9780511620850
65. de Ruiter, J.P., Bangerter, A., Dings, P.: The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. Topics in Cognitive Science **4**(2), 232–248 (2012). DOI 10.1111/j.1756-8765.2012.01183.x
66. Russell, S., Norvig, P.: Classical planning. In: Artificial Intelligence: A Modern Approach, third edn., chap. 10. Pearson (2010)
67. Sandini, G., Sciutti, A., Rea, F.: Movement-based communication for humanoid-human interaction. In: Humanoid Robotics: A Reference. Springer
68. Schraft, R.D., Meyer, C., Parlitz, C., Helms, E.: PowerMate – A safe and intuitive robot assistant for handling and assembly tasks. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 4074–4079. Barcelona, Spain (2005). DOI 10.1109/ROBOT.2005.1570745
69. So, W.C., Kita, S., Goldin-Meadowa, S.: Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. Cognitive Science **33**(1), 115–125 (2009). DOI 10.1111/j.1551-6709.2008.01006.x
70. Townsend, W.T., Salisbury, J.K.: Mechanical design for whole-arm manipulation. In: Robots and Biological Systems: Towards a New Bionics?, *NATO ASI Series*, vol. 102. Springer (1993). DOI 10.1007/978-3-642-58069-7_9
71. Traum, D.R., Larsson, S.: The information state approach to dialogue management. In: J. van Kuppevelt, R.W. Smith (eds.) Current and New Directions in Discourse and Dialogue, *Text,*

Speech and Language Technology, vol. 22, pp. 325–353. Springer (2003). DOI 10.1007/978-94-010-0019-2_15

72. Ureche, A.L.P., Umezawa, K., Nakamura, Y., Billard, A.: Task parameterization using continuous constraints extracted from human demonstrations. IEEE Transactions on Robotics **31**(6), 1458–1471 (2015). DOI 10.1109/TRO.2015.2495003
73. Versace, J.: A review of the severity index. Tech. Rep. 710881, SAE International (1971). DOI 10.4271/710881
74. Yokoyama, K., Handa, H., Isozumi, T., Fukase, Y., Kaneko, K., Kanehiro, F., Kawai, Y., Tomita, F., Hirukawa, H.: Cooperative works by a human and a humanoid robot. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 2985–2991. Taipei, Taiwan (2003). DOI 10.1109/ROBOT.2003.1242049
75. Ziaie, P., Müller, T., Foster, M.E., Knoll, A.: A naïve Bayes classifier with distance weighting for hand-gesture recognition. In: Proceedings of the International CSI Computer Conference. Kish Island, Iran (2008)