Finding the FOO

A Pilot Study for a Multimodal Interface*

Dennis Perzanowski, Derek Brock,
William Adams, Magdalena Bugajska,
Alan C. Schultz, J. Gregory Trafton
Naval Research Laboratory
Washington, DC, U.S.A.
<dennisp | adams | magda | schultz>@aic.nrl.navy.mil
<brock | trafton>@itd.nrl.navy.mil

Sam Blisard, Marjorie Skubic
University of Missouri-Columbia
Columbia, MO, U.S.A.
snbfg8@mizzou.edu
skubicm@missouri.edu

Abstract *– In our research on intuitive means for humans and intelligent, mobile robots to collaborate, we use a multimodal interface that supports speech and gestural inputs. As a preliminary step to evaluate our approach and to identify practical areas for future work, we conducted a Wizard-of-Oz pilot study with five participants who each collaborated with a robot on a search task in a separate room. The goal was to find a sign in the robot's environment with the word "FOO" printed on it. Using a subset of our multimodal interface, participants were told to direct the collaboration. As their subordinate, the robot would understand their utterances and gestures, and recognize objects and structures in the search space. Participants conversed with the robot through a wireless microphone and headphone and, for gestural input, used a touch screen displaying alternative views of the robot's environment to indicate locations and objects.*

Keywords: Dynamic autonomy, gesture, human-robot interaction, intelligent communication, intelligent systems, mixed-initiative, multimodal interface, natural language, Wizard-of-Oz.

1 Introduction

As robots are used more and more to work with humans, issues involving human-robot interaction become increasingly important. Terms such as mixed-initiative, cooperation, collaboration, and dynamic autonomy have become significant concepts for roboticists. Likewise, robotics engineers concerned with facilitating human-robot interaction must address these same issues when implementing their user interfaces. The more they succeed in addressing these issues and solving related problems, the more they will ensure that the modalities of the interface provide easy and natural means to obtain users' goals.

In our research in human-robot interfacing [3,4], we set out to develop intuitive ways to interact with intelligent, mobile robots. Our underlying assumption was that a user interface that provides such modes of interaction as natural language and gestures, which people normally use with each other and are readily familiar with, greatly reduces the learning curve in human-robot interactions. In other words, if people are able to transfer communicative skills they already possess to their interactions with robots, there will be greater ease of use and the need to learn new strategies can be minimized.

In our applications, we anticipate that humans and robots will need to be able to interact in both basic and nonbasic settings [1]. Basic settings are proximal settings in which the participants of an activity can communicate with each other in a face-to-face manner. Such settings are basic because they only require people to employ perceptual and language skills that are widely taken to be fundamental; namely, hearing, seeing, speaking, and gesturing. Basic settings fully support the use of these skills. In contrast, nonbasic settings require people to compensate with specialized, secondary procedures and skills to coordinate activities they pursue together. For humans and robots, a characteristic nonbasic setting is one in which the parties are sufficiently removed from each other to prevent the full use of their face-to-face interaction skills. Thus, for example, arm, head, and eye gestures, while appropriate in a basic setting, may not be appropriate in a nonbasic setting simply because the participants in the activity cannot see each other.

Even though the communication abilities of today's robots, including our own, remain far from those of their human counterparts, characterizing human-robot interaction in terms of basic and nonbasic communicative settings is an established, well-defined framework in which to couch the structure and goals of our multimodal interface research. Our human-robot interface, therefore, incorporates a range of interaction modes to accommodate

* 0-7803-7952-7/03/$17.00 © 2003 IEEE.

both fundamental and specialized communication skills that users can employ. In particular, speech and natural gestures can be used in basic, or face-to-face, settings. In nonbasic settings in which the parties are distant from each other, commands and deictic gestures can be made using a graphical user interface and pointing device on a personal digital assistant (PDA) or some other form of end-user terminal (EUT), such as a touch screen (Figure 1).

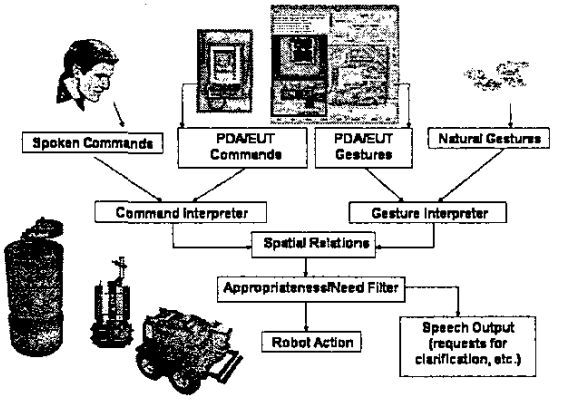

Figure 1. System diagram of multimodal interface

In the preliminary study we describe in this paper, the task our participants and our mobile robot carried out took place in a nonbasic setting. Consequently, we did not employ all of the interaction modes shown in Figure 1. In particular, the part of the Gesture Interpreter for recognizing and interpreting natural gestures, such as arm movements and pointing motions in face-to-face settings, was not utilized. Instead, we have focused on verbal communication and gestural input through a touch screen. While a significant focus of our previous interface work has been on facilitating the use of physically demonstrative gestures, in the present study, our intent is to explore how people choose to communicate with a robot in a remote location (in this case, in another room). Thus, our participants were allowed to freely converse with the robot, but their demonstrative gestures were limited to touching the screen and making explicit deictic references to objects and locations on the shared views of the robot's environment.

For this exercise, then, we used our multimodal interface as modified above. Some information about objects and locations in the robot's environment was derived by the Spatial Reasoning Component [5] of our interface, since we were very much interested in seeing how people referred to objects when giving directions and trying to maneuver a mobile robot around objects. The interface already permits rather complex commands and interactions involving spatial relations and objects, such as "Go behind the pillar" [6]. Therefore, we wanted to see how much of the spatial information that was already available in the interface the participants would use and how they would use it. Our findings in this regard are presented below.

2 The pilot study

Our decision to use a nonbasic setting for our multimodal interface study is motivated by scenarios that envision the use of robots as proxies in real-world tasks that currently expose people to unacceptable levels of risk. If semi-autonomous robots are to carry out dangerous reconnaissance operations, for instance, or to work in hazardous environments, their activities will need to be supervised and directed remotely. We naturally expect people's remote coordination strategies with each other to be different from those they would use in face-to-face settings. Therefore, one of our primary goals is to develop a principled understanding of how the affordances of our current interface necessarily affect people's language, gestures, and expectations. We are particularly interested in interactions with a highly advanced robot in a remote operational context.

As a preliminary exercise to carrying out a more formal study, we conducted a pilot study. For this study, we asked five volunteers, naïve to our purposes and our robot's interface and its abilities. Their individual task was to carry out, with as little instruction and coaching as possible, a remotely guided search for an object. We wished to observe how they would use the various modalities of the interface.

After some demographic information was collected, participants were told they could talk freely with an intelligent, mobile robot (Figure 2) named Coyote. Coyote was located in another room via a wireless microphone and headset they were given to wear. They were also told that Coyote's natural language understanding was good—about as good as a native speaker of English—and that Coyote was familiar with its environment in the sense that it knew what things were, such as boxes, tables, a pillar, etc.

Figure 2. Coyote, the mobile robot

Participants were seated in front of a touch-sensitive computer screen that showed two views of the robot's

environment (Figure 3). They were told that the view on the left was a real-time video camera display of the robot's view of its environment. The view on the right showed a floor plan or map of the environment that was updated in real-time by the robot. Participants were told that they could point to objects and locations on either display as they wished.

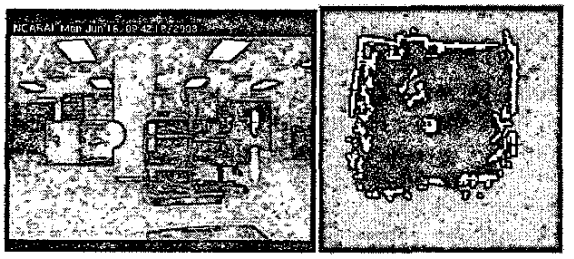

Figure 3. Touch screen display of robot's eye view (left) and mapped representation (right) of environment. (The large (red) dot on the pillar in the left display indicates where participants touched the screen.)

Finally, participants were told that their job was to get the robot to find a hidden object in its environment, namely, a large yellow card with the word "FOO" written on it. They were shown a card identical to the real one and were assured that the card could be read easily on the computer display when and if it was found. This was essentially all the instruction participants received. They were asked if they had any questions about the task or about how to interact with the robot. Once we were confident that the participants understood everything, the robot verbally announced that it was "Good to go" and initiated the exercise, which continued until the sign marked FOO was located. All of the exercises were successfully completed. Participants were then asked to complete an exit questionnaire and were also verbally interviewed.

We chose a Wizard-of-Oz approach rather than using our interface off the shelf, simply because we already knew what its limitations were and wanted our participants to believe that the robot was indeed capable of a wide variety of actions and understanding. We believed that if participants suspected the robot was in any way limited or lacked some skill, we would not elicit intuitive interactions. Therefore, instead of employing a dynamically autonomous robot that could process natural language, translate speech to meaningful navigational commands, interpret locations and objects, interact with a person, and navigate around a complex environment, the robot was controlled by two research scientists, or "wizards," who have worked on the interface and on the robot's navigational and understanding systems for over five years. One of the wizards acted as the navigational controller of the robot, controlling its motions with a joystick according to the directions and gestures that were made by the participant in the other room. Both wizards were able to see all of the participants' gestures on a monitor mounted on top of the robot, mirroring each of the participant's actions on the touch screen. Whenever the participants touched either view, a large red dot appeared in both the participants' and the wizards' displays (Figure 3).

Both wizards also wore wireless headsets so they could hear the participants' utterances. However, the second wizard acted as the spoken language understanding and generation system for the robot. Using a wireless microphone and a sound modulator, this wizard responded to each participant's commands and queries as needed, either to correct errors, explain some difficulty encountered in navigation, or inform the participant that the robot understood what it was expected to do. We felt that such feedback was necessary in any natural communicative act. The wizard's voice was modulated so that it did not sound particularly human. Rather, it sounded more like a synthesized voice and succeeded in convincing participants that they were indeed speaking and interacting with a mechanical device.

3 The results

All five participants (two females and three males) were adults over 21 years of age. The participants' interaction protocols were transcribed. From these transcriptions, counts of utterances, gestures, use of the two kinds of visual displays (the video camera view and the floor plan view), and certain other aspects of each participants' task performance were made (Table 1). We found notable differences in how the participants used the interface and approached the problem of communicating with the robot to accomplish the task. Participants made an average of 57 utterances and 11.2 gestures. Utterance counts ranged from a low of 35 to a high of 94. Counts of gestures ranged from 0 to 19. Participants making the most utterances tended to make the fewest gestures and vice versa. Most gestures, but not all, were combined with utterances. When isolated gestures were made, they were apparently intended to be corrections or refinements of previous gestures. As expected, nearly all of the participants' utterances were directives. The most notable exceptions were questions, whose counts are also shown in Table 1. Participants averaged 14.6 definite references to objects and/or locations in the robot's environment. Many of these were presented as combined utterances and gestures, as in "Move close to this chair" or "Move here." The references *this chair* and *here* were qualified by touch screen interactions. Counts of adjacency pairs, which are simply conversational exchanges in which a participant's utterance is matched by the robot's response to indicate construal, indicate the amount of discourse that took place in each exercise. For instance, a speaker's command to the robot "Coyote, stop" was followed by the robot's response "Stopping."

Table 1. Various measures of participants' task performance

Participants:	P1	P2	P3	P4	P5	means
utterances	35	53	42	94	61	57
gestures	10	18	19	0	9	11.2
combined utterances and gestures	10	15	15	0	9	9.8
adjacency pairs	35	50	33	91	57	53.2
gestures on video camera view	10	18	15	0	2	9
gestures on floor plan view	0	0	4	0	7	2.2
definite references	10	4	17	33	9	14.6
questions	1	0	0	14	1	3.2

Finally, the averages of gestures made on the video camera view versus the floor plan view (respectively, 9 and 2.2) show a pronounced preference for the former, with only one subject choosing to use the floor plan view more often.

The participants' protocols also revealed a range of qualitative differences in how they approached their interactions with the robot. All of the participants except one quickly became comfortable talking to the robot, but each chose to speak to it in a different way. One formally prefaced nearly every utterance with the robot's name and tried to tightly manage the robot's actions throughout the exercise. The others appeared to be less concerned with formality in their addresses, but differed in their degree of control. Two of the participants, for instance, asked the robot on several occasions to turn an unusually precise number of degrees left or right, as in "Turn left twenty-five degrees." But these participants also chose to use more commonly heard and imprecise adverbs in many of their other turning directives, such as "Turn slightly right." In contrast, two of the others used no directional modifiers at all, and yet another, used almost none.

Participants also differed in their use of objects in the search environment as landmarks for navigation. Three directed the robot without naming any of the objects in the room. Instead, objects were identified gesturally. For the most part, objects were associated with deictic words, such as *here* and *there*, and were used as indirect means for indicating locations the robot was to move to. Another participant also used objects in this way but chose, in nearly every instance, to both point to them on the touch screen and name them, as in "Move to this bookcase." The remaining participant also made use of objects in giving directions, but always named them verbally and never identified them gesturally. In fact, this participant made no use of the touch screen at all. Unlike any of the others, in several instances, this individual specifically used objects as navigational aids in giving directions, such as "Go between the pillar on your left and the bookcase on your right." While this participant's instructions were consistently the most specific of any of the participants, they were also the most demanding of the robot's spatial reasoning abilities and autonomy.

From the exit questionnaire and interview, the following opinions were gathered from the participants. All of the participants seemed to enjoy working with the robot to carry out the task. When asked, all rated the interaction experience positively and gave no indication that they doubted they were truly interacting with a machine. All of the participants had more than a passing knowledge of computers or used computers in their daily work. None, however, were involved in robotics research.

Participants also differed concerning the kind of visual display they preferred to use. Whereas one participant felt that both displays were extremely helpful, another participant felt that the map view of the task was completely unnecessary. Yet another interacted almost entirely with the map view and felt the camera view was only needed to identify the sign with the word FOO printed on it.

One participant claimed to have been reluctant to speak freely with the robot due to recent experiences with a commercial operating system that permits a simpler and more constrained set of verbal interactions with the computer's desktop. In hindsight, this participant regretted not having been less cautious once it became apparent that the robot's speech interface was much more robust than the one available in the commercial system. This participant's reticence and mistrust of computational systems was also apparent during the exercise. Despite the preliminary instructions, this participant's session had to be halted briefly in order to go back over the purpose and use of the touch screen and some of the robot's navigational abilities.

As mentioned above, we were interested in seeing how the spatial information that was available in our interface would be used by the participants. When the interface is

operating fully and using sensor information [3,5,6], it is capable of providing easy to understand spatial descriptions for the user, such as "The pillar is to my right and mostly in front of me." Thus, the user can get a good linguistic description of what the robot is sensing about its world. Furthermore, it can understand linguistically appropriate queries and commands involving landmarks and spatial relations between objects. Users of the interface, for example, can tell the mobile robot to navigate around objects in straightforward terms, such as "Coyote, go between the carton and the table." In addition, users can provide deictic information through gestures when appropriate. We, therefore, wanted to see if our participants would spontaneously talk about and refer to objects in the robot's world in this way.

On the basis of the protocols, we would characterize most of the utterances as terse directives that generally lack any verbal identification of landmarks. A common directive, for example, was "Move here." Such utterances were almost always accompanied by a deictic gesture on the touch screen that identified the location of *here*. On the other hand, two of the participants did supply landmark information verbally, such as "Move to this column," and "Coyote, stop before the chair." However, only one of these latter participants accompanied these utterances with corroborating deictic gestures.

We are not sure how to interpret this finding in this pilot study. Perhaps the participants didn't really believe the robot was all that linguistically competent and capable of understanding more complex commands involving landmarks and spatial relations involving objects. On the other hand, this finding may be due to some other aspect of the setting or the task which we have yet to identify.

This preliminary finding is similar to the research results of Moratz, et al. [2]. In this study, about half of the participants directed the robot using landmarks. However, the other half decomposed the control actions into simpler path segments, such as "Drive a bit forward" and "Come ahead to the right." The authors hypothesized that the participants may have assumed that the robot did not have the capability of understanding references to landmarks.

4 Conclusions

In this Wizard-of-Oz pilot study, participants directed a mobile robot in another room to find a sign with the word FOO written on it. They interacted with the robot both verbally and gesturally. The participants did not know that the robot's collaborative behavior was being orchestrated by two of the researchers involved in the study.

While the participants enjoyed their interactions with the robot, and believed they were interacting with a rather intelligent robot, they appeared to have some reservations. On the basis of the participants' interaction protocols, it is apparent to the authors that the participants generally felt that they had to guide the robot constantly. In turn, this strategy possibly affected the types of utterances they tended to use in their interactions with the robot. If, for example, every single step in a task has to be enumerated or specifically stated, then utterances may become shortened. This suggests that a more dynamically autonomous robot could foster more complex and linguistically intricate interactions.

The task basically involved finding a sign with a word printed on it. Because of this, our participants may have thought of the robot as only a mobile camera and/or a sensory device to manipulate into position for viewing a goal. Thus, it was also their job to interpret whatever was observed. Characteristic utterances, such as "Turn left," "Move here," and "Look here," may have been the result of the participants thinking that they were dealing with a servant robot, rather than with a collaborative agent in solving the task. By making the task more complex, the participants may be forced to rely more on the mobile robot's interpretation of actions and visual cues, rather than fall back on their own interpretations of those events and simply command the robot to move about to get a better view of something.

In terms of the kind of task that the participants were asked to perform, we feel we did elicit spontaneous interactions. However, in the future, we intend to manipulate certain factors, such as the frame rate of the video camera view. Also, specifically to test the Spatial Relations Component of the interface, we will test and compare the various modalities of the interface individually. For example, as we outlined elsewhere [5], one group of participants will perform a particular task using natural language, while another group will use a joystick for the same task. We can then calculate success or failure of task completion on several parameters, such as total time to perform the task, time spent searching for the object, time spent navigating for the object, and user satisfaction with the various modalities. With results from these kinds of manipulations, we can hypothesize about the appropriateness or intuitiveness of a particular modality.

The chief purpose of this pilot study was to mature and validate the Wizard-of-Oz techniques that are planned for our forthcoming formal study on human-robot interaction. The result of this pilot study was a number of design improvements made and lessons learned. Critical hardware and software problems were identified. Some were successfully addressed, but others will require further consideration. Iterative evaluation of the whole

process between participants allowed us to catch numerous omissions, and to refine the consistency of the Wizards' interaction behavior and the overall conduct of the experiment. Careful consideration of the preparation and exit processes also allowed us to improve the participant training phase and to develop a more thorough exit questionnaire, respectively. All of this work will ensure the integrity of the future formal study.

Acknowledgments

The pilot study was funded by DARPA. The authors would also like to thank Scott Thomas for his assistance in the administration of the study.

References

[1] H. H. Clarke, *Using language*, Cambridge University Press: New York, NY, 1996.

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

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

[4] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a Multimodal Human-Robot Interface," *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 16-21, January/February 2001.

[5] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial Language for Human-Robot Dialogs," *IEEE Transactions on Systems, Man and Cybernetics*, (to appear).

[6] M. Skubic and S. Blisard, "Go to the Right of the Pillar: Modeling Unoccupied Spaces for Robot Directives," Technical Report, *AAAI 2002 Fall Symposium, Human-Robot Interaction Workshop*, November, 2002.