Toward Multimodal Human-Robot

Cooperation and Collaboration

Dennis Perzanowski,[*] Derek Brock[†]
Naval Research Laboratory, Washington, DC, 20375

Magdalena Bugajska, Scott Thomas, Donald Sofge, William Adams,[‡]
Naval Research Laboratory, Washington, DC, 20375

Marjorie Skubic[§], Sam Blisard[**]
University of Missouri, Columbia, MO, 65211

Nicholas Cassimatis[††]
Rensselaer Polytechnic Institute, Troy, NY, 12180

J. Gregory Trafton[‡‡] and Alan C. Schultz[§§]
Naval Research Laboratory, Washington, DC, 20375

Our multimodal interface integrates speech recognition, natural language understanding, spatial reasoning and human cognitive models for completing specific tasks and for perspective-taking in locative oriented tasks. With natural language and gestures, we believe human-robot interaction and communication is facilitated. Instead of concentrating on the various modalities of the interface, users can concentrate on the task at hand. Likewise, by incorporating human cognitive models for handling spatial information and perspective-taking, as well as for specific task completion, a better match with the expectations that humans acquire from their human-human interactions should be obtained, further facilitating cooperation and collaboration in human-robot interactions.

Nomenclature

ATRV = All TeRrain Vehicle
EUT = End User Terminal
PDA = Personal Digital Assistant

I. Introduction

OUR research in human-robot interaction is based on two essential ideas. First, we expect that natural means of communication, such as people's ability to use natural language and gestures, will reduce the user's learning curve for learning the interface, and reduce the cognitive burden on the user in subsequent regular use. Secondly, we believe that cognitive models are essential to interpret human speech and actions well, and to produce robotic

[*]Computational Research Linguist, Intelligent Multimodal/Multimedia Systems, Code 5512, and Professional Member.
[†]Computer Scientist, Interface Design and Evaluation, Code 5513.
[‡]Computer Scientists, Intelligent Systems, Code 5515.
[§]Associate Professor, Electrical and Computer Engineering Department and Computer Science Department.
[**]Graduate Student, Electrical and Computer Engineering Department.
[††]Assistant Professor, Department of Cognitive Science.
[‡‡]Computer Scientist, Intelligent Systems, Code 5515.
[§§]Program Manager, Intelligent Systems, Code 5515.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Toward Multimodal Human-Robot Cooperation and Collaboration	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory ,Intelligent Multimodal/Multimedia Systems, Code 5512,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
AIAA 1st Intelligent Systems Technical Conference. American Institute of Aeronautics and Astronautics, Chicago, IL, 20-22 Sep 2004.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

responses that seem natural and felicitous to the human members of the interaction. Humans anticipate how other humans will act/react in situations. They become baffled when these anticipated results are thwarted or are lacking. We believe that our robots, by being able to rationalize about situations and react in certain ways similar to human behavior, will act more autonomously and/or cooperatively because human users of the system will anticipate that the robots can act, react and even reason about situations in ways similar to humans.

In this paper, we outline the various modalities of our human-robot interface and summarize how the various modules are integrated on several mobile platforms. Our research involves collaboration with several other laboratories and universities. In-house, we use mobile Nomad 200 and B21r research robots named Coyote, Roadrunner, GRACE[1,2], and George[3***]. The latter two are also used in collaboration with Carnegie Melon University. We also have a team of more robust in-house ATRV-Jrs. named Magneto and Wolverine. In collaboration with research at MIT, we work with a stationary robot, Leonardo[4]. Finally, in collaborative work with the Media Lab at MIT, the University of Massachusetts at Amherst, the University of Southern California at Los Angeles, Vanderbilt University, and NASA-Johnson Space Center, we use a humanoid robot, Robonaut[5,6]. Depending upon the platform and application, we are collaborating on interface and cognitive issues[7,8] in order to achieve easy interaction and facilitate cooperation and collaboration. No matter the application or platform, human users can interact with our robots by speaking commands or requesting information. Spoken input is natural, and the robustness of our natural language understanding system permits a wide variety of utterances that allow a high degree of paraphrasability. Context is also stored on a very rudimentary level and limited dialoging capabilities are available.

In most of our applications, we permit two natural gestures employing the user's arm or hand. Pointing gestures can be used to indicate locations and objects, both in the real world and on touch-sensitive monitors or handheld devices. Gestures can also segment space, indicating distances, paths to follow, or areas to explore. Visually recognized objects can be labeled for future ease of reference. For this latter work, we have integrated a Spatial Reasoning component with our natural language capabilities; we also employ human cognitive models for task completion and perspective-taking. Finally, to test our multimodal interface empirically, we have been conducting human-subject experiments. A preliminary analysis of the data seems to indicate that humans find the naturalness of our interface lends itself to cooperative behaviors and interactions with a mobile robot.

II. The Multimodal Interface

A. Overview

The focus of our work in multimodal interactions with robot platforms[9,10] has been to provide natural and/or easy modes of interaction for humans. The human operator can command and query the robot using everyday English. Depending upon the particular application and robot, the natural language component allows the user to direct the robot to navigate to and around different locations and interact with objects, identifying their presence and location in responses phrased in everyday English, and even allow the human collaborator to name the objects for future ease of reference. Where appropriate to the application, the interface can allow the user to ask a robot (such as Leonardo) to touch buttons, or pick up wrenches and turn bolts (as with Robonaut). Users can also query the robot about information, accessible through links to different knowledge bases, either available on the internet or specially constructed for a particular application. Queries of this sort can be questions about the weather, or in another context, requests for user information about a conference or convention that the user and the robot are attending (e.g. GRACE and George). In this latter regard, the robot can act as a receptionist or even act as a guide around a convention center. In each of the various applications, however, the same basic architecture is used. Modifications usually take the form of adding or subtracting specific domain information to or from the natural language processing components, which then must be mapped to specific function calls appropriate for that domain when the interface is ported to the new application.

For gestures, the interface permits both "natural" and "synthetic" gestures. So-called natural gestures are those that people normally make while conversing, whereas synthetic gestures are actions such as strokes and taps that can be made on devices such as PDAs and touch screens on EUTs. We limit ourselves to two types of natural gestures that signify vectoring and segmenting actions. Vectoring gestures usually accompany an utterance like "Go over there," for which people will typically use finger-pointing, arm movements, and even movements of the head and eyes to indicate the intended direction. In this regard, we further limit our natural gestures to arm movements.

[***]Several of our robots are named after cartoon or comic book characters. CMU named GRACE. The name is an acronym for Graduate Robot Attending the ConferencE. George was named as GRACE's counterpart in keeping with two classic American icons of comedy. MIT named their robot Leonardo in honor of Leonardo daVinci.

Segmenting gestures usually accompany utterances such as "Move forward this far," in which humans typically demonstrate an imaginary line segment in front of them with both hands. We use gestures solely to disambiguate utterances of the types just mentioned. If the robot needs further information or isn't able to understand the initial utterance, the interface responds verbally, using voice synthesis.

Figure 1 shows the system architecture for our multimodal interface for autonomous mobile robots. Slightly different versions, ones which do not employ the PDA/EUT components, for example, are used when interacting with Leonardo, Robonaut or with one of our "conference robots," GRACE and George.

Because spatial information and the location of objects play important parts in navigation, spatial relations and spatial knowledge play important parts in the analysis of human-robot communication. Even when navigation per se is not a factor, in such stationary robots as Robonaut, or in the receptionist version of GRACE, spatial relations are still important considerations for human-robot interactions. For instance, if a person asks a stationary robot where a tool is on a workbench in front of it, it must be able either to point to it or to describe its location. In addition, it is important for this information to be provided in a form that is easy for the person to interpret and use. Our Spatial Relations Component provides sensor readings using its array of on-board sensors[11]. This information is mapped to the real-time plan view map in one of the robotic subcomponents. (See Fig. 2). However, verbal information is also provided to the user via a subcomponent of the Spatial Relations Component[12], which gives the human collaborator natural and easily understood information about the environment. Instead of returning information about objects in cryptic digital arrays, radians, or sets of x,y coordinates, this information is mapped in the Spatial Relations Component to utterances that are clearly modeled on English responses. For example, if a human asks one of the mobile robots, "What objects do you see?" the robot returns a comprehensible utterance such as, "There are two objects. Object number one is directly in front of me, and object number two is in front of me but slightly to my right." This provides the relevant information in terms that are consistent and more easily mapped to the user's normal way of representing spatial relations. Of course, when and if the need arises, the user always has access to more precise information, for purposes of comparison, which can be displayed on a computer terminal both numerically and graphically, and can also listen to a synthesized verbal presentation of sensor information. Furthermore, to make interactions with the robot and the objects in its environment even easier for human collaborators, we have incorporated a labeling functionality that allows users to name objects in the robot's environment to facilitate future references. After telling the robot, for instance, that "Object number two is a pillar," the user can subsequently converse with the robot about the pillar instead of having to remember to refer to it as "object two."

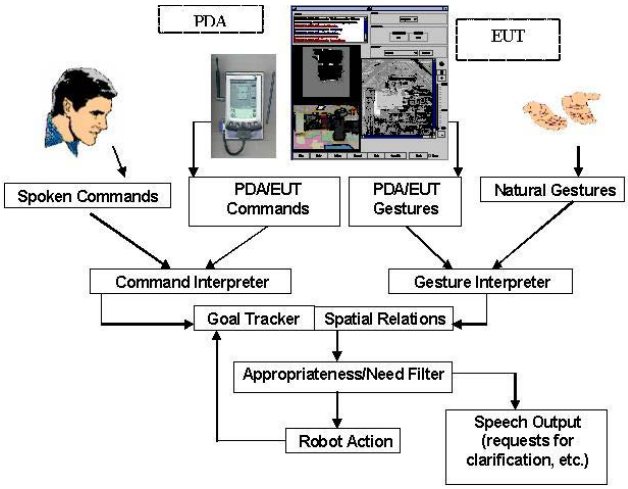

Figure 1. Architecture of the Multimodal Interface.

B. Understanding Gestures

In the past, two general types of gestural interfaces have been developed. One gestural interface[13] uses stylized gestures of arm and hand configurations; another[14,15] uses the strokes of a stylus on a PDA display to indicate gestures. In our interface we combine both of these approaches, providing both the "natural" gestures of arm movements and hand configurations with the "synthetic" gestures on a PDA display. For the so-called "natural" gestures, a structured-light rangefinder that emits a horizontal plane of laser light is used. This component of the interface detects the positions of the user's hands over several consecutive frames to generate trajectories for gesture interpretation. A camera mounted on the robot just above the laser uses an optical filter tuned to the frequency of the laser. The camera registers the reflection of the laser light off of objects in the room and generates a depth map (XY) based upon location and pixel intensity. Data points for bright pixels (indicating closeness to the robot) are clustered. If a cluster is significantly closer to the robot than background clusters, it is interpreted as a hand. Hand locations are stored from several consecutive frames, and the positions of the hands are used to generate trajectories for gestural commands. For both types of gestures, natural and synthetic, trajectories are analyzed to determine if

they represent valid gestures. The command corresponding to a matched gesture is then queued so that the multimodal interface, upon receiving another command, can retrieve the gesture from the gesture queue and combine it with the verbal command queued in the Command Interpreter.

Regarding "synthetic" gestures, human users can interact with a PDA display: they can point to locations on a map of the environment, draw paths, and encircle areas and objects. We are currently involved in expanding the "synthetic" gestural component by allowing users to point to a computer touch screen (Fig. 2). Users will be able to gesture to objects, areas, and paths in both a real time video display of the robot's view of its environment and a planar map view of the environment based on its sensor data, which is also updated in real time. Depending on the view, a large red dot or a large blue X is used as feedback to indicate where the user has gestured in the display.

Figure 2. Graphical displays of robot's environment.

The left panel shows the real time video display with a red dot indicating where the user has pointed. The right panel shows the planar map view of the robot's environment. The circle in the right display indicates the position of the robot and a radial line inside the circumference indicates the robot's orientation. A large blue "X" (not shown here) is displayed in the planar view when a gesture is made using that view. Objects in the planar view can be labeled either by fiat or by the user.

C. Natural Language Interface

The natural language component of our interface combines a commercial speech recognition front-end with an in-house parsing system. Natural language input can also be typed at a keyboard if one is provided. ViaVoice™ is used to translate the speech signal into a text string, which is then passed to our natural language understanding system, NAUTILUS[16], to produce both syntactic and semantic interpretations. We employ so-called "deep" parsing because we feel that ambiguities inherent in natural language can be more easily handled by parsing robustly. Furthermore, the kinds of detailed representations that we obtain also provide structures that aid in contextual and dialog analysis. Initial semantic interpretations, positional and gestural inputs are compared, matched and resolved (see Fig. 1), and depending upon the appropriateness and necessity of co-occurring gestures with particular utterances, the representations are mapped to an appropriate robot action, such as navigating to a particular location, or an appropriate error message is spoken to aid the user in interpreting what either went wrong or was misinterpreted by the modules discussed thus far.

In terms of the natural language and spatial components of the interface, let us consider the processing of an utterance such as "How many objects do you see?"

ViaVoice™, a commercial off-the-shelf speech recognition system, analyzes the speech signal, mapping the acoustic signal to a text string. NAUTILUS, the in-house natural language understanding system, parses the string syntactically, and maps the grammatical structures to domain predicates with their corresponding arguments, producing a representation something like the following, simplified here for expository purposes:

```
(1)  (ASKWH
        (MANY N3 (:CLASS OBJECT) PLURAL)
           (PRESENT #:V7791
              (:CLASS P-SEE)
                 (:AGENT (PRON N1 (:CLASS SYSTEM) YOU))
                 (:THEME N3)))
```

(1) is a semantic representation obtained from an analysis of the text string. In general, the various verbs or predicates of an utterance (e.g. *see*) are mapped into corresponding semantic classes (*p-see*) that have particular argument structures (*agent, theme*); for example "you" (which is considered a "system" in the particular domain of the example) is the agent (the grammatical "doer" or subject) of the *p-see* class of verbs in this domain and "objects" is the theme (the grammatical object) of this verbal class. In English, these questioned elements are "fronted" in sentences, thereby resulting in the semantic construction *ASKWH* which further binds the *N3* index of the *theme* to the fronted syntactic element "how many objects." N3 is a variable bound to the class OBJECT near the beginning of the representation. Essentially, the form asks how many N3s there are, where N3 is an object seen by "you". If the spoken utterance requires a gesture for disambiguation, as in for example the sentence "Look over there," the gesture components obtain and send the appropriate gesture to the Goal Tracker/Spatial Relations component which combines the linguistic and gesture information.

If an appropriate gesture accompanies the utterance, the various navigational and sensory components of the robot (not shown here but embedded in the Robot Action component of Fig. 1) are activated to yield the appropriate action. If, on the other hand, an inappropriate gesture accompanies the utterance, the robot utters a meaningful response, such as "I'm sorry, I didn't understand that gesture," or "That gesture didn't make any sense with that command." If no gesture is perceived in this example, the robot will simply utter, "Where?" We believe that providing adequate verbal output for the various errors that may occur provides the user with a much more "habitable" interface, thereby creating a feeling that the robot is trying to be cooperative, rather than simply reacting when appropriate or remaining silent if an error is obtained.

Our robots Coyote, Roadrunner, Magneto, and Wolverine are used in command and control applications where humans remotely direct robots around areas. We envision that the various kinds of interfaces developed for different scenarios will be helpful in working with robots in hazardous environments such as a battlefield or in the presence of hazardous materials. Such robots could also be used in remote search and rescue operations, allowing rescue personnel in safe locations to interact with them. Using our video displays, we imagine that humans and robots will be able to cooperate and collaborate in identifying both friendly and enemy personnel, as well as in locating and identifying victims in hazardous environments.

In information retrieval type tasks, such as asking for information about the weather or asking a robot to provide information about a particular conference site or to act as a guide in a particular locale, gestures are not as important; however, the natural language analysis proceeds along similar lines. Our representations provide structures that can be mapped to various query languages to access information in online databases. For certain applications, such as acting as a guide at a convention, domain-specific databases are created and our natural language representations map directly to database query languages. For example, GRACE acted as a robot-receptionist at a recent AAAI conference, while her companion robot, George, acted as a guide, escorting individuals around the San Jose Conference and Convention Center. [3] Conference attendees were able to ask GRACE about times and locations of particular speeches and to obtain information about who was speaking. In addition, GRACE was able to answer questions about the convention center, its various facilities, and about such amenities as local restaurants near the center, having also been provided with an appropriate database in which to look up this information. George, on the other hand, directed individuals to particular locations at the center if they asked GRACE to provide them with an escort. GRACE simply transmitted the request to George, who was ready to perform the task when requested.

We have also ported our natural language modules to both Leonardo and Robonaut, allowing users to interact with these dexterous robots and have them either point at objects or to pick them up. For example, Leonardo can be told to "point to the red button until it turns red," and Robonaut can be employed as an assistant in tool-oriented tasks, and be told to "pick up the wrench over there." The focus of these collaborative efforts is to permit ease of communication, thereby facilitating cooperation and collaboration between humans and robots. Another way in which we seek to promote cooperation and collaboration is by incorporating human cognitive models of behavior. We will turn to this issue below in Sections III.C-D.

As mentioned earlier, we believe that a multimodal interface incorporating natural verbal and gestural channels of communication allows users to concentrate on tasks, rather than on the manipulation of the interface itself. The initial learning curve in using the interface will be reduced, and natural cooperation and collaboration between

humans and robots will be fostered. To test these beliefs, we are currently analyzing data from a human-subject experiment in which individuals were asked to collaborate with a robot in a remote location in order to achieve a particular goal. A pilot study was conducted in which human users were seated in a different location from a robot stationed in another room. Humans were asked to interact with the robot in finding a hidden object, namely a large placard with the word "FOO" written on it. Participants were allowed to direct the robot vocally and point to the real time video and planar map views of its environment, shown in Fig. 2. Some preliminary results have been reported[17].

III. Cognitive Behaviors

A. Understanding Spatial Language

Spatial reasoning is an important aspect of a human-robot interface because humans often think and converse with others in terms of relative spatial positions and relationships. While it may be efficient and computationally effective to store and compute locative information in Cartesian or polar coordinates, humans do not typically think or converse with each other about spatial matters in those terms. Instead, people use natural language expressions that employ a rich structure of locative prepositions, such as *on, near, over, behind, through*, etc., and name other relationships with compounds such as *to the left of, slightly to the right of, almost on top of*, etc. Similar expressions can often be used to construct non-locative expressions, such as "on Tuesday" or "Monday through Friday"[18]. Consequently, our Spatial Reasoning Component allows human collaborators to use these more intuitive expressions to control a robot or a team of robots. For example, we may want to give the robot a command such as "Go down the road fifty feet. Turn right behind the building, and proceed forward for twenty more feet. Then patrol the perimeter." Or, in an office setting, "Go between the table and the chair, through the doorway, and down the hall to the left fifty feet." Allowing human collaborators to imbed spatial terms and locative information in commands such as these allows them to control the robot in a manner that is already familiar and natural to them.

Likewise, the Spatial Relations Component of the interface allows the robot to provide feedback derived from its onboard sensor detection of objects and evidence grid maps to the human operator in a very natural way. Thus, both human and robot exhibit characteristics of members working in a team, conversing and exchanging information in very natural ways, making it easier for the human to participate in the dialog.

Consider the following excerpt of a dialog (2) in which a human user and a robot exchange information about the robot's environment.

(2) Human: "What objects do you see?"
Robot: "I see three objects."
Human: "Where are they?"
Robot: "Object 1 is in front of me. Object 2 is behind me and to my right. Object 3 is to my left."
Human: "Object 1 is a pillar. Object 2 is a computer, and object 3 is a table."
Robot: "I now know that object 1 is a pillar, object 2 is a computer, and object 3 is a table."
Human: "Go between the pillar and the table."
Robot: "Going to that location."

In (2), the robot presents sensor information in natural terms, such as "in front of me" and "behind me and to my right." We maintain that it is easier in certain contexts for humans to understand the information presented in this fashion, rather than if it were presented in numerical or in some other symbolic representation which might require additional interpretation on the part of the human. Also, the human can refer to objects in a very natural way, such as renaming Object 1 as a "pillar." Again, permitting natural human communication makes it easier for the human to interact with the robot. These capabilities foster a more collaborative and cooperative environment for human-robot interactions.

B. Perspective-taking

Perspective-taking, or the ability to see or imagine something from someone else's point of view, is also an important part of interpreting locative information. For example, if the human commands the robot, "Turn left," the robot must understand whose left is being referred to, the human's or the robot's. As has been discussed elsewhere[19], commands typically employ "addressee-centered" perspectives. In order to facilitate and understand commands and directions, humans typically phrase their utterances in ways that are easy for addressees to interpret. Therefore, unless additional information is presented, as for example in the utterance "Turn to my left," commands in our interface usually take the form of addressee-centered utterances. However, phrasing utterances involving

perspective in this way alone is not always sufficient. Frequently, participants in a dialog need to reason about or know how their counterparts view or perceive the environment. For example, in a situation where someone asks a robot to pick up a particular tool that it cannot see, the robot should not first question the tool's existence, but instead should assume that the request is felicitous[20] and reason that the person making the request is referring to a tool that is salient or visible from his or her vantage point but happens to be obscured from its own. Next, it should reason spatially about the immediate physical environment from the perspective of the person making the request. This should then be compared with its own perspective in order to determine where the occluded tool is likely to be located, at which point, the robot should move to carry out the request. In related work, using Polyscheme[21], cooperative perspective-taking was incorporated into a domain-specific application in which a robot participated in both an analogous scenario to the preceding one and another scenario that involved determining which of two identical objects a human collaborator was referring to when one of the objects was obscured from the collaborator's perspective. In both scenarios, Polyscheme was used to resolve the collaborator's ambiguous references, by constructing a model of the world from the point of view of its human counterpart, and then reasoning from this perspective. Thus, in an environment where two orange traffic cones were visible to one of our robots but only one was visible to one of the authors, the robot, when asked to "Go to the cone," was consistently able to choose the cone that was visible to both itself and its collaborator. A more detailed discussion of this research is currently under review[22].

C. Hide and Seek

Humans communicate with each other with a great deal of shared knowledge and tacit understanding of each other's behavior about situations and actions. Cognitive models allow us to capture that knowledge and the associated processes in a cognitively plausible way to support human-robot interaction. Along these lines, we have built a cognitive model of a particular task using ACT-R, a theory for simulating and understanding human cognition[23], modeled after the child's game of Hide and Seek.

Basing our model on the human model of the game[24,25], one of our robots was taught how to play the game. The robot learned about hiding by playing a series of games with a human. After each game, it was also given feedback as to its performance just as a human might be given corrective actions while learning to play the game. For example, it could be told not to hide in the open, or that a chair is too small to hide behind. Using this model, the robot was then asked to play the game with another human. It performed as expected and located the human in an area best suited for a person to hide. Since the object of the game is to remain as hidden as possible from another player, the robot employed this strategy to find the best hiding place in which to hide from its fellow player. For instance, the robot learned, just as humans do, that objects good for hiding tend to be enclosures or obstacles in or behind which they can be "hidden" from view.

We then constructed a reverse model of the game in which the robot was asked to seek a hidden human. Using this reverse model, the robot was able to find the hidden person successfully. The robot was able to reason where the human was hiding based upon the rules it had previously learned. Knowing how and where to hide provided the reverse rules to help the robot seek and find.

D. Other Task Models

When communicating with another individual concerning the completion of a task, humans utilize contextual information, world knowledge, domain-specific knowledge, etc. not only to present the information but also to interpret the actions as they happen and to take corrective measures based upon how they perceive a situation to be unfolding. To exhibit how this information can assist humans and robots to work more cooperatively and collaboratively, our latest cognitive model written for Robonaut addresses the task of tightening a wheel's lug nuts.

In this activity, a human and Robonaut act as co-workers, attempting to tighten four lug nuts. The robot keeps track of which parts of the nut-tightening task have been done and compares them to the model of the task to be accomplished, as taught by the human at an earlier time. During this activity, the human interacts with Robonaut, telling the robot what actions are being performed. For example, in a typical nut-tightening scenario, the human tells Robonaut what step is being completed: "Robonaut, I am tightening nut number one." Robonaut observes the actions, listening to the verbal information, and updates the task model accordingly. The human could also ask Robonaut to perform part of the task by saying, "Tighten nut two."

Suppose, however, that at a certain point in this activity, the human is called away, after all but the last nut on the wheel has been tightened. As he is leaving, the human says, "Robonaut, tighten the last one." To cooperate with the human's request and collaborate on the completion of the task, we supplied Robonaut with a cognitive model specific to this task, allowing it to resolve the situation and correctly interpret task-related language ambiguities; i.e., it decides that "the last nut" must refer to the one nut that has yet to be tightened, eliminating other possibilities,

such as the last nut touched. In this task-completion exercise, the robot acts as a valued co-worker, knowledgeable about the overall task, aware of the current state, and capable of completing the task on its own. Thus, it steps in and completes the task without any extra instruction or confusion over which nut needs tightening.

IV. Conclusion

In order to facilitate cooperative and collaborative behavior in human-robot interactions, we have integrated natural modalities of communication. Thus, speech recognition and natural language understanding are integrated with gesture understanding, thereby permitting human users of the interface to concentrate on a task rather than on the modalities for interacting with one or more of our robots. Incorporating human cognitive models of spatial reasoning, task completion, and perspective-taking further allows us to build robotic systems that are easier to interact with. With the various modalities of the interface which we provide our users, humans feel they are given the opportunity to interact with teammates, rather than feel they are laboriously teleoperating or directing an uncooperative or seemingly uncooperative agent through a task. Incorporation of human cognitive models of task completion and perspective-taking further facilitates these interactions.

Acknowledgments

This research has been funded by the DARPA IPTO MARS Program (MIPR #04-L697), the ONR Intelligent Systems Program (Work Order #N0001404WX20210), and an NRL Research Option (Work Certificate #IT-015-09-4C & 4D). The authors would also like to thank Farilee Mintz for her transcriptions and coding being used in an analysis of the FOO data.

References

[1] Perzanowski, D., Schultz, A., Adams, W., Bugajska, M., Abramson, M., MacMahon, M., Atrash, A., and Coblenz, M., "'Excuse me, where's the registration desk?' Report on Integrating Systems for the Robot Challenge AAAI 2002," *Human-Robot Interaction: Papers from the 2002 AAAI Fall Symposium*, Technical Report FS-02-03, AAAI Press, Menlo Park, CA., Fall 2002, pp. 63-72.

[2]Simmons, B. R., Schultz, A., Kortenkamp, D., Maxwell, B., Goldberg, D., Goode, A., Montemerlo, M., Roy, N., Sellner, B., Urmson, C., Abramson, M., Adams, W., Atrash, A., Bugajska, M. M., Coblenz, M. , MacMahon, M., Perzanowski, D., Wolfe, B., and Milam, T. "GRACE: An Autonomous Robot for AAAI Robot Challenge," *AI Magazine*, Summer 2003, pp. 51-72.

[3]Mink, J., Rosenthal, S., Caffrey, K., Thomas, S., and Perzanowski, D, "Integrating Natural Language Processing for Social Robots," Naval Research Laboratory, (forthcoming).

[4]Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Mulanda, D., "Humanoid Robots as Cooperative Partners for People," *International Journal of Humanoid Robots*, (in review).

[5]Breazeal, C., Hoffman G., and Lockerd, A., "Teaching and Working with Robots as a Collaboration," *Autonomous Agents and Multi-Agent Systems*, (in review).

[6]Sofge, D., Bugajska, M., Trafton, J.G., Perzanowski, D., Thomas, S., Skubic, M., Blisard, S., Cassimatis, N., Brock, D., Adams, W., and Schultz, A., "Collaborating with Humanoid Robots in Space," *International Journal of Humanoid Robotics CIRAS Special Issue on Humanoid Robotics,* World Scientific Publishing, Singapore, 2004, (in review).

[7]Sofge, D., Trafton, G., Cassimatis, N., Perzanowski, D., Bugajska, M., Adams, W., Schultz, A., "Human-Robot Collaboration and Cognition with an Autonomous Mobile Robot," *Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8)*, University of Amsterdam, Amsterdam, March 2004, pp. 80-87.

[8]Sofge, D., Perzanowski, D., Bugajska, M., Adams, W., Schultz, A., "An Agent Driven Human-centric Interface for Autonomous Mobile Robots," In *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI2003)*, Volume III, July 2003, International Institute of Informatics and Systemics, pp. 223-228.

[9]Perzanowski, D., Schultz, A., and W. Adams, W., "Integrating Natural Language and Gesture in a Robotics Domain," *Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference*, Gaithersburg, MD: National Institute of Standards and Technology, September 1998, pp. 247-252.

[10]Perzanowski, D., Schultz, A., Adams, W., Bugajska, M., Marsh, E., Trafton, G., Brock, D., Skubic, M., and Abramson, M. "Communicating with Teams of Cooperative Robots," *Multi-Robot Systems: From Swarms to Intelligent Automata,* edited by A.C. Schultz and L.E. Parker, Proceedings from the 2002 NRL Workshop on Multi-Robot Systems, Kluwer, The Netherlands, 2002, pp. 185-193.

[11]Schultz, A., Adams, W., and Yamauchi, B., "Integrating Exploration, Localization, Navigation and Planning Through a Common Representation," *Autonomous Robots*, Vol. 6, No. 3, 1999, pp. 293-308.

[12]Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., and Bugajska, M., "Spatial Language for Human-Robot Dialogs," *IEEE Transactions on Systems, Man, and Cybernetics: Part C: Applications and Reviews*, Vol. 34, No. 2, May 2004, pp. 154-167.

[13]Kortenkamp, D., Huber E., and Bonasso., P, "Recognizing and Interpreting Gestures on a Mobile Robot," *Proceedings of Thirteenth Nation Conference on Artificial Intelligence,* AAAI Press, Menlo Park, CA, August 1996, pp. 915-921.

[14]Fong, T., Thorpe, C., and Baur, C., "Advanced Interfaces for Vehicle Teleoperation: Collaborative Control, Sensor Fusion Displays, and Remote Driving Tools," Autonomous Robots, Vol. 11, 2001, pp. 77-85.

[15]Fong, T., and Thorpe, C., "Robot as Partner: Vehicle Teleoperation with Collaborative Control," *Multi-Robot Systems: From Swarms to Intelligent Automata,* edited by A.C. Schultz and L.E. Parker, Proceedings from the 2002 NRL Workshop on Multi-Robot Systems, Kluwer, The Netherlands, 2002, pp. 195-202.

[16]Wauchope, K. "Eucalyptus: Integrating Natural Language Input with a Graphical User Interface," Technical Report NRL/FR/5510-94-9711, Naval Research Laboratory, Washington, D.C, 1994.

[17]Perzanowski, D., Brock, D., Blisard, S., Adams, W., Bugajska, M., Schultz, A., Trafton, G., Skubic, M., "Finding the FOO: A Pilot Study for a Multimodal Interface," *Proceedings of the IEEE Systems, Man, and Cybernetics Conference,* IEEE, Piscataway, NJ, October 2003, pp. 3218-3223.

[18] Jackendoff, R., *Semantics and Cognition*, MIT Press: Cambridge, MA, 1983.

[19]Taylor, H.A., and Tversky, B., "Perspective in spatial descriptions," *Journal of Memory and Language,* Vol. 35, 1996, pp. 371-391.

[20]Grice, H.P., "Logic and Conversation," *Syntax and Semantics: Speech Acts,* Vol. 3, edited by P. Cole and J. Morgan, Academic Press: New York, 1975, pp. 43-58.

[21]Cassimatis, N.L., "Polyscheme: A Cognitive Architecture for Integrating Multiple Representation and Inference Schemes," Ph.D. Dissertation, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[22]Trafton, J. G., Cassimatis, N. L., Brock, D. P., Bugajska, M., Mintz, F., and Schultz, A. C., "Enabling effective human-robot interaction using perspective-taking in robots" *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, (in review).

[23]Anderson, J.R., and C. Lebiere, C., *The Atomic Components of Thought*, Lawrence Erlbaum, Mahwah, NJ, 1998.

[24]Trafton, J.G., Cassimatis, N., Hiatt, L., Schultz, A.C., Perzanowski, D., Bugajska, M.D., Adams, W., Brock, D.P., "Using similar representations to improve human-robot interaction" Naval Research Laboratory, (forthcoming).

[25]Trafton, J.G., Schultz, A.C., Perzanowski, D., Bugajska, M.D., Adams, W., Cassimatis, N.L., and Brock, D.P., "Children and robots learning to play hide and seek," Naval Research Laboratory, (forthcoming).