Achieving Collaborative Interaction with a Humanoid Robot

D. Sofge[1], D. Perzanowski[1], M. Skubic[3], N. Cassimatis[2], J. G. Trafton[2],
D. Brock[2], M. Bugajska[1], W. Adams[1], A. Schultz[1]
[1,2]Navy Center for Applied Research in Artificial Intelligence,
Naval Research Laboratory
Washington, DC 20375
[1]{sofge, dennisp, magda, adams, schultz}@aic.nrl.navy.mil
[2]{cassimatis, trafton, brock}@itd.nrl.navy.mil

[3]Department of Computer Engineering and Computer Science
University of Missouri-Columbia
Columbia, MO 65211
skubicm@missouri.edu

Abstract

One of the great challenges of putting humanoid robots into space is developing cognitive capabilities for the robots with an interface that allows human astronauts to collaborate with the robots as naturally and efficiently as they would with other astronauts. In this joint effort with NASA and the entire Robonaut team we are integrating natural language and gesture understanding, spatial reasoning incorporating such features as human-robot perspective taking, and cognitive model-based understanding to achieve this high level of human-robot interaction.

1. Introduction

As we develop and deploy advanced humanoid robots such as Robonaut to perform tasks in space in collaboration with human astronauts, we must consider carefully the needs and expectations of the human astronauts in interfacing and working with these humanoid robots, and to endow the robots with the necessary capabilities for assisting the human astronauts in as useful and efficient a manner as possible. By building greater autonomy into the humanoid robot, the human burden for controlling the robot will be diminished and the humanoid will become a much more useful collaborator with a human astronaut for achieving mission objectives in space.

In this effort we build upon our experience in designing multimodal human-centric interfaces and cognitive models for dynamically autonomous mobile robots. We argue that by building human-like capabilities into Robonaut's cognitive processes, we can achieve a very high level of interactivity and collaboration between human astronauts and Robonaut. Some of the necessary components for this cognitive functionality addressed in this paper include use of cognitive architectures for humanoid robots, natural language and gesture understanding, and spatial reasoning with human-robot perspective-taking.

2. Cognitive Architectures for Humanoids

Most of Robonaut's activities involve interaction with human beings. We base our work on the premise that embodied cognition, using cognitive models of human performance to augment a robot's reasoning capabilities, facilitates human-robot interaction in two ways. First, the more a robot behaves like a human being, the easier it will be for humans to predict and understand its behavior and interact with it. Second, if humans and robots share at least some of their representational structure, communication between the two will be much easier. For example, both in language use [1] and other cognition [2], humans use qualitative spatial relationships such as "up" and "north". It would be difficult for a robot using real number matrices to represent spatial relationships and transformations without also endowing it with qualitative representations of space. In [3] and [4] we used cognitive models of human performance of the task to augment the capabilities of robotic systems.

We have decided to use two cognitive architectures based on human cognition for certain high-level control mechanisms for Robonaut. These cognitive architectures are ACT-R [5] and Polyscheme [6].

ACT-R is one of the most prominent cognitive architectures to have emerged in the past two decades as a result of the information processing revolution in the

cognitive sciences. Also called a unified theory of cognition, ACT-R is a relatively complete theory about the structure of human cognition that strives to account for the full range of cognitive behavior with a single, coherent set of mechanisms. Its chief computational claims are: first, that cognition functions at two levels, one symbolic and the other subsymbolic; second, that symbolic memory has two components, one procedural and the other declarative; and third, that the subsymbolic performance of memory is an evolutionarily optimized response to the statistical structure of the environment. These theoretical claims are implemented as a production-system modeling environment. The theory has been successfully used to account for human performance data in a wide variety of domains including memory for goals [7], human computer interaction [8], and scientific discovery [9]. We will use ACT-R to create cognitively plausible models of appropriate tasks for Robonaut to perform.

Second, we will use Cassimatis' Polyscheme [6] architecture for spatial, temporal and physical reasoning. The Polyscheme cognitive architecture enables multiple representations and algorithms (including ACT-R models), encapsulated in "specialists", to be integrated into inference about a situation. We will use an updated version of the Polyscheme implementation of a physical reasoner to help keep track of Robonaut's physical environment.

2.1. Perspective-taking

One feature of human cognition that is very important for facilitating human-robot interaction is "perspective-taking". There is extensive evidence that human perspective-taking is an important cognitive ability even for young children. In order to understand utterances such as "the wrench on my left", the robot must be able to reason from the perspective of the speaker what "my left" means. We will use the Polyscheme cognitive architecture, integrated with an ACT-R model, to endow Robonaut with the ability to conceive of task-oriented goals and knowledge of another person. This will allow Robonaut to more easily predict and explain its behavior, making it a better partner in a collaborative activity.

Polyscheme has a simulation mechanism, called a "world", which we will use to endow Robonaut with perspective-taking capabilities. Polyscheme will allow Robonaut to use multiple representations to reason from the perspective of what it sees in its immediate environment. Using worlds, Polyscheme can simulate the perspective it would have at other times, different places and in hypothetical worlds and use its specialists to make inferences within those perspectives. Polyscheme uses worlds to implement algorithms such as counterfactual reasoning, backtracking search, truth-maintenance and stochastic simulation. We will use and extend the world mechanism to reason about the perspective of *other people*. This will enable Polyscheme to predict and explain other people's behavior by using its perceptual, motor, procedural, memory, spatial and physical specialists from the perspective of another person's mind.

3. Multimodal Interface

We use a multimodal interface to process the various interactions with the robot. While there are a wide variety and many examples of multimodal interfaces, too numerous to site here, there are a few multimodal interfaces that focus on the kinds of interactions with which we are concerned; namely, gestural and natural language modes of interaction. For example, one gestural interface uses stylized gestures of arm and hand configurations [10] while another is limited to the use of gestural strokes on a PDA display [11]. Other interactive systems, such as [12,13], process information about the dialog using natural language input. Our multimodal robot interface is unique in its combination of gestures and robust natural language understanding coupled with the capability of generating and understanding linguistic terms using spatial relations.

4. Understanding Language and Gestures

Any interface which is to support collaboration between humans and robots must include a natural language component. We currently employ a natural language interface that combines a ViaVoice speech recognition front-end with an in-house developed deep parsing system [14]. This gives the robot the capability to parse utterances, providing both syntactic representations and semantic interpretations. The semantic interpretation subsystem is integrated with other sensor and command inputs through use of a command interpretation system. The semantic interpretation, interpreted gestures from the vision system, and command inputs from the computer or other interfaces are compared, matched and resolved in the command interpretation system.

Using our multimodal interface (Figure 1), the human user can interact with a robot, using natural language and gestures. The natural language component of the interface embodied in the Spoken Commands and Command Interpreter modules of the interface uses ViaVoice to analyze spoken utterances. The speech signal is translated to a text string that is further analyzed by our natural language understanding system, Nautilus, to produce a regularized expression. This representation is linked, where necessary, to gesture information via the

Gesture Interpreter, Goal Tracker/Spatial Relations component, and Appropriateness/Need Filter, and an appropriate robot action or response results.

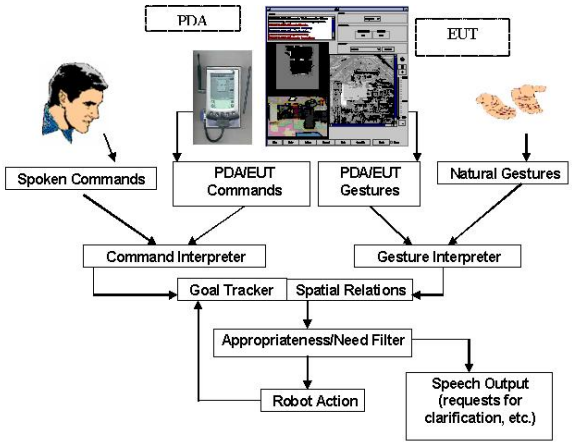

Figure 1 Multimodal Interface for Robot Collaboration.

For example, the human user can ask the robot "How many objects do you see?" ViaVoice analyzes the speech signal, producing a text string. Nautilus parses the string and produces a representation something like the following, simplified here for expository purposes.

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

The parsed text string is mapped into a kind of semantic representation, shown here, in which the various verbs or predicates of the utterance (e.g. *see*) are mapped into corresponding semantic classes (*p-see*) that have particular argument structures (*agent*, *theme*); for example "you" is the agent of the *p-see* class of verbs in this domain and "objects" is the theme of this verbal class, represented as "N3"—a kind of co-indexed trace element in the theme slot of the predicate, since this element is fronted in English wh-questions. If the spoken utterance requires a gesture for disambiguation (e.g. the sentence "Look over there"), the gesture components obtain and send the appropriate information to the Goal Tracker/Spatial Relations component where linguistic and gesture information are combined.

Both natural and so-called "symbolic" gestures are input to the multimodal interface. Users can gesture naturally by indicating directions, measurements, or specific locations with arm movements or they can use more symbolic gestures, by indicating paths and locations on a metric-map representation of the environment or video image on a PDA screen or end-user terminal (EUT). Users of this modality can point to locations and objects directly on the EUT monitor, thereby permitting the following kinds of utterances: "Go this way," "Pick up that object/wrench," or "Explore the area over there" using a real-time video display. If the gesture — whatever its source — is valid, a message is sent to the appropriate robotics module(s) to generate the corresponding robot action. If the gesture is inappropriate, an error message is generated to inform the user. Where no gesture is required or is superfluous, the linguistic information maps directly to an appropriate robot command. In the example above (1), no further gesture information is required to understand the question about the number of objects seen.

Thus far we have been interacting with several non-humanoid mobile robots. As we move in the direction of working with humanoid robots, we believe natural gestures will become more prevalent in the kinds of interactions we study. Gesturing is a natural part of human-human communication. It disambiguates and provides information when no other means of communication is used. For example, we have already discussed the disambiguating nature of a gesture accompanying the utterance "Look over there." However, humans also gesture quite naturally and frequently as a non-verbal means of communicating information. Thus, a human worker collaborating with another worker in an assembly task might look in the direction of a needed tool and point at it. The co-worker will typically interpret this look and gesture as a combined non-verbal token indicating that the tool focused on and gestured at is needed, should be picked up and passed back to the first co-worker. In terms of the entire communicative act, both the look and the gesture indicate that a specific object is indicated, and the context of the interaction, namely assembly work, dictates that the object is somehow relevant to the current task and should therefore be obtained and handed over.

A verbal utterance might also accompany the foregoing non-verbal acts, such as "Get me that wrench" or simply "Hand me that." In the case of the first utterance, the object in the world has a location and a name. Its location is indicated by the deictic gestures perceived (head movement, eye gaze, finger pointing, etc.), but its name comes solely from the linguistic utterance. Whether or not the term "wrench" is already known by the second co-worker, the latter can locate the object and complete the task of handing it to the first co-worker. Further, even if the name of the object is not part of the second co-worker's lexicon, it can be inferred from the gestural context. Gestures have narrowed down the

possibilities of what item in the world is known as a "wrench." In the case of the second utterance above, the name of the item is not uttered, but the item can still be retrieved and handed to the first co-worker. In this case, if the name of the item is unknown, the second co-worker can ask "What's this called?" as the co-worker passes the requested item.

We envision such interactions and behaviors as those outlined above as elements of possible scenarios between humans and Robonaut. Thus far, in our work on a multimodal interface to mobile robots, we have shown how various modes of our interface can be used to facilitate communication and collaboration. However, we would like to extend such capabilities to a humanoid robot, as well as add learning, such as learning the name of an object previously unknown based on contextual (conversational and visual) information.

5. Spatial Reasoning

Building upon the existing framework of natural language understanding with semantic interpretation, and utilizing the on-board sensors for detecting objects, we are developing a spatial reasoning capability on the robot [15,16,17,18]. This spatial reasoning capability will be fully integrated with the natural language and gesture understanding modules through the use of a spatial modeling component based on the histogram of forces [19]. Force histograms are computed from a boundary representation of two objects (extracted from sensory data) to provide a qualitative model of the spatial relationship between the objects. Features extracted from the histograms are fed into a system of rules [20] or used as parameters in algorithms [17] to produce linguistic spatial terms. The spatial language component will be incorporated into the cognitive framework of the robot through a perspective-taking capability implemented using the Polyscheme architecture.

5.1. Spatial Language

Spatial reasoning is important not only for solving complex navigation tasks, but also because we as human operators often think in terms of the relative spatial positions of objects, and we use such relational linguistic terminology naturally in communicating with our human colleagues. For example, a speaker might say, "Hand me the wrench on the table." If the assistant cannot find the wrench, the speaker might say, "The wrench is to the left of the toolbox." The assistant need not be given precise coordinates for the wrench but can look in the area specified using the spatial relational terms.

In a similar manner, this type of spatial language can be helpful for intuitive communication with a robot in many situations. Relative spatial terminology can be used to limit a search space by focusing attention in a specified region, as in "Look to the left of the toolbox and find the wrench." It can be used to issue robot commands, such as "Pick up the wrench on the table." A sequential combination of such directives can be used to describe and issue a high level task, such as, "Find the toolbox on the table behind you. The wrench is on the table to the left of the toolbox. Pick it up and bring it back to me." Finally, spatial language can also be used by the robot to describe its environment, thereby providing a natural linguistic description of the environment, such as, "There is a wrench on the table to the left of the toolbox."

In all of these cases the spatial language increases the dynamic autonomy of the system by giving the human operator a less restrictive vernacular for communicating with the robot. However, the examples above also assume some level of object recognition by the robot. Although there has been considerable research on the linguistics of spatial language for humans, there has been only limited work done in using spatial language for interacting with robots. Some researchers have proposed a framework for such an interface [21]. Moratz et al. [22] investigated the spatial references used by human users to control a mobile robot. An interesting finding is that the test subjects consistently used the robot's perspective when issuing directives, in spite of the 180-degree rotation. At first, this may seem inconsistent with human to human communication. However, in human to human experiments, Tversky et al. observed a similar result and found that speakers took the listener's perspective in tasks where the listener had a significantly higher cognitive load than the speaker [23].

To address the object recognition problem, we use the spatial relational language to assist in recognizing and labeling objects, through the use of a dialog. Once an object is labeled, the user can then issue additional commands using the spatial terms and referencing the named object. An example is shown below:

Human: "How many objects do you see?"
Robot: "I see 4 objects."
Human: "Where are they located?"
Robot: "There are two objects in front of me, one object on my right, and one object behind me."
Human: "The nearest object in front of you is a toolbox. Place the wrench to the left of the toolbox."

Establishing a common frame is necessary so that it is clear what is meant by spatial references generated both by the human operator as well as by the robot. Thus, if the human commands the robot, "Turn left," the robot must know whether the operator refers to the robot's left or the operator's left. In a human-robot dialog, if the robot places a second object "just to the left of the first object," is this the robot's or the human's left?

Currently, commands using spatial references (e.g., go to the right of the table) assume an extrinsic reference frame of the object (table) and are based on the robot's viewing perspective to be consistent with Grabowski's "outside perspective" [24]. That is, the spatial reference assumes the robot is facing the referent object.

There is some rationale for using the robot's viewing perspective. In human-robot experiments, Moratz et al. found that test subjects consistently used the robot's perspective when issuing commands [22]. We are currently investigating this through use of human-factors experiments where individuals who do not know the spatial reasoning capabilities and limitations of the robot provide instructions to the robot for performing various tasks where spatial referencing is required. The results of this study will be used to enhance the multimodal interface by establishing a common language for spatial referencing which incorporates those constructs and utterances most frequently used by untrained operators for commanding the robot.

5.2. Spatial Representation

In our previous work, we have used both 2D horizontal planes (e.g., an evidence grid map, built with range sensor data) and 2D vertical planes (using image data), but thus far they have not been combined. For Robonaut, we will combine them to create a 2½D representation. To achieve the type of interaction described above, it is not necessary to build a full 3D representation of the environment. Rather, we assert that a more useful strategy is to obtain range information for a set of objects. Human spatial language naturally separates the vertical and horizontal planes, e.g., the wrench is on the table, vs. the wrench is to the left of the toolbox. Our linguistic combination utilizes both prepositional clauses, e.g., the wrench is on the table to the left of the toolbox. Processing the spatial information as two (roughly) orthogonal planes provides a better match with human spatial language.

Range information is extracted from stereo vision; the vision-based object recognition can assist in determining the correct correspondence between stereo images by constraining the region in the image. We do not need to label everything in the scene, but only those objects or landmarks that provide a basis to accomplish the robot's task. The position of recognized objects can be stored in a robot-centric frame such as the Sensory Ego Sphere [25]; global position information is not necessary.

6. Conclusion

Humanoid robots such as Robonaut offer many opportunities for advancing the use of robots in complex environments such as space, and for development of more effective interfaces for humans to interact with robots. Once a sufficiently high level of interaction between robots and humans is achieved, the operation of and interaction with these robots will become less of an additional burden for the humans, and more of a collaboration to achieve the objectives of the task-at-hand. In this paper we describe our plans to endow Robonaut with cognitive capabilities which will support collaboration between human astronauts and Robonaut. We build upon our experience in natural language understanding, gesture recognition, spatial reasoning and cognitive modeling in achieving these goals.

Acknowledgements

Support for this effort was provided by the DARPA IPTO Mobile Autonomous Robot Software (DARPA MARS) Program. Thanks also to Sam Blisard for his contributions to this effort.

References

[1] G. A. Miller and P. H. Johnson-Laird. *Language and Perception*. Harvard University Press. 1976.

[2] B. Tversky, "Cognitive maps, cognitive collages, and spatial mental model," In A. U. Frank and I. Campari (Eds.), *Spatial information theory: Theoretical basis for GIS*, Springer-Verlag, 1993.

[3] M. Bugajska, A. Schultz, T. J. Trafton, M. Taylor, and F. Mintz, "A Hybrid Cognitive-Reactive Multi-Agent Controller," In *Proceedings of 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, EPFL, Switzerland, 2002.

[4] J. G. Trafton., A. Schultz, D. Perzanowski, W. Adams, M. Bugajska, N. L. Cassimatis, and D. Brock, "Children and robots learning to play hide and seek," In *Proceedings of the IJCAI Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico; August 2003.

[5] J. R. Anderson and C. Lebiere. *The atomic components of thought.* Lawrence Erlbaum, 1998.

[6] N. L. Cassimatis., Polyscheme: A cognitive architecture for integrating multiple representation and inference schemes. PhD dissertation. MIT Media Laboratory, 2002.

[7] E. M. Altmann and J. G. Trafton, "An activation-based model of memory for goals." In *Cognitive Science*, 39-83, 2002.

[8] J. R. Anderson, M. Matessa, and C. Lebiere. "ACT-R: A theory of higher level cognition and its relation to visual attention," In *Human-Computer Interaction*, 12 (4), 439-462), ASME Press, 763-768, 1997.

[9] C. D. Schunn and J. R. Anderson, "Scientific discovery," In J. R. Anderson, and C. Lebiere (Eds.), *Atomic Components of Thought*. Lawrence Erlbaum, 1998.

[10] D. Kortenkamp, E. Huber and P. Bonasso, "Recognizing and Interpreting Gestures on a Mobile Robot," In *Proceedings of AAAI*, 1996.

[11] T. W. Fong, F. Conti, S. Grange and C. Baur, "Novel Interfaces for Remote Driving: Gesture, haptic, and PDA," In SPIE 4195-33, *SPIE Telemanipulator and Telepresence Technologies VII*, Nov. 2000.

[12] C. Rich, C. L. Sidner, and N. Lesh, "COLLAGEN: Applying collaborative discourse theory to human-computer interaction," In *AI Magazine*, vol. 22, no. 4, pp. 15-25, 2001.

[13] J. F. Allen, D. K. Byron, M, Dzikovska, G. Ferguson, L. Galescu and A. Stent, "Toward conversational human-computer interaction," In *AI Magazine*, vol. 22, no. 4, pp. 27-37, 2001.

[14] D. Perzanowski, A. Schultz, W. Adams and E. Marsh, "Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy," In *Proceedings of the 1999 IEEE Intl. Symposium on Computational Intelligence in Robotics and Automation*, pp. 208-213, 1999.

[15] M. Skubic, D. Perzanowski, A. Schultz and W. Adams, "Using Spatial Language in a Human-Robot Dialog," In *Proceedings of the IEEE 2002 International Conference on Robotics and Automation*, pp. 4143-4148, 2002.

[16] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska and D. Brock, "Spatial Language for Human-Robot Dialogs," In *IEEE Transactions on SMC*, Part C, Special Issue on Human-Robot Interaction, 2003.

[17] M. Skubic and S. Blisard, "Go to the Right of the Pillar: Modeling Unoccupied Regions for Robot Directives," In *2002 AAAI Fall Symposium, Human-Robot Interaction Workshop*. AAAI Technical Report FS-02-03, 2002.

[18] M. Skubic, P. Matsakis, G. Chronis and J. Keller, "Generating Multi-Level Linguistic Spatial Descriptions from Range Sensor Readings Using the Histogram of Forces," In *Autonomous Robots*, vol. 14, no. 1, pp. 51-69, Jan. 2003.

[19] P. Matsakis and L. Wendling, "A New Way to Represent the Relative Position of Areal Objects," In *IEEE Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 634-643, 1999.

[20] P. Matsakis, J. Keller, L. Wendling, J. Marjamaa, and O. Sjahputera, "Linguistic Description of Relative Positions of Objects in Images," In *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 31, No. 4, pp. 573-588, 2001.

[21] R. Muller, T. Rofer, A. Landkenau, A. Musto, K. Stein, and A. Eisenkolb, "Coarse Qualitative Descriptions in Robot Navigation," In *Spatial Cognition II. Lecture Notes in Artificial Intelligence* 1849, C. Freksa, W. Braner, C. Habel and K. Wender (Eds.) Springer-Verlag, 2000, pp. 265-276.

[22] R. Moratz, K. Fischer and T. Tenbrink, "Cognitive Modeling of Spatial Reference for Human-Robot Interaction," In *Intl. Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 589-611, 2001.

[23] B. Tversky, P. Lee and S. Mainwaring, "Why Do Speakers Mix Perspective?" In *Spatial Cognition and Computation*, vol. 1, pp. 399-412, 1999.

[24] J. Grabowski, "A Uniform Anthropomorphological Approach to the Human Conception of Dimensional Relations," In *Spatial Cognition and Computation*, vol. 1, pp. 349-363, 1999.

[25] R. A. Peters II, K. Hambuchen, K. Kawamura, and D. M. Wilkes, "The Sensory Ego-Sphere as a Short-Term Memory for Humanoids," In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2001.