

Spontaneous, Short-term Interaction with Mobile Robots¹

Jamieson Schulte Charles Rosenberg Sebastian Thrun

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Human-robot interaction has been identified as one of the major open research directions in mobile robotics. This paper considers a specific type of interaction: short-term and spontaneous interaction with crowds of people. Such patterns of interactions are found when service robots operate in public places, for example information kiosks, receptionists, tour-guide robots applications. We describe our approach to spontaneous short-term interaction: a robot designed to be a believable social agent. The approach has been implemented using a mobile robot with a motorized face as focal point for interaction, an architecture that suggests the robot has moods, and a method for learning how to interact with people. Our system was recently deployed at a Smithsonian museum in Washington, DC. During a two week period it interacted with thousands of people. The robot's interactive capabilities were essential for its high on-task performance, and thus its practical success.

1 Introduction

Human-robot interaction has been identified as one of the major open research directions in mobile robotics.[4] Interactivity is essential for an upcoming generation of service robots, which will directly interact with people. For example, these robots may assist elderly or handicapped people, assist humans in search-and-rescue missions, or perform janitorial services in environments populated by humans. Thus, interfaces for human-robot interaction are important for the practical success of such systems. In each of the cases just described, the human-robot interaction is typically one-on-one and it is possible to train the user (and the robot) in the vocabulary of the interface. However, in certain service robot applications, such as robotic *receptionists*, *information kiosks*, or *tour-guides*, it is necessary for the robot to interact spontaneously with completely untrained people.

In this article we focus on *spontaneous short-term human-robot interaction*, typical of the robotic applications just described. These robots are often approached by groups of uninformed people, and typical interactions

last for 10 minutes or less. This is in contrast to human-robot interfaces proposed by various researchers which utilize gesture, speech, clapping, and natural language based interfaces. These interfaces are generally effective for a specific class of interactions. Gestures, for example, are well-suited for directing a mobile robot to manipulate (e.g. pickup) specific objects.[8, 11, 16] Speech input has been demonstrated to be highly effective for tasks such as teleoperating robots, or attaching names to places in unknown environments.[1] However, such interfaces are targeted toward scenarios where a *single* person interacts with a robot and typically fail in *crowds* of people.

Our approach to spontaneous, short-term interaction is a robot which operates as a believable social agent. Specifically, we employ a tour-guide robot application to investigate the utility of our approach. Tour-guide robots are usually approached by crowds of people, most of whom have never interacted with a robot before. In our system, a tour-guide robot has three main goals during its operation:

- **Traveling** from one exhibit to the next during the course of a tour.
- **Attracting** people to participate in a new tour between tours.
- **Engaging** people's interest and maintaining their attention while describing a specific exhibit.

The main functional components necessary for the robot to accomplish these goals during its operation are *navigation* and *interaction*. By *navigation*, we mean the ability of the robot to localize itself in a map, plan a motion path to a target, and avoid obstacles. In many robotic systems *navigation* and its related subcomponents alone might be sufficient for the robot to accomplish its goals. However, in the tour-guide application, the robot is in an environment crowded with people and its primary function is to provide a service to the people in its environment. To this end, *interaction* is as essential as *navigation*.

To make interaction effective, our approach is to create a system which acts in a *believable* manner while interacting with people in the context of *spontaneous short-term interaction*. A believable agent creates the impression that it is self-determining, and is an idea that has been previously

¹To appear in ICRA'99.

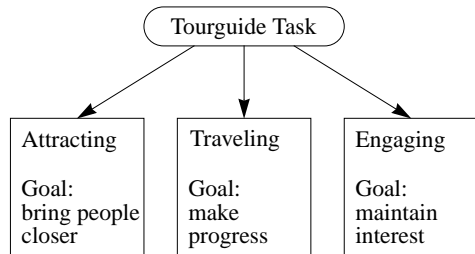


Figure 1: Decomposition of the tour-guide interaction problem.

considered in both software [2] and robotic [5] agents. We have created and tested a user interface for a robot with the goal of allowing it to act as a reasonable social agent in the specific context of the application described here, not under all possible conditions. In our approach, the three cornerstones which together create the impression of a *believable* agent are:

- **Focal Point**
- **Emotional State**
- **Adaptation**

The *focal point* provides people with a single location on which to focus their attention during interaction. In our implementation the *focal point* for human interaction was realized by a specific hardware interface consisting of a motorized face with pan and tilt control on top of the robot. The system communicates an *emotional state* to the people around it as a means of conveying its *intention* in a way that is easily understood in the context of a believable social agent. For example, a robot tour-guide might have the *intention* of making progress while giving a tour. In our system the expression displayed on the motorized face and the contents of the recorded speech playback communicate this information. *Adaptation* is the ability of the system to learn from its interactions with people and modify its behavior to elicit the desired result.

We recently designed and installed such a tour-guide robot, called Minerva, in the entrance area of the Smithsonian National Museum of American History, where it interacted with thousands of people over a two-week period. In this paper we describe its basic architecture and survey the results obtained in the museum. The Minerva robot had two basic intents: (1) to attract people to whom it could give a tour, and (2) to make progress while giving a tour. Both intents are somewhat orthogonal: for the former, the robot brings people closer to motivate their interest, while the latter requires people to clear the way, hence stay behind the robot. This paper describes a number of mechanisms that were found essential in the pursuit of these two goals. It first describes a specific hardware interface, which served as the *focal point* of human-robot interaction. It then describes two quite complimentary solutions, one for each goal described above. To make progress, the robot communicates an *emotional state* or “mood” to the viewers. To

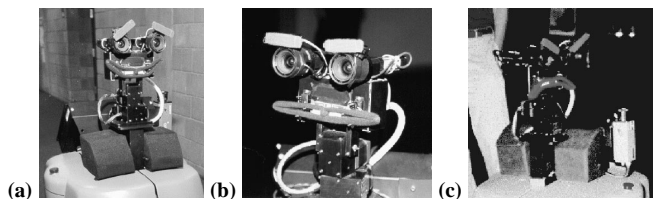


Figure 2: Minerva’s face with (a) happy, (b) neutral, and (c) angry facial expressions.

attract people, Minerva uses a learning algorithm that adaptively determines the best action out of a pool of possible actions (consisting of speech acts, head motion primitives, and facial expressions).

To evaluate the utility of the proposed methods for spontaneous short-term interaction, this paper compares Minerva with a different robot, called Rhino [3], which was built by the same group of researchers. In mid-1997, Rhino was installed as a robotic tour-guide in the Deutsches Museum Bonn. Both robots essentially use the same control and navigation software; the major difference lies in the nature of the interaction: Rhino did not possess any of the human-robot interfaces described in this paper. As a result, Rhino’s ability to attract people was much inferior when compared to Minerva, and it was much less effective when giving tours, as reflected by the rate of progress when moving from exhibit to exhibit. We largely attribute these differences to the interface, which proved essential for Minerva’s success and effectiveness.

2 Approach: Minerva The Robot

Our approach to the problem of making Minerva a believable agent that uses interaction to reach its goals contains three parts. First, a face is used to define a focal point for interaction. Second, the robot is supplied with an “emotional” state, expressed outwardly by facial expressions and sounds. Third, adaptation occurs in one of the interaction tasks using a memory based learner. We describe these aspects of Minerva, with an explanation of how each contributes to the goals of specific tasks.

2.1 The Face

At this point in time, there exists little precedence for robotic interaction with novice users upon which to build a new system. Hence, to engage museum visitors, it was in our interest to present as recognizable and intuitive an interface as possible: a caricature of a human face.[9, 10, 15] It was important that the face contain only those elements necessary for the degree of expression appropriate for a tour-guide robot. A fixed mask would be incapable of visually representing mood, while a highly accurate simulation of a human face would contain numerous distracting details beyond our control. An iconographic face consisting of two eyes with eyebrows and a mouth is almost universally recognizable, and can portray the range of simple emotions

useful for tour-guide interaction. Figure 2 shows three possible expressions realized by different configurations of the face hardware.

We determined, also, that a physically implemented face would be more convincing and interesting than a flat display.[9] Reasons for this include the expectation that moving objects require intelligent control, while flat moving images likely result from the playback of a stored sequence as in film or television. Additionally, a three-dimensional face can be viewed from many angles, allowing museum visitors to see it without standing directly in front of the robot.

The face has four degrees of freedom which were implemented via servo motors controlled by a serial port interface. One degree of freedom was used to separately control each eyebrow and two degrees of freedom were used to control the mouth. The face control motors were mounted on and arranged around a central box. The “eyes” of the robot were a pair of color CCD cameras. These cameras were not used for navigation or obstacle avoidance, but were present for the sole purpose of transmitting a robot’s eye view of the museum to web visitors. The eyebrows, consisting of blue rectangles, are mounted directly above the cameras. The eyebrows can independently move ± 90 degrees from horizontal. The mouth consisted of a red elastic band. Each end of the band was mounted to a servo control arm and its motion was constrained by three pins. Even though both sides of the mouth could be controlled independently, they were controlled in a coordinated, symmetrical fashion to bring the “mouth” into a smiling or frowning configuration. Because of the mouth design and low bandwidth of the actuators, it was not possible to make the “lips” move in synchronization to the speech generated by the robot. Instead, a bar graph LED display was mounted behind the mouth, which illuminated in response to the speech generated by the robot. Two such displays were mounted in mirror image fashion back to back such that when the robot spoke, the length of the displayed bar increased symmetrically from the center of the mouth. The head was mounted on a pan/tilt unit which allowed it to be rotated approximately ± 90 degrees from center and tilted slightly from the horizontal.

The face hardware installed on Minerva served a second purpose beyond communicating its *intent*; it provided a *focal point* for the interaction between the human and the robot. By *focal point*, we refer to a place for a human to focus attention and better understand that the system will follow some basic social conventions. People focused attention on Minerva’s face when interacting with it. As anecdotal evidence, visitors tended to take photographs of just Minerva’s face, whereas in the case of Rhino, people tended to take pictures of the entire robot. People understood that the robot’s face pointed in the direction it intended to go, even when the robot was stopped. Sim-

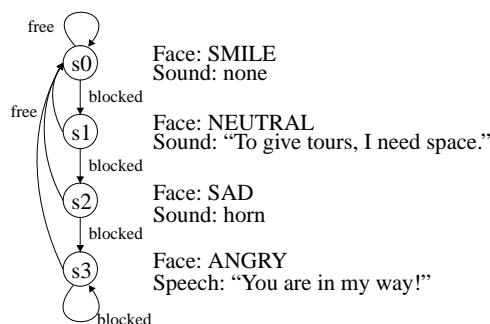


Figure 3: State diagram of Minerva’s emotions during travel. “Free” and “blocked” indicate whether a person stands in the robot’s path.

ilarly, the LEDs placed behind the mouth provided a focal point when speech was generated, localizing the sound there, even though it was produced by speakers on the robot base.

2.2 Emotional State

Minerva’s emotional state is the basis of its travel-related interaction. Travel occurs between stops in a tour when Minerva moves through the museum and finds its way to the next exhibit to discuss. To navigate through crowded spaces, the robot decides whether an obstacle is a human by using of an entropy filter applied to the laser range data and the museum map.[7] If the robot is being blocked by a person, it communicates its *intent* to those who are in the way. Possibly, the most effective way to do this would be to loudly and aggressively state that everyone should step away. However, another implicit objective of our robot is to interact in a friendly and socially acceptable manner. To communicate its intent to make progress in a particular direction, Minerva utilizes its interface: an expressive face, a pan/tilt head, and speech output. It is with these “effectors” that Minerva manipulates the environment around it. Our solution combines these behaviors in a simple state machine, where state is represented externally as a mood. Please note that by mood, we do not suggest that this system has the highly complex property of “emotion,” we simply use the term to indicate an emotional state that the person observing Minerva would impart to it.[5, 13] In this work we view “mood” from an engineering view point — it is nothing more than a means to an end. We feel this sets Minerva apart from other agents which utilize emotion as part of their interface [6, 12, 14].

The emotional state machine encodes the complete travel interaction behavior in a total of four states, as shown in Figure 3. Minerva starts in a “happy” state, smiling while traveling between tour stops, until first confronted by a human obstacle that cannot be trivially bypassed. At this point, the robot kindly announces that it is giving a tour and changes from a smiling to neutral expression, while

Feature	Values
facial expression	happy, neutral, sad, angry
face pointing target	closest person, center of mass of people, least populated area, random direction
sound output	happy speech, “clap your hands”, neutral speech, horn, aggressive speech

Table 1: During the adaptive behavior, an action is performed by setting each of the three features to one of the pre-defined values listed above.

pointing its head in the direction it needs to travel. If this does not bring success, Minerva adopts a sad expression, and may ask the obstructing person to stand behind it. This usually makes sense in context, since the direction the head points suggests a “front” and “back” of the robot. If the person still does not move, then Minerva frowns and becomes even more demanding.

Emotional state helps Minerva achieve navigational goals by enhancing the robot’s believability. Observation of interaction with museum visitors suggests that people are generally unconcerned about blocking the path of a passive, mute robot. A change of facial expression and sudden utterance by Minerva usually results in a quick response from anyone in the way. (One side-effect is that some people wish to find out how much they can perturb the robot, and will intentionally prevent it from moving.) Our subjective interpretation of the effect of emotional state is that the increasing “frustration” of the robot produces feelings of empathy in many people and coerces them to move. This empathy is possible, we think, because the timely and exaggerated transition of moods lends Minerva a believable personality in this limited context.

2.3 Adaptation

Between tours, Minerva spends approximately one minute generating interaction behaviors with the goal of attracting people to follow it on the next tour. We chose to experiment with learning interactive behaviors by having the robot select actions, then evaluate them based on the movement of people in the period of time following the new action. An action is defined to be a joint setting of three features: a facial expression, a pan/tilt target for pointing the face, and a sound type. The state is a discretization of the distance and density of people standing near the robot. A memory-based learner (MBL) is used to store the results of interaction experiences in order to make future decisions when confronted with the same task. A performance function maps the sequence of movements by people following an action into a single scalar value that we refer to as a *reward*, indicating the relative success of the behavior. The function was defined such that an increase in closeness and density of people around the robot was rewarded and a decrease was penalized.

Interaction with humans by a robot presents a unique and challenging learning problem. The realm of possible actions with different meanings in an interaction setting is

enormous. Subtle changes in the speech timing and volume, or in the intensity of a facial expression can affect the quality of interaction significantly. The effect of a given action is not constant, and much of the state that could help define specific state/action pairs is hidden to a robot with limited sensing capability. In particular, our robot is unable to detect anything more about the humans with whom it is interacting than their distances and spatial densities. Given this, we chose a very biased and limited, but learnable, space of overall interaction possibilities. The range of possible robot behaviors was selected to include obviously “good” and “bad” actions, and the overall cadence of interaction and selection of spoken phrases was fixed.

Specifically, Minerva enters an “attraction interaction” state for one minute between museum tours, where the goal is to attract people in preparation for the next tour. In this state, an action is initiated consisting of facial expression, face pointing direction, and sound output. This action persists for 10 seconds, after which a new action is selected. During this interval, the distances and densities of people around the robot are monitored and used to evaluate the effect of the action. The evaluation result, or reward, is stored by the MBL. The next action is selected by choosing that which maximizes the expected reward given the learner’s previous experiences and the current state. Some features of this new action are occasionally randomized to ensure that new regions of the action space are explored. The action space for this phase of interaction is outlined in Table 1.

After some experimentation we chose a very simple learning strategy. The MBL chooses an action a such that:

$$\max_{a \in A} m(a, s)$$

where A is the set of all 80 possible action combinations, and $m(a, s)$ is simply the weighted mean of previous rewards following action a in states near s . If no experiences with a have been recorded, then $m(a, s)$ returns zero, which corresponds to the reward following an action that produces no change, positive or negative, in the distribution of people around the robot. The simplicity of this approach reflects the difficulty of collecting sufficient data in a noisy environment. The algorithm described above is *on-line* in the sense that learning occurs continuously and the results of experiments immediately affect future actions, without human intervention or the execution of a separate training step.

3 Results

3.1 Travel Interaction

In the museum environment, a tour-guide robot is often surrounded by people which impede its forward progress (Figure 3). An examination of the average speed of Minerva (38.8 cm/s) showed it to navigate more quickly than

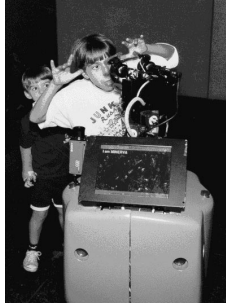


Figure 4: Interaction helps Minerva navigate through crowded environments.

the Rhino robot (33.8 cm/s), even though Minerva operated in a considerably more populated environment. We attribute this to the fact that Minerva could more efficiently and clearly indicate its intended direction of travel. Also, in terms of entertainment value, Minerva’s behavior during this time is more interesting to the people who follow the robot. Others have also found interfaces similar to Minerva’s to have entertainment value.[12, 14]

From observation, it was clear that museum visitors understood the changes in mood brought about by obstructing Minerva. While not everyone chose to move, the robot’s *intentions* were quite clear. In the case of the faceless robot Rhino, a horn sound was used to clear people from its path obstructed. People found this signal to be ambiguous, and did little to impart the believability that helped Minerva influence people.

3.2 Attraction Interaction

Minerva performed 201 attraction interaction experiments, and over time become a more “friendly” robot that attracted people more successfully. The distances of people around the robot is an inherently noisy measure of the success of an interaction behavior. Nevertheless, we have seen promising indications that some basic adaptation and parameter tuning within a pre-defined behavior can work to make an agent more flexible.

Ultimately, we expect that this flexibility can enhance believability. Figure 5 shows the learned expected reward for different types of behavior at the end of the experiments. The first plot compares “negative” and “positive” actions. Negative actions are those for which Minerva makes a demand of the visitors in a stern voice while frowning. Positive actions consist of friendlier comments and a neutral or happy facial expression. The numbers were produced by taking a weighted average of the value of the expected reward function $m(a, s)$ for all actions belonging to the category being analyzed, over all states. The larger confidence interval for “negative” actions reflects the fact that less data was collected by Minerva in this less promising region of the action space, since the exploration strategy was biased toward successful actions.

The second plot in Figure 5 compares the expected reward resulting from the five categories of sound that Minerva can produce. Here we can see a clear tendency

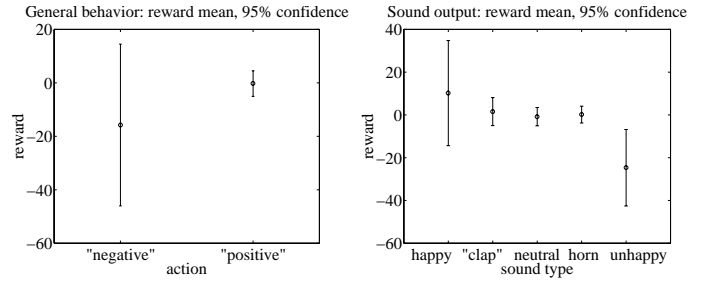


Figure 5: Minerva’s expected reward for different sets of actions. *Left*, A comparison of “positive” (friendly) and “negative” (unfriendly) actions. *Right*, five different categories of sounds produced by Minerva, with reward averaged over all other action features.

for happy sounds to produce greater reward than neutral sounds, and for upset sounds to result in a penalty. The fact that the horn sound falls in the neutral reward category sheds some light on the difficulty that Rhino had convincing people to move in previous research. While these figures are of limited significance, there is a promising trend of increasing reward with friendlier behavior. Due to the noisiness of the data relative to the number of experiments, and the fact that we could perform only one training session, a plot of the performance increase over time would not be meaningful.

It should be noted that the performance function finally settled upon is slightly different from that used during early training. Thus the results presented are not exactly those learned on-line in the museum. During the initial experiments, there was a penalty for actions that brought people within several inches of the robot, so some actions that were most successful at engaging visitors were only modestly rewarded. The impact on these results is that “friendly” actions may have a lower reward estimate and greater reward variance than would be the case if the updated performance function were used throughout the learning process.

3.3 Visitor Surveys

To measure the subjective concept of Minerva’s believability, we asked a sampling of 63 museum visitors to answer a short questionnaire. Perhaps the most interesting estimate of believability results from answers to the question: “As far as intelligence is concerned, what would be the most similar animal? (amoeba, fish, dog, monkey, or human)” Figure 6 compares responses for the two age groups: 0 to 10 years and greater than 10 years. The bar between “monkey” and “human” is a count of respondents that suggested that Minerva fell somewhere between the two categories. Clearly, young children were more likely to attribute human-like intelligence to the robot. Most of this group (65%) also felt that Minerva was “alive”, while very few others would make this assertion. For the questions that we asked, gender played little role in perception of Minerva. The notion of intelligence does not directly correspond to believability, but it is encouraging to find Minerva

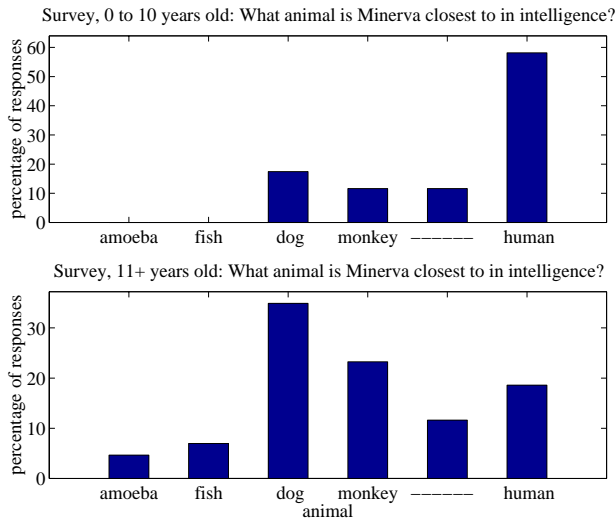


Figure 6: Survey responses comparing Minerva’s intelligence to that of 5 animals for respondents (top) 0-10 years old, and (bottom) 11+ years old. Responses labeled “——” indicate “between monkey and human.”

frequently compared to animals that we recognize as complex social creatures.

4 Summary and Conclusions

Interfaces for human-robot interaction are essential for an upcoming generation of service robots, which will have to directly interact with people. In this paper we focus on interfaces targeted toward *spontaneous, short-term interaction*. The Minerva tour-guide robot described in this paper is an example of a robot which interacts with people in this way.

Our experiments have demonstrated the usefulness of our approach for building such an interface. In our system this included: an expressive face, a head with pan and tilt control, and speech output. These systems allowed Minerva to be perceived as a *believable agent* and effectively communicate its *intent* to the individuals interacting with it. The Minerva robot was able to make progress through the museum during tours at the same rate as the Rhino robot, even though the Minerva robot encountered an order of magnitude more people. Both robots were similar, with the exception of the interaction component.

We experimented with both a hand coded solution and a learning based solution to action selection for this interface and found both to be effective. Because the space of possible interaction behaviors is so large, learning necessarily occurs within a limited action space. Nevertheless, we found that Minerva successfully learned to select actions that improved the effectiveness of interaction, using an on-line algorithm.

In conclusion, we have demonstrated that a robot system, with an interface that represents the robot as a *believable* social agent, can effectively exploit traditional social interactions between humans, to communicate *intent*

during *spontaneous, short-term interaction*. We view this mode of interaction as another tool in the interface designer’s tool box when building systems which need to interact with uninformed users and in environments where people may impede the robot in achieving its goals.

Acknowledgments

We would like to thank the Lemelson Center of the National Museum of American History for providing us a venue for conducting this research, and the Minerva team for its support. We would also like to gratefully acknowledge DARPA, for partially funding this research via AFMSC (contract number F04701-97-C-0022), TACOM (contract number DAAE07-98-C-L032), and Rome Labs (contract number F30602-98-2-0137).

References

- [1] H. Asoh, S. Hayamizu, H. Isao, Y. Motomura, S. Akaho, and T. Matsui. Socially embedded learning of office-conversant robot Jijo-2. *IJCAI-97*.
- [2] J. Bates. The role of emotion in believable agents. TR CMU-CS-94-136, Carnegie Mellon University, 1994.
- [3] W. Burgard, A. Cremers, D. Fox, D. Häehnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. The interactive museum tour-guide robot. *AAAI-98*.
- [4] J. Crisman and G. Bekey. Grand challenges for robotics and automation: The 1996 ICRA panel discussion.
- [5] K. Dautenhahn. The role of interactive conceptions of intelligence and life in cognitive technology. *CT-97*.
- [6] C. Breazeal (Ferrell). A motivational system for regulating human-robot interaction. *AAAI-98*.
- [7] D. Fox, W. Burgard, S. Thrun, and A. Cremers. Position estimation for mobile robots in dynamic environments. *AAAI-98*.
- [8] R. Kahn, M. Swain, P. Prokopowicz, and R. Firby. Gesture recognition using the Perseus architecture. *CVPR-96*.
- [9] W. King and J. Ohya. The representation of agents: Anthropomorphism, agency, and intelligence. *CHI-96*.
- [10] T. Koda. Agents with faces: The effect of personification. In *5th IEEE International Workshop on Robot and Human Communication*, 1996.
- [11] D. Kortenkamp, E. Huber, and P. Bonasso. Recognizing and interpreting gestures on a mobile robot. *AAAI-96*.
- [12] P. Maes. Artificial life meets entertainment: Interacting with life-like autonomous agents. *Communications of the ACM*, 38(11), 1995.
- [13] S. Penny. Embodied cultural agents: At the intersection of art, robotics and cognitive science. In *AAAI Socially Intelligent Agents Symposium*, 1997.
- [14] C. Rosenberg and C. Angle. IF: An interactive animatronic prototype. IS Robotics Inc. internal development project, <http://www.cs.cmu.edu/~chuck/robotpg/itpg/>, 1994.
- [15] A. Takeuchi and T. Naito. Situated facial displays: Towards social interaction. *CHI-95*.
- [16] S. Waldherr, S. Thrun, R. Romero, and D. Margaritis. Template-based recognition of pose and motion gestures on a mobile robot. *AAAI-98*.