In *Proceedings of the16th IFAC Symposium on Automatic Control in Aerospace*, Elsevier Science Ltd, Oxford, UK, 2004

COGNITIVE TOOLS FOR HUMANOID ROBOTS IN SPACE

Donald Sofge[1], Dennis Perzanowski[1], Marjorie Skubic[2], Magdalena Bugajska[1],
J. Gregory Trafton[1], Nicholas Cassimatis[1], Derek Brock[1],
William Adams[1], Alan Schultz[1]

[1]Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375

[2]Electrical and Computer Engineering Department
University of Missouri-Columbia
Columbia, MO 65211

Abstract: The effective use of humanoid robots in space will depend upon the efficacy of interaction between humans and robots. The key to achieving this interaction is to provide the robot with sufficient skills for natural communication with humans so that humans can interact with the robot almost as though it were another human. This requires that a number of basic capabilities be incorporated into the robot, including voice recognition, natural language, and cognitive tools on-board the robot to facilitate interaction between humans and robots through use of common representations and shared humanlike behaviors.

Keywords: Cognitive Systems, Co-operative Control, Speech Recognition, Natural Language, Human-Machine Interface, Autonomous Mobile Robots.

1. INTRODUCTION

Humanoid robots are now being built to assist humans in a variety of activities. The Robonaut platform (Ambrose, *et al.*, 2000) is a humanoid robot designed to assist human astronauts working in space (Figure 1). Demonstrated teleoperative capabilities of Robonaut include dexterous grasping, tool handling, and cooperation with human teammates on manual tasks that require both information exchange and physical interaction between robot and human. Current development efforts are focused upon the implementation of autonomous behaviors and capabilities for Robonaut.

Effective collaboration between robots and humans requires the use of an efficient interface whereby a human can communicate and interact with a robot almost as easily as with another human. In this interaction the human may act as a supervisor and/or collaborator with the robot. Human-robot collaboration is facilitated by a number of capabilities built into the interface and the robot itself, including voice recognition, natural language and gesture understanding, and behaviors supporting dynamic autonomy (Sofge, *et al.*, 2003). The inclusion of cognitively plausible representations and processes incorporated into the robot provides a further basis for facilitating collaboration between humans and robots, thereby reducing the human effort required to adapt to limitations of the robot as a non-human collaborator. Use of a cognitive model aboard the robot facilitates better communication and interaction between the human and the robot through use of a common representational framework for the environment and objects within it, processing of sensor information, and joint problem solving involving both humans and robots.

In this paper we focus on the use of cognitive models of spatial reasoning capabilities aboard Robonaut to enhance communications between human astronauts and Robonaut, thereby enabling collaborative teams consisting of human astronauts and humanoid robots.

2. COGNITIVE HUMANOID ROBOTS

Achieving effective collaboration between humans and robots will require the use of cognitive models on-board the robots. *Embodied cognition*, as we call it, uses cognitive models of human performance to augment a robot's reasoning capabilities, and facilitates human-robot interaction in two ways. First, we hypothesize that 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 the representational structures in their interactive communications or activities, we further hypothesize that communication among them will be much easier. For example, in tasks requiring direction generation, humans naturally use qualitative spatial relationships (Miller and Johnson-Laird, 1976; Tversky, 1993) such as "left," "up", "east," or "north." Interaction with a robot capable of manipulating the same representation instead of traditional real number matrices would be more natural and efficient. In (Bugajska, *et al*., 2002) and (Trafton, *et al*., 2003) we used cognitive models of human performance of the task to augment the capabilities of robotic systems.

We are investigating the use of two cognitive architectures based on human cognition for certain high-level control mechanisms aboard Robonaut. These cognitive architectures are ACT-R/S (Harrison and Schunn, 2003) based upon the ACT-R architecture (Anderson and Lebiere, 1998) and Polyscheme (Cassimatis, 2002).

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 sub-symbolic; second, that symbolic memory has two components, one procedural and the other declarative; and third, that the sub-symbolic performance of memory is optimized in 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 (Altmann and Trafton, 2002), human computer interaction (Anderson, *et al*., 1997), and scientific discovery (Schunn and Anderson, 1998).

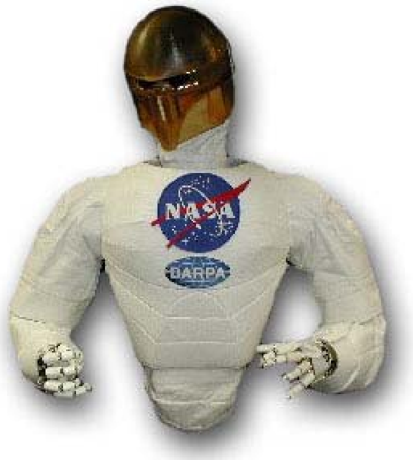

Fig. 1. Robonaut – NASA's Humanoid Robot

The ACT-R/S system uses three different spatial represent-tations suggested in (Harrison and Schunn, 2003). Briefly, they suggest that people use a focal representation for object identification that consists of non-metric geons, a manipulative representation for grasping and tracking that consists of metric geons, and a configural representation for navigation that consists of rectangular bounding regions. While a coarse representation is adequate for obstacle avoidance in navigation, in order to do perspective taking, we must spatially transform an object, focusing on the configural representation to determine that object's spatial references. These transformations must then be mapped onto the user's representations so that actions can be performed. We use ACT-R/S to create cognitively plausible models of human performance of tasks to be performed by the robots.

Furthermore, we are using Cassimatis' Polyscheme architecture (Cassimatis, 2002) 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 inferences that agents make about a situation or about possible situations. Finally, we use an updated version of the Polyscheme implementation of a physical reasoner to help keep track of the robot's physical environment.

2.1 Perspective-Taking

One of the features of human cognition that facilitates natural human-robot interaction is "perspective-taking". 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 to resolve the meaning of "my left". We will use the Polyscheme architecture and ACT-R models to endow the robot with the ability to conceive of task-oriented goals and knowledge of another person. This will allow the robot to more easily predict and explain its own behavior, as well as the behavior of others, making it a better partner in a collaborative activity.

Polyscheme has a simulation mechanism, called a "world" in which the robot is endowed with perspective-taking capabilities. Polyscheme allows the robot to reason about what it sees in its immediate environment from different perspectives. Using worlds, Polyscheme can simulate the perspective it would have at other times, different places and in other hypothetical worlds, and use its specialists to make inferences within those perspectives. Polyscheme uses reasoning algorithms such as counterfactual reasoning, backtracking search, truth-maintenance and stochastic simulation. We have created a specialist to reason about the perspective(s) of *other people*. This allows Polyscheme to predict and explain other people's behavior, using its perceptual, motor, procedural, memory, spatial and physical specialists from the perspective of another person.

For both ACT-R/S and Polyscheme we have created preliminary models that can perform simple spatial perspective-taking tasks. There seem, however, to be advantages and disadvantages to both systems: For example, ACT-R/S has more difficulty doing large scale simulations, but has a large amount of historical cognitive plausibility (e.g., there have been a large number of empirical and psychological studies validating ACT-R/S), while Polyscheme has comparatively less of a cognitive history. Additionally, because the representations and operations of each system are a bit different, their behaviors are different and various tasks may be easier or more straightforward to model for one system than for another.

3. NATURAL LANGUAGE UNDERSTANDING

While ACT-R/S and Polyscheme enable us to embody a common cognitive model for tasks and spatial reasoning, we employ a natural language and gesture understanding system by which a human and robot can communicate this information to each other. Our rationale here is that the use of as natural an interface as possible again facilitates interaction. Our natural language interface combines a commercial speech recognition front-end with an in-house developed deep parsing system, NAUTILUS (Wauchope, 1994). ViaVoice™ is used to translate the speech signal into text, which is then passed to our natural language understanding system, Nautilus, to produce both syntactic and semantic interpretations. 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 2) the human user can interact with the robot using both natural language and gestures. The semantic interpretation 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.

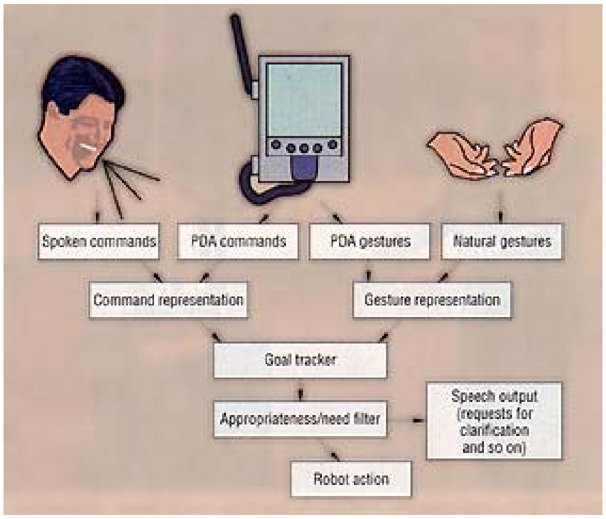

Fig. 2. NRL Multimodal Interface

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

The parsed text string is mapped into a kind of semantic representation (1). 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" 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 syntactically fronted in English wh-questions. 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 component which combines linguistic and gesture information.

4. SPATIAL LANGUAGE

As human operators we 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 and vice versa. However, the examples above also assume some level of object recognition by the robot.

To address the object recognition problem, we use natural language to address spatial relations, thereby assisting humans interacting with robots in recognition and labeling of objects (Skubic, *et al,.* 2002). Given our natural language interface, a human can easily communicate with the robot about objects and spatial relations in the environment through the use of a dialog that is easy and natural for the human. Furthermore, once an object is labeled, the user can then issue additional commands using natural spatial terms and by referencing the named object. An example is given in (2):

Human: "How many objects do you see?" (2)
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 is referring to the robot's left or the operator's left. Likewise, in a human-robot dialog, if the robot places a second object "just to the left of the first object," both need to know if the goal location is to the left of the robot or the human.

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" (Grabowski, 1999). That is, the spatial reference assumes the robot is facing the referent object.

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 (Muller, *et al*., 2000). Moratz (2001) 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 (1999) 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.

The experiments by Moratz (2001) provide rationale for using the robot's viewing perspective. We are currently investigating this further 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 (Perzanowski, *et al*., 2003). 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. CONCLUSIONS

Humanoid robots are being designed and built to provide assistance to humans in complex and challenging work environments, such as outer space. Achieving effective use of these humanoid robots in space will depend upon the difficulty of the tasks required of human astronauts interacting with robots. The use of cognitive tools aboard the robots provides a number of benefits, such as shared representations, behaviors, and modes of interaction between humans and robots, thereby easing the cognitive load on the part of the human. The key to achieving effective interaction is to provide the robot with sufficient skills for natural communication with humans so that humans can interact with the robot almost as easily as with another human

This paper describes the design, implementation, and capabilities of a robotic system architecture for a robot which can be used (at some level) to collaborate with a human. The capabilities required of the robot include voice recognition, natural language understanding, gesture recognition, spatial reasoning, and cognitive modeling with perspective-taking. These represent a small subset of potential capabilities humans utilize with one another in collaborating to perform a task in a complex environment, and barely scratches the surface of capabilities we might want to build into an intelligent, collaborative robot.

The capabilities described above have been successfully implemented and demonstrated on several mobile robotic platforms (Sofge, *et al.*, 2004), and we are now porting them to Robonaut. We are also extending the capabilities of the cognitive architectures (both ACT-R/S and Polyscheme) and their perspective-taking cognitive models. Future work will focus on enhancing the cognitive models through expanded rulesets and cognitively plausible behaviors and reasoning mechanisms, and adding learning capabilities to the models so that the robots may be able to acquire new knowledge and skills through interaction with humans and while performing tasks. Parts of this architecture have already been extended to several robots designed specifically for enhanced human interaction, such as MIT's robot Leonardo (Breazeal, 2003) (Figure 3). While Leonardo is not a humanoid, it is being developed with human-like characteristics and functionalities. We are also extending the architecture and methodology to include and study collaboration between teams of robots and humans.

ACKNOWLEDGEMENTS

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

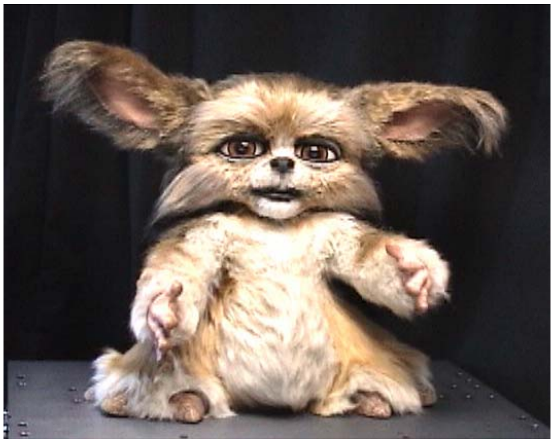

Fig. 3. MIT's Leonardo Robot (photo courtesy Cynthia Breazeal, © MIT Media Lab, 2002)

REFERENCES

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

Ambrose, R. O., H. Aldridge, R. S. Askew, R. R. Burridge,W. Bluethmann, M. Diftler, C. Lovchik, D. Magruder, F. Rehnmark (2000). "Robonaut: NASA's space humanoid," IEEE Intelligent Systems, *IEEE Intelligent Systems*, vol. 15, no. 4 , pp. 57-63.

Anderson, J. R. and C. Lebiere (1998). *The atomic components of thought.* Lawrence Erlbaum.

Anderson, J. R., M. Matessa, and C. Lebiere (1997). "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.

Breazeal, C. (2003). "*Towards sociable robots*," Robotics and Autonomous Systems, vol. 42, no. 3-4.

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

Cassimatis., N. L. (2002). "Polyscheme: A cognitive architecture for integrating multiple representation and inference schemes," PhD dissertation, MIT Media Laboratory.

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

Harrison, A., and Schunn, C. D. (2003). "Segmented spaces: Coordinated perception of space in ACT-R." In F. Detje, D. Dorner & H. Schaub (Eds.), *The logic of cognitive systems: Proceedings of the Fifth International Conerence. on Cognitive Modeling,* pp. 307, Bamberg, Germany.

Miller, G. A., and P. H. Johnson-Laird (1976). *Language and Perception*. Harvard University Press.

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

Muller, R., T. Rofer, A. Landkenau, A. Musto, K. Stein, and A. Eisenkolb (2000). "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, pp. 265-276.

Perzanowski, D., A. Schultz, W. Adams, and E. Marsh, (1999). "Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy," In *Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Monterey, CA.

Perzanowski, D., D. Brock, S. Blisard, W. Adams, M. Bugajska, A. Schultz, G. Trafton, M. Skubic, (2003), "Finding the FOO: A Pilot Study for a Multimodal Interface," In *Proceedings of the IEEE Systems, Man, and Cybernetics Conference*, Washington, DC.

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

Skubic, M., D. Perzanowski, A. Schultz, and W. Adams (2002). "Using Spatial Language in a Human-Robot Dialog," In *Proceedings 2002 IEEE Conference on Robotics and Automation*, IEEE.

Sofge, D., D. Perzanowski, M. Skubic, N. Cassimatis, J. G. Trafton, D. Brock, M. Bugajska., W. Adams, and A. Schultz (2003). "Achieving Collaborative Interaction with a Humanoid Robot," In *Proceedings of the Second International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS)*.

Sofge, D., J. G. Trafton, N. Cassimatis, D. Perzanowski, M. Bugajska, W. Adams, and A. Schultz (2004). "Human-Robot Collaboration and Cognition with an Autonomous Mobile Robot," In *Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8)*.

Trafton., J. G., A. Schultz, D. Perzanowski, W. Adams, M. Bugajska, N. L. Cassimatis, and D. Brock (2003). "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.

Tversky, B. (1993). "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.

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

Wauchope, K. (1994). "*Eucalyptus: Integrating Natural Language Input with a Graphical User Interface,"* Naval Research Laboratory Technical Report NRL/FR/5510-94-9711, Washington, DC.