Goal Tracking and Goal Attainment: A Natural Language Means of Achieving Adjustable Autonomy

Dennis Perzanowski, Alan C. Schultz, Elaine Marsh, and William Adams

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Codes 5512 and 5514
Washington, DC 20375-5337
< dennisp | schultz | marsh | adams > @aic.nrl.navy.mil

Abstract

Intelligent mobile robots that interact with humans must be able to exhibit adjustable autonomy, that is the ability to dynamically adjust the level of autonomy of an agent depending on the situation. When intelligent robots require *close* interactions with humans, they will require modes of communication that enhance the ability for humans to communicate naturally and that allow greater interaction. Our previous work examined the use of multiple modes of communication, specifically natural language and gestures, to disambiguate the communication between a human and a robot. In this paper, we propose using *context predicates* to keep track of various goals during human-robot interactions. These context predicates allow the robot to maintain multiple goals, each with possibly different levels of required autonomy. They permit direct human interruption of the robot, while allowing the robot to smoothly return to a high level of autonomy.

Introduction

We have been involved in tasks that require tight human and robot interactions. The combined human/robot system requires that goals and motivations can originate either from the human or from the robot. It may be necessary for either of these agents (the human or the robot) to assume the responsibility of instantiating goals which direct the combined system towards completion of its task. We refer to systems with this property as *mixed-initiative systems*, i.e. the initiative to dictate the current objective of the system can come from the robot itself or from a human.

In the context of mixed-initiative systems, *adjustable autonomy* is a critical requirement. Adjustable autonomy allows systems to operate with dynamically varying levels of independence, intelligence, and control. In these systems, a human user, another system, or the autonomous system itself may adjust the system's "level of autonomy" as required by the current situation. Our research addresses the case of human-robot interactions, where human interaction with the robot will require the robot to smoothly and robustly change its level of autonomy.

The need for adjustable autonomy is clear in situations where intelligent mobile robots must interact with humans.

Consider the following examples:

Several dozen micro air vehicles are launched by a Marine. These vehicles will have a mission to perform, but depending on the unfolding mission, some or all of the vehicles may need to be redirected on the fly, at different times, and then be autonomous again.

Groups of autonomous underwater vehicles involved in salvage or rescue operations may start by autonomously searching an area, but then need to be interrupted by a human or another robot to be redirected to specific tasks.

A planetary rover interacts with human scientists. Because of the communication time lag in this situation, autonomy is critical to the safety of the vehicle. However, the human must be able to exert lower levels of control to perform various experiments.

In many tasks, the human will be exerting control over one or more robots. At times, the robots may be acting with full autonomy. However, situations will arise where the human must take low-level control of individual robots for short periods, or take intermediate level of control over groups of robots, for example, by giving them a new short-term goal which overrides their current task. The robots must be able to smoothly transition between these different modes of operation.

Intelligent mobile robots that require *close* interaction with humans will require natural modes of communication, such as speech and gestures. Our previous work examined the use of multiple modes of communication to disambiguate the communication between a human and a robot.

In this research, we explore the use of *context predicates* to keep track of various goals during human-robot interactions. These context predicates allow the robot to maintain multiple goals, each with possibly different levels of required autonomy. They permit direct human interruption of the robot, while allowing the robot to smoothly return to a high level of autonomy.

In the following paper, we will describe the robot platform and supporting software. Next, we will describe our previous work on multi-modal communication that involved attaining single goals. Next, we will describe our proposed use of context predicates to track multiple goals. We will conclude with some general thoughts on how our current work can be applied to achieving adjustable autonomy.

Robotic Platform

The methods by which gestures are perceived and interpreted and the natural language input integrated to produce appropriate robot commands are discussed in our previous work (Perzanowski, Schultz, and Adams 1998), but we outline them briefly here.

For our research in developing a natural language and gestural interface to a mobile robot we have been employing a Nomad 200 robot (see Figure 1), equipped with 16 Polaroid sonars and 16 active infrared sensors.

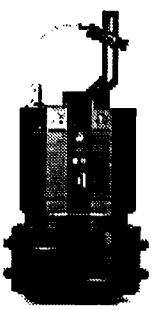

Figure 1: A Nomad 200 mobile robot with mounted camera

Gestures are detected with a structured light rangefinder emitting a horizontal plane of laser light. A camera mounted above the laser is fitted with a filter tuned to the laser frequency. The camera observes the intersection of the laserlight with any objects in the room, and the bright pixels in the camera's image are mapped to XY coordinates.

Periodically, the data points from one camera frame are used to compute an average distance from the objects seen and then sorted into clusters. Any cluster sufficiently closer than the average and of appropriate size is designated as a hand. Hand locations are stored for multiple frames until no hands are found or a maximum number of frames are used.

The hand locations across the frames are ordered into one or two trajectories. Completed trajectories are checked to see if they are in motion or are stationary, and then logically compared to determine if the overall gesture is valid and if so which gesture was made. The valid gestures are queued and when the multimodal software needs to check for a gesture, it queries the gesture process which returns the most recent gesture from the queue.

Multi-Modal Communication: Single Goals

The first stage of our interface was built relying on the interaction between natural language and gesture to disambiguate commands and to provide complete information where one of the two channels of communication lacked some specific or required information. Thus, for example, the utterance, "Go over there" may be perfectly understandable as a human utterance, but in the real world, it does not mean anything if the utterance is not accompanied by some gesture to indicate the locative goal.

For this work, we assumed that humans frequently and naturally use both natural language and gesture as a basis for communicating certain types of commands, specifically those involved in issuing directions. When a human wishes to direct a mobile robot to a new location, it seemed perfectly natural to us to allow the human the option of using natural language or natural language combined with gesture, whichever was appropriate and produced a completely interpretable representation which could then be acted upon.

Coincidentally, we did not incorporate any hardware devices, such as gloves (McMillan 1998), for inputting gesture information. In order to keep our interactions as "natural" as possible, we have not included such devices as gloves which would, in some sense, restrict the human in interacting with the robot.

Furthermore, we did not permit gestures in isolation because we believed that their use took the communicative act out of the natural realm and made it a more symbolic act of communication, which we did not wish to pursue at this point. We, therefore, are not ruling out isolated, symbolic gestures or symbolic gestures in combination with speech as possible means of efficient interaction with mobile systems. We simply leave their consideration for future work.

In our previous work (Perzanowski, Schultz, and Adams 1998), we outlined how natural language and gesture were used largely to disambiguate verbal input for command and control of a mobile robot, and in a more general way with work in spatial cognition and computation (Wauchope et

al. 1997). This work was based on the premise that human-machine interaction can benefit from the natural way in which humans interact with each other during normal dialogs; namely, by utilizing both natural language and gestural input as complementary means to communicate commands.

Just as others, such as (Konolige and Myers 1998), have attempted to incorporate gesture recognition as part of the attention process in human-robot interactions, we have incorporated it naturally, along with natural language. However, we restrict the types of communication in this interface to a model of communication characterized as a *push* mode (Edmonds 1998). By this, we mean to characterize our interface as one in which the human basically provided all the input, and the mobile robot acted as a passive agent, reacting only to those commands issued by the human participant.

However, even though our interface remains in the *push* mode to date, despite the adaptations outlined below, we still believe we will be able to produce a much more independently reactive robotic agent. Furthermore, as our research continues, we believe it will open up opportunities to achieve the kinds of autonomy we discuss here.

In the earlier version of our system, utterances are fully parsed syntactically and a semantic interpretation is obtained, utilizing our in-house natural language processing system (Wauchope 1994). Given the vision capability on the Nomad 200 robot and the processing as outlined above, when the sensors on the robot detect a vector within the limitations of its light striping sensor, and a command is sent to move in some direction, a query is made of the gesture process on the robot to see if some gesture has been perceived. The two inputs, the semantic interpretation mapped into a command interpretation and the gesture signal, are then mapped to a message, which is then sent to the robot in order to produce an appropriate action. The mapping of the speech input and the perceived gesture is a function of the appropriateness or inappropriateness and the presence or absence of a gesture during the speech input.

Previously, in our work and in (Yamauchi et al. 1997) we showed how both natural language and gesture could be employed to provide a more "natural" or human means of interaction and an efficient method of communication, thereby integrating the two channels of communication in this domain. However, input was restricted to commands that involved achieving only one goal. If any interruptions occurred, or if intervening goals made it necessary for the primary directive to be kept on hold, that system was incapable of performing appropriately.

Despite these noted limitations, a brief overview of how that system performed can be helpful.

A Brief Overview of System Capabilities

If a human wants the robot to move to a new or different location, the human can either utter a sentence, such as one of the sample set of sentences in (1), or the human can utter a sentence along with performing an appropriate gesture.

(1)(a) Go to the left.
(b) Move to the right.
(c) Move this way.
(d) Go to the waypoint over there.

Assuming optimum performance of the system, and complete input being provided by the human participant, an appropriate robot response is obtained. However, we immediately noticed that we needed to incorporate some sort of mechanism for the robot to recover from erroneous or incorrect input.

Thus, for example, if (1a) or (1b) are uttered while the human points in some direction other than in the appropriate direction, the system needs to inform the human that contradictory input is being received and further action can only be taken upon correction.

Likewise, if (1c) or (1d) are uttered with no accompanying gesture, the system should return a request for additional information, either verbal or gestural clarification of the "deictic" or referred-to direction or object in the sentence. In our work, however, we do not employ any symbolic referents, as in for example (Wilkes et al. 1998) or (Kortenkamp, Huber, and Bonasso 1996).

Our gestures are perfectly natural and indicate directions and distances in the immediate vicinity of the two participants of the interaction, namely the human and the robot. Therefore, if (1d) is uttered with no accompanying gesture, the system responds that it needs information about the location of the waypoint, and furthermore, if the human utters (1d) and gestures in a direction in which no waypoint is located, the robot responds appropriately that there is no waypoint in that direction.

This first version of the interface, therefore, permitted a natural way for humans to interact with a mobile robot that had a well-defined but limited vision capability. It permitted recovery from error, but it constrained the human to certain types of verbal input and ultimately constrained the capabilities of the robot in a very specific sense, namely the system could only process one command or goal at a time. Intervening interruptions or unforeseen goals simply caused the system to fail.

We now turn to our proposal to use context predicates to enhance the system's capabilities in goal achievement, thereby introducing a capability to provide greater autonomy in human-robot interactions.

Multi-Modal Communication: Multiple Goals

As a first step in our attempt to provide greater autonomy in robotic control, the natural language and gestural interface was enhanced to enable the processing of incomplete and/or fragmentary commands during human-robot interactions. This enhancement has enabled us to keep track of various goals during human-robot interactions by instantiating context predicates, which are basically the topical predicates at various stages of the

interaction. (We will narrow this broad definition in the following discussion.) By utilizing these context predicates, a discourse component of the interface tracks the goals of the interaction, and records exactly which and to what extent each goal was achieved. With this information and by performing certain logical operations on semantic information of the context predicates, the robot can continue to achieve any unaccomplished goals on its own, no matter at what point or in what state the system is currently.

Tracking of the goals, by means of these context predicates, permits the system to work independently on achieving previously stated, but as yet uncompleted, goals. In this sense greater autonomy is achieved, since users can expect the robotic system to be able to continue its performance and accomplish previously stated goals or subsequent logical goals, without the user having to explicitly state or re-state each expected or desired action.

As part of the continuing research in natural interfaces to mobile robots, we propose the addition of certain discourse capabilities which will expand the semantic component of the natural language interface and interact in a rather unique way with the gestural interface to the mobile robot. These additions, we believe, provide greater independence on both the part of the human and the mobile robot in command and control or task-oriented interactions. We believe that this increase in independence for both participants in these interactions provides opportunities for allowing robotic systems to be more autonomous, while at the same time being aware of what the human partner wishes.

Thus, depending upon what tasks have been completed or are still to be completed, the robot can go off on its own to perform those tasks as initially directed, or it can be interrupted, either permanently or temporarily by the human partner as so desired. Proceeding either with a new task, or returning to previous uncompleted actions can then be accomplished by the system's knowledge of what actions have been accomplished or still need to be accomplished, based on the system's knowledge of the status of the tasks thus far issued. We also use a kind of prioritization of tasks to determine which actions need to be accomplished when several tasks remain to be completed.

We mention this stage of our research here, because it was the completion of this step that led us directly to our consideration of achieving greater freedom of interaction, independence, and ultimately, a way of addressing adjustable autonomy in human-robot interactions.

We noticed that while we were trying to emphasize naturalness in human-machine interactions, many of our interactions were unnatural in the following sense.

During human-human interactions, the participants in a dialog frequently rely on fragmentary responses, rather than repeating grammatically full or complete utterances. For example, the dialog of (2) seems much less natural than its corresponding (3), which incorporates fragmented utterances.

(2)
Participant I: Go to the waypoint over there.
Participant II: Where?
Participant I: Go to the waypoint over there.
(with accompanying gesture.)

(3)
Participant I: Go to the waypoint over there.
Participant II: Where?
Participant I: Over there. (with accompanying gesture.)

In Participant I's final utterance of (3), a sentence fragment or incomplete thought is uttered. It is juxtaposed here with its corresponding complete sentence or utterance in (2).

Natural language systems require the information found in complete sentences, such as those in (2), for interpretations to be obtained and subsequent actions to be taken in a task-oriented application, like interacting with a mobile robot. Fragmentary input, such as the last utterance in (3), does not provide sufficient information for an appropriate message to be passed to the robot and subsequent action to be taken. The fragment simply does not contain all the necessary information—namely the action to be taken. Only if the information in the discourse is somehow kept or tracked, will the more natural interchange of (3) be acceptable in human-machine interactions.

Therefore, if a natural language/gesture interface keeps track of the various actions, embodied here in the verbal predicates of the various utterances, then not only can the utterance be parsed and interpreted in its fullest sense, but a record of what actions have and/or have not yet been accomplished can be kept. This information can be used for later purposes, as we further discuss.

We decided to keep track of the various predicates in our interactions, since the utterances in this application tended to be grouped into task-oriented or goal-oriented actions or behaviors. Most, if not all, of the actions dealt with movement, such as turning, or obtaining some goal, such as going to a particular waypoint. We instantiated what we call context predicates.

A context predicate is basically the verbal or action-part of an utterance. In linguistic terms, the context predicate is the predicate of the sentence. It also contains a prioritization code to which we turn later. For example, the context predicate in (4) is the expression "move left."

(4) Move to the left.

Notice that the context predicate does not contain some of the lexical items in (4). While (4) is a rather trivial example of a context predicate, basically the context predicate embodies the action item of the sentence, as well as any of the arguments associated with that action. Therefore, an interpretation of Participant I's incomplete utterance in (3) above relies upon the system's knowledge

that the context predicate for this sentence incorporates information from the previous utterance of Participant I; namely, "go to the waypoint."

Typically, fragmentary utterances in human dialogs can be interpreted by selecting the predicate from the participant's immediately preceding utterance (we will refer to this as the context predicate hereafter) and by eliminating any redundancies between the current and the former utterances. Human dialog does not typically permit fragmenting to occur over intervening complete utterances.

For example, the following is not a well-formed dialog:

(5)
(a) Participant I: Go to the waypoint over there.
(b) Participant II: Where?
(c) Participant I: Move to the left.
(d) Participant II: <moving to the left>
(e) Participant I: Over there. (with accompanying gesture.)

There is simply no way that Participant I in (5) can jump back in the conversation and provide a fragmented utterance, such as (5e), to correct an incomplete utterance (5a) after another utterance (5c) intervenes. The only way that Participant I can get Participant II to go to the referred-to waypoint is to repeat the entire utterance again, and produce an appropriate gesture if necessary.

Schematically, we can represent analysis of obtaining context predicates in Figure 2.

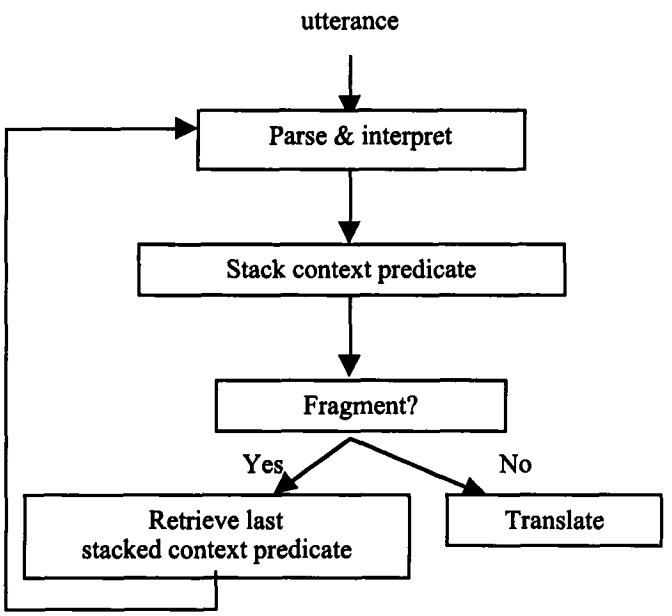

Figure 2: Schematic flowchart for obtaining context predicates

If a fragmentary utterance is produced during some interaction, it can be completed by checking a proximate context predicate in the discourse, thus providing a complete utterance for final interpretation and subsequent translation to an appropriate robot action.

So, context predicates initially provided us with a way of retrieving information in fragmentary input. This certainly allowed the human to produce more natural utterances during interactions with the robot, but we also discovered that if we kept track of these predicates, and checked our semantics for concomitant actions required to achieve a particular context predicate and stacked them, and introduced a prioritization schema to those actions, we could actually provide the robot with a greater degree of freedom to perform actions.

For example, our semantics stipulates that going to a waypoint consists of three concomitant actions:

1. (get gesture)
2. turn in direction of goal
3. go to location indicated

With this semantics for going to a waypoint, therefore, the robotic system in some sense knows what actions are necessary to complete an action and achieve a goal. (The first step in getting to a waypoint is in parenthesis simply to indicate that it is optional if the actual waypoint is stipulated, as in "Go to waypoint 2," but is necessary if the waypoint is merely referred to by some sort of deictic element, as in "Go to the waypoint over there.")

Of course, the question now is: specifically how does the use of context predicates facilitate independence or provide for greater autonomy on the part of the robot?

We answer this question by calling upon the system's ability to track goals, as they are embodied in the context predicates.

After being given a command, such as "Go to the waypoint over there," the system is free to go off and achieve this goal. However, suppose the system is interrupted in trying to achieve this goal, say for example, after step 2 in going to a waypoint; namely, after the robot has turned in the direction of the goal. For some reason, at this point the human stops the robot and wants it to move over a little bit. Of course, the robot should obligatorily move over, as the human has requested. Once this interrupting goal is achieved, and the human issues a command to continue what it was doing, the robot should be able to pick up where it left off. In this example, the robot was in the middle of trying to get to a waypoint. Achieving that previously stated goal should, therefore, proceed.

The ability of the system to interrupt action, achieve interrupting goals, and return to previous uncompleted goals and tasks is accomplished in our system by utilizing context predicates and tracking the steps in achieving the various goals. By stacking the goals, as they are embodied in the context predicates, and checking the domain semantics for sub-goals, any action can be interrupted at

any point and can be returned to for completion at a later time, because the system has kept a history of unachieved goals.

Once the interrupted goal is achieved, of course, the system can then retire attained goals and work on the more immediate situation as it is presented.

We believe this does exhibit system autonomy in the sense that even though the robot may be interrupted at any point in its attempts to achieve a goal, it can return to a prior as yet unattained goal and achieve it, without having specifically to be told what to do by restating any previous point, which may or may not be remembered by the human in the interaction.

Although we have outlined a way in which our natural language/gesture interface can achieve goal tracking and provide a way of achieving autonomy, the system is not yet fully implemented. We are currently working on fine tuning our semantics component so that it is capable of handling both context predicates and the concomitant actions required in achieving particular goals in our robotics domain. We will, as a result, also have to coordinate this semantic interpretation with the actual steps that the robot must take in order to achieve not only the ultimate goal of a command, but the necessary steps to achieve that goal.

Conclusions

We are currently investigating ways to utilize context predicates and goal tracking to permit humans and robots to act more independently of each other. As situations arise, humans may interrupt robot agents in accomplishing previously stated goals. Context predicates allow us to keep track of those goals and the necessary steps in achieving them. After interruptions, therefore, the system can return to complete interrupted actions, because the system has kept a history of which goals have or have not been achieved. This capability of our system allows both the human and the robot in these interactions to work at varying levels of autonomy when required. Humans are not necessarily required to keep track of robot states. The system does, and the robot is capable of performing goals as they are issued, even if an intervening interruption prevents an immediate satisfaction of that goal.

We intend to conduct experiments on the enhanced system in the near future with the intention of incorporating empirical results of those studies for future pulication.

The incorporation of context predicates to track goals will be a necessary capability to allow adjustable autonomy in robots, which in turn permits the kinds of interactions and communication in the mixed-initiative systems we are developing.

Acknowledgements

This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

References

Edmonds, B. 1998. Modeling Socially Intelligent Agents. *Applied Artificial Intelligence* 12:677-699.

Konolige, K. and Myers, K. 1998. The Saphira Architecture for Autonomous Mobile Robots. In Kortenkamp, D., Bonasso, R. P., and Murphy, R. eds. *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, 211-242, Menlo Park, CA: AAAI Press.

Kortenkamp, D., Huber, E., and Bonasso, R.P. 1996. Recognizing and Interpreting Gestures on a Mobile Robot. In Proceedings of the Thirteenth National AAAI Conference on Artificial Intelligence, 915-921. Menlo Park, CA: National AAAI Conference on Artificial Intelligence.

McMillan, G.R. 1998. The Technology and Applications of Gesture-based Control. In RTO Lecture Series 215: Alternative Control Technologies: Human Factors Issues, 4:1-11. Hull, Québec: Canada Communication Group, Inc.

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

Wauchope, K. 1994. Eucalyptus: Integrating Natural Language Input with a Graphical User Interface, Technical Report, NRL/FR/5510--94-9711, Navy Center for Applied Research in Artificial Intelligence. Washington, DC: Naval Research Laboratory.

Wauchope, K., Everett, S., Perzanowski, D., and Marsh, E. 1997. Natural Language in Four Spatial Interfaces. In Proceedings of the Fifth Conference on Applied Natural Language Processing, 8-11. San Francisco, CA: Morgan Kaufmann.

Wilkes, D.M., Alford, A., Pack, R.T., Rogers, T., Peters II, R.A., and Kawamura, K. 1998. Toward Socially Intelligent Service Robots. *Applied Artificial Intelligence* 12:729-766.

Yamauchi, B., Schultz, A., Adams, W., Graves, K., Grefenstette, J., and Perzanowski, D. 1997. ARIEL: Autonomous Robot for Integrated Exploration and Localization. In Proceedings of the Fourteenth National AAAI Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, 804-805. Menlo Park, CA: National AAAI Conference on Artificial Intelligence.