Human-Robot Collaboration and Cognition with an Autonomous Mobile Robot

Donald SOFGE, J. Gregory TRAFTON, Nicholas CASSIMATIS,
Dennis PERZANOWSKI, Magdalena BUGAJSKA,
William ADAMS, Alan SCHULTZ
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375

Abstract. 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 efficiently as he/she would 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 robot and robot interface, including voice recognition, natural language and gesture understanding, and behaviors supporting dynamic autonomy. The inclusion of cognitively plausible representations and processes aboard 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.

Introduction

Effective collaboration between robots and humans in accomplishing complex tasks requires the use of an efficient interface whereby a human can communicate and interact with a robot almost as efficiently as he/she would with another human. This level of interaction requires a number of capabilities not often found in deployed robotic systems today. These include voice recognition with integrated natural language understanding, recognition of human gestures (such as pointing to objects), and built-in behaviors for sequencing and executing tasks requiring various levels of control by — and interaction with — a human supervisor (we refer to this as dynamically adjustable autonomy, or dynamic autonomy). Use of cognitive models aboard the robots may further enhance the human-robot interaction through use of a common set of representations, process steps and process times for processing sensory data, and expectations shared by both human and robot.

1. Dynamic Autonomy

Dynamic autonomy allows the robot to dynamically adjust its behaviors depending upon and appropriate to the task(s) at hand [1]. The human operator is able to interact with the robot in a human-centric manner by providing verbal commands and gestures to the robot to perform tasks requiring varying levels of human interaction. Some circumstances may require very fine-grained level operator control, while others may require less precision. Dynamic autonomy used in mobile robots provides a more flexible and operator-friendly interface and makes the robots more versatile.

We support dynamic autonomy in our system through a number of robot behaviors of varying complexity including collision-free navigation, path following, exploration, automatic prioritization of multiple command directives, and feedback from the robot to the operator. Feedback is provided by voice synthesis and through text strings requesting clarification if the robot isn't able to understand the command(s). Operation of the natural language interface with the gesture interpretation process and other command input modes is discussed in greater detail in the section describing the robot's integrated goal-driven architecture. Figure 1 shows our human-centric multimodal interface for autonomous mobile robots.

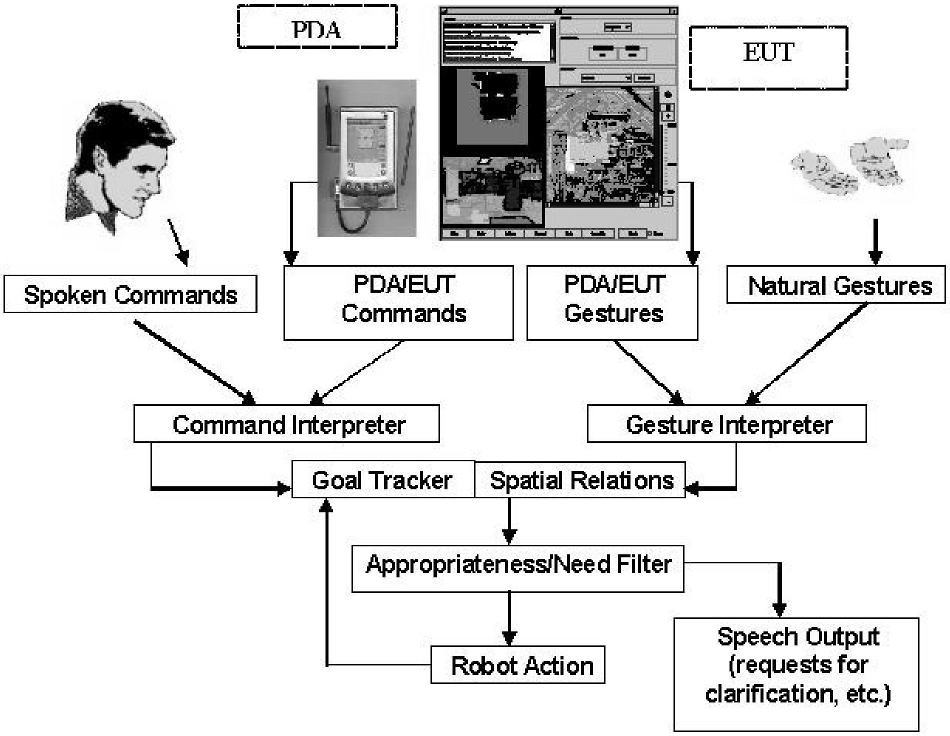

Figure 1. Human-Centric Multimodal Interface

Using this interface, commands may be communicated to the robot in a variety of human-centric ways including verbally, through touch (on a PDA, touch tablet or keyboard), and through gesturing with hands and arms. The robot passes a variety of information back to the operator such as sensor readings, video, navigation maps built using its array of on-board sensors [2], and the status of commands sent to it.

2. Understanding Gestures

One of the key modes of interaction with the robot is through the use of gestures. Several types of gestural interfaces have been developed and used in the past. For example, one gestural interface uses stylized gestures of arm and hand configurations ("natural" gestures) [3], while another is limited to the use of gestural strokes on a PDA display ("synthetic" gestures) [4]. In our system we combine both of these approaches, allowing both "natural" and "synthetic" gestures. Our stylized gesture interface utilizes a structured-light rangefinder to detect the positions of the hands over several consecutive frames to generate trajectories for the gesture command. The structured-light rangefinder emits a horizontal plane of laser light. A camera mounted on the robot just above the laser is fitted with an optical filter which is 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. The data points for bright pixels (indicating closeness to the robot) are clustered. If a cluster is significantly closer to the robot than background scenery, it is interpreted as being a hand. Hand locations are stored from several consecutive frames, and the positions of the hands are used to generate trajectories for the gesture command. Each trajectory is analyzed to determine if it represents a valid gesture. The command corresponding to the 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 in the command interpretation system.

3. Natural Language Interface

Our natural language interface combines a commercial speech recognition front-end with an in-house developed deep parsing system [1]. 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 (Figure 1).

Using our multimodal interface 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. 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 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 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/Spatial Relations component which combines linguistic and gesture information.

The effectiveness of the natural language and gesture interfaces are further enhanced through use of a spatial reasoning component which allows humans and robots to communicate using spatial terms. The spatial reasoning component is described next.

4. Spatial Reasoning

Spatial reasoning is an important element of a human-centric interface because humans often think in terms of relative spatial positions, and use such relational linguistic terminology naturally in communicating with one another. Our spatial reasoning component builds upon an existing framework of natural language understanding with semantic interpretation [5], and utilizes on-board sensors for detecting objects and map-building through use of evidence grids.

Understanding spatial linguistic terms allows for more efficient and natural control of a dynamically autonomous mobile robot. For example, we may want to give the robot a command such as "Go down the road 100 feet, turn right behind the building and proceed ahead 20 feet. Then go into active surveillance mode." Or, in an office setting, "Go between the table and the chair, through the doorway, and down the hall to the left 50 feet." Spatial reasoning increases the dynamic autonomy of the system by giving the operator a less restrictive vernacular for commanding the robot.

The spatial reasoning component of the multimodal interface allows the robot to provide feedback to the human operator using natural spatial terminology. The human is able to query the robot about the relative spatial positions of objects in the environment, and the robot is able to respond using spatial terms. This is demonstrated in the following dialogue.

> Human: "Tell me what you see."
> Robot: "I see 3 objects."
> Human: "Where are they located?"
> Robot: "Object A is 5 feet in front of me. Object B is 10 feet in front of me and to my right. Object C is 20 feet to my left."

This natural spatial language is used to disambiguate spatial references by both humans and robots [5]. It provides a common interpretation for location expressions, such as "left" and "right", as well as other relative directions. 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. Use of spatial language between humans and robots is currently under investigation by our group at NRL through human-factors experiments [6] where novice users provide instructions to the robot for performing various tasks where spatial referencing is required. This work will result in development of a common language for spatial referencing geared to the needs and expectations of untrained and non-expert operators. This common spatial language will be incorporated into the multimodal interface.

5. Cognitive Robots

Spatial reasoning is only one aspect of incorporating greater cognitive capabilities into the robots. Achieving effective collaboration between humans and robots will require the use of cognitive models on-board the robots. 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 [7] and other cognition [8], 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 [9] and [10] we used cognitive models of human performance of the task to augment the capabilities of software agents.

In this effort we used two cognitive architectures based on human cognition for certain high-level control mechanisms in our robots. These cognitive architectures are ACT-R [11] and Polyscheme [12].

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 [13], human computer interaction [14], and scientific discovery [15]. In our system we use ACT-R to create cognitively plausible models of appropriate tasks for the robots to perform.

Second, we used Cassimatis' Polyscheme architecture [12] 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 used an updated version of the Polyscheme implementation of a physical reasoner to help keep track of the robot's physical environment.

5.1 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. Our Polyscheme system uses a combination of AI techniques (called specialists) including reactive systems, neural networks, constraint graphs, rule-based systems, and category hierarchies. These specialists are able to simulate other times, places, perspectives and possible worlds. For example, using Polyscheme's mental simulation capabilities to perform perspective taking, if someone says "Give me the wrench on my left," Polyscheme creates a "possible world" for the speaker, identifies what wrench is on its left, and "gives" it to the speaker.

For both ACT-R and Polyscheme we have created preliminary models that can perform simple spatial perspective taking tasks. There seem to be advantages and disadvantages to both systems: ACT-R 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), while Polyscheme has comparatively less 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.

6. Mobile Robot Integrated Goal-Driven Architecture

The natural language, gesture, spatial reasoning, and perspective-taking modules described previously must be integrated into a single coherent system in order to operate on the mobile robots. Figure 2 shows our mobile robot integrated goal-driven architecture. This architecture is organized around providing integration and arbitration for goals presented though various interface modules. Outputs for speech recognition, natural language understanding, gesture interpretation, and other interface modules are cached; command prioritization and resolution are then performed.

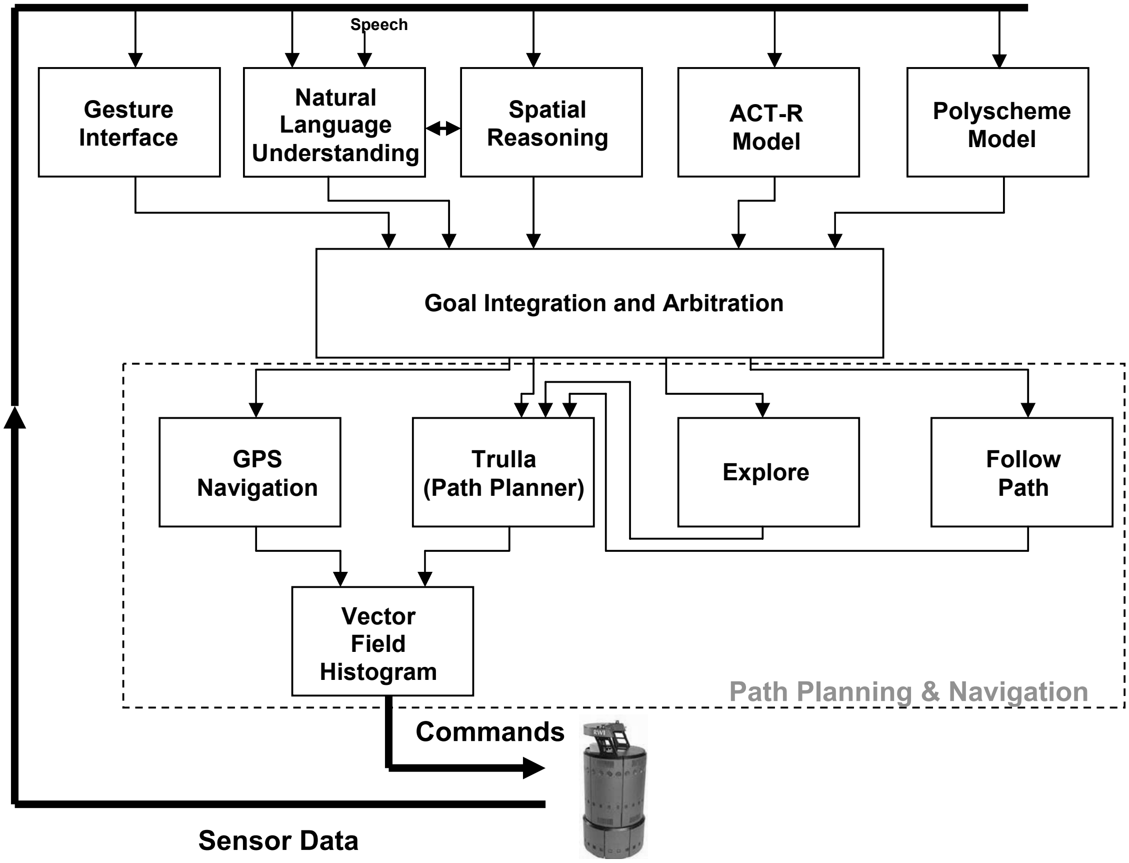

Figure 2. Mobile Robot Integrated Architecture

Once goals are interpreted and resolved, they are passed to the *Path Planning and Navigation* routines, where they are integrated with low-level behaviors such as obstacle avoidance, exploration and path planning using the Vector Field Histogram (VFH) method [16]. The architecture maintains both short-term and long-term maps (not shown in Figure 2), which are also important for several of the other processes such as *Spatial Reasoning*, *PDA Interface*, and *Robot GUI*.

7. Conclusions

This paper describes a robotic system architecture for a robot which can be used 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. We are currently performing human-subject experiments on cooperation in task performance — both with and without robots — in order to gain a better understanding of which capabilities are most critical aboard the robot.

Most of the capabilities described above have been successfully implemented and demonstrated on several robotic platforms. We have most recently added cognitive models (using both ACT-R and Polyscheme) to provide perspective-taking capabilities. Future work will focus on enhancing the cognitive models through expanded rulesets and cognitively plausible (in human terms) behaviors and reasoning mechanisms, and by adding learning capabilities to the models. Parts of this architecture are also being extended to several robots designed specifically for enhanced human interaction, namely NASA's humanoid robot Robonaut [17] and MIT's clearly non-humanoid robot Leonardo [18]. We are also extending the architecture and methodology to include and study collaboration between teams of robots and humans.

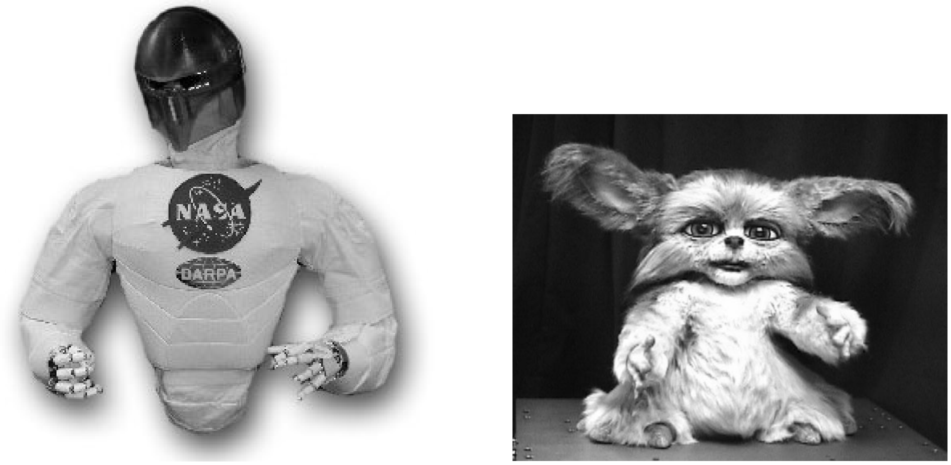

Figure 3. Technology transitions being undertaken for this effort include NASA's Robonaut and MIT's Leonardo
(Leonardo photo courtesy Cynthia Breazeal, © MIT Media Lab, 2002)

Acknowledgements

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

References

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

[2] Schultz, A., W. Adams, and B. Yamauchi (1999). “Integrating Exploration, Localization, Navigation and Planning Through a Common Representation,” In *Autonomous Robots*, 6(3), Kluwer.

[3] Kortenkamp, D., E. Huber, and P. Bonasso, P. (1996). “Recognizing and Interpreting Gestures on a Mobile Robot,” In *Proceedings of AAAI.*

[4] Fong, T. W., Conti, F., Grange, S. and Baur, C. (2000), “Novel Interfaces for Remote Driving: Gesture, haptic, and PDA,” *SPIE 4195-33, SPIE Telemanipulator and Telepresence Technologies VII*, Boston, MA.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

[14] 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.

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

[16] Borenstein, J. and Y. Koren (1991). “The Vector Field Histogram – Fast Obstacle Avoidance for Mobile Robots,” *IEEE Transactions on Robots and Automation***,** IEEE: New York.

[17] 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.

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