In *IEEE Proceedings of the International Workshop on Robot and Human Interactive Communication (RO-MAN)*
Pages 416-421, Aug 13 -15 2005, Nashville, TN

A Multi-Modal Approach to Selective Interaction in Assistive Domains

David Feil-Seifer
Interaction Laboratory
University of Southern California
University Park, Los Angeles, California, USA
dfseifer@pollux.usc.edu

Maja J Matarić
Interaction Laboratory
University of Southern California
University Park, Los Angeles, California, USA
mataric@usc.edu

***Abstract*— Human-robot interaction (HRI) is an important area of robotics research; it is related to human-computer interaction (HCI), but contains a key difference: HRI allows embodied systems to utilize physical context and mobility. Most current HRI systems, however, do not yet utilize mobility for interactive purposes. In this paper, we describe the design and evaluation of a robot system aimed at embodied HRI communication. In it, the robot selects a target for interaction based on the perceived desire of human subjects to interact with it. We use an interaction policy inspired by Joint Intention Theory to shape the interaction between the users and the robot. One of the intended purposes of the system is for use in crowded classrooms for selecting students most desirous of interaction and help.**

***Index Terms*— Human-Robot Interaction, Joint Intention Theory**

I. Introduction

Embodiment and mobility are key properties of human-robot interaction (HRI) that distinguish it from human-computer interaction (HCI). These properties must be explored and studied in a directed fashion in order to fully realize the potential of HRI and its relative strengths over non-embodied alternatives (e.g., PDAs) in specific application domains. Our work focuses on the assistive domain, where the robot must satisfy the combined and at times conflicting goals of, on the one hand, engaging the user and, on the other, achieving the needed progress in training/exercise/rehabilitation.

We believe that physical embodiment provides HRI with a unique avenue for success in assistive domains. Our focus is on *socially assistive robotics*, a sub-area of assistive robotics in which robots assist humans in a variety of settings (including hospitals, schools, rehabilitation centers, and homes) through social interaction, instead of through physical contact [9]. Our work focuses on using the physical embodiment of the robot through HRI in order to develop effective non-contact assistance.

In this paper we describe a selective interaction policy for target selection in a mobile HRI system. Our approach is based on the conversational policy of Joint Intention Theory [6]. Rather than strictly using linguistic communication as the basis for a policy, our system uses speech recognition, body language in the form of gestures, and observation and interpretation of the use of space and body language of the users being communicated with.

The motivation for developing the described system comes from the challenge of creating assistive interactive robots capable of naturally approaching only the users who wish to be approached and interacted with. Furthermore, this capability will allow individuals in supervisory roles (e.g., teachers, health care staff) to accurately and naturally guide robots toward desired targets for interaction.

II. Related Work

We briefly summarize work in Joint Intention Theory, HRI, and human activity tracking that relates directly to our described research.

Joint Intention Theory (JIT) describes the nature of communication between two agents as sequences of communicative acts within a conversation policy [6]. Smith et al. [15] describe an agent conversation policy based on JIT; the policy was based on the Cohen-Levesque analysis of joint intentions. The authors determined formal ways for two agents to form a team with a mutual goal. The idea of team formation through the use of joint intentions is the foundation of the research presented in this paper.

Breazeal et al. used the Leonardo robot platform to study JIT [2] with the goal of teaching the robot a sequential button-pushing task. Unlike the work of Smith et al., Leonardo is an embodied system, but not a mobile one. It employed visual detection of point gestures as well as the state of the button panel. It also used a speech recognition system with a limited grammar for accurate recognition.

Kanda et al. [11] used radio frequency ID (RFID) tags to track children in an elementary school using a mobile robot. The robot tracked multiple children at one time, but only actively interacted with them when approached.

The CMU Nursebot project used a mobile robot to help guide elderly residents at a nursing home to their daily obligations [13]. The project coupled human-robot dialog with navigation also used on MINERVA [18] and RHINO [4]. The

robot used a hierarchical POMDP for dialog management in the face of uncertainty in speech recognition [14].

Human activity tracking and understanding is a well-studied problem in machine vision [16], typically not used for on-line real-time applications such as HRI. Laser range finders have also been used for user tracking, with both stationary and mobile robot platforms. Fod et al. [10] used a Kalman Filter to track people from the combination of several overlapping laser scans. In a complementary approach, Yan & Matarić [19] used a particle filter approach to track people and to model multi-robot and human movement in similar indoor environments. Both methods were intended for on-line activity modeling.

Lang et al. [12] implemented a system combining data from a laser range-finder, face recognition, and a microphone array for tracking multiple people using a mobile robot. The robot maintained its focus of attention on one speaker at a time until the speaker yielded to another.

In the work described here, we use proximity data, by combining laser tracking and simple activity modeling, to assess the desire of human subjects to interact with the robot.

III. Approach

Fig. 1. The mobile robot used in the experiments

A. Overview

The system we designed consists of a policy (developed off-line) for determining candidates for interaction, a set of sensors to determine *beliefs* about the people that the robot is attempting to interact with, and a set of behaviors for exhibiting the robot's desires and intentions to the selected targets for interaction. Thus, the robot's beliefs serve as inputs into the selective interaction policy and the behaviors act as the outputs of that policy.

B. Selective Interaction Policy

Joint intention theory provides a rich framework for HRI policies and the foundation for our Selective Interaction Policy. We begin by stating the formal definitions involved in the policy, first developed by Smith et al. [15]. In JIT, teams are formed when an agent x has a *Persistent Weak Achievement Goal* (PWAG) with another agent y to achieve p. By definition, if a PWAG(x, y, p, q) is held, then the following conditions are met:

- p does not hold
- p is currently achievable
- the relativizing condition q still holds

In JIT, q is a relativizing condition for PWAG(x, y, p, q) when x and y can only agree on a goal p if q is true. For our policy, the goal, p, is the interaction between a human user and the robot. The condition q is that the human desires to interact with the robot. To form a team, that is to commit to interaction between the robot and the human, there must exist a PWAG(x, y, p, q) where x is the robot and y is a person, as well as a PWAG(y, x, p, q). The constructed conversation policy must be able to satisfy those conditions in order to establish a proper commitment by both the robot and a human.

An INFORM, written as INF(x, y, e, q), indicates a mutual belief between the listening agent y and the informing agent x that proposition q is true. The action e is an event demonstrating that belief.

SOFFER(x, y, a), a standing offer, is the means of forming a team, as follows:

- x makes a conditional offer to y to do a, revealing that x will have a PWAG toward y to do a;
- y can either INFORM x about his/her intentions toward x or stay silent. Silence implies a refusal to do the proposed task;
- if y confirms to x (using an INFORM) that y will do a, then a team is formed from mutual PWAGs.

For the purposes of our experiment, x is the robot and y is a human. The robot makes a standing offer to a human it observes and wishes to interact with (p). If the person then acknowledges that s/he desires to interact with the robot (q), then a team (consisting of the human and the robot) is formed. If the human stays silent or declines, then the robot assumes that s/he does not desire to interact. As a final alternative, the person can refer the robot to another person on the scene.

For this experiment, p refers to the robot approaching the chosen human target in order to initiate an interaction. However, p could be anything that the robot designer assigns as the goal. If multiple team activities can be accomplished by the human-robot team, the robot and human can negotiate among a variety of possible p's.

C. Sensing

The sensors on the robot are used to determine whether the human (y) is trying to inform the robot (x) about her/his goals. This information falls into one of five categories:

- INFORM-YES: The person wants to interact with the robot. The condition q is true for y.
- INFORM-NO: The human does not want to interact with the robot. The condition q is false for y.
- REFER-LEFT: The person to the left of y wants to interact with the robot. The condition q is false for y, but true for the human z to the left of y.
- REFER-RIGHT: The person to the right of y wants to interact with the robot. The condition q is false for y, but true for the human z to the right of y.
- SILENCE: No reaction from y. This is an important condition, since, due to perceptual uncertainty, objects in the environment and other robots could potentially be confused for people. No reaction over time makes the robot move on to a more responsive target.

Each of the following sensors can classify the observed behavior of y into one of the above categories, thereby producing the belief of x based on the perceived state of y.

1) Laser-based person tracking: We used laser-based leg finding as a robust means of detecting people. The leg tracker looks for appropriately-sized leg-shaped occlusions in the laser scan. When legs are identified, they are grouped into pairs that are classified as a fiducial. The fiducial is then translated from robot-centric coordinates to world coordinates for ease of tracking while the robot and the target are both moving.

Detected leg fiducials are tracked continually. Any newly found fiducials are added to the tracker and correlated with existing ones. We have validated tracking of up to four people. A reasonable moving speed ($<$ 1.5 meters/second) is used as a heuristic to eliminate false positives.

Based on the fiducial tracking, if a person is seen to be moving closer to the robot, that action is interpreted as INFORM-YES. This classification is based on the assumption that a person moves closer to the robot to indicate that s/he wants to interact with it.

Analogously, if a person is observed to be moving away from the robot, that action is interpreted as an INFORM-NO, based on the assumption that moving away is a sign of avoiding interacting with the robot. Finally, if the person remains relatively still (heuristically set at moving $<$ 50 cms/second), that lack of action is interpreted as SILENCE.

The laser range finder provides no means for the robot to sense gestures or other user indications of interacting with another person, i.e., one to the left or right of the user. While the laser is not useful for the PERSON-LEFT and PERSON-RIGHT conditions, it was the only sensor available that could attend to more than one target at once, given its 180-degree field of view. This capability is critical for interaction target selection, the goal of our HRI system.

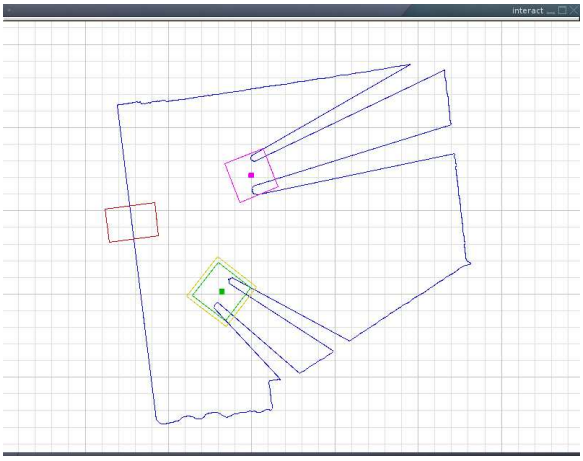

Fig. 2. People (boxes) identified from a laser scan (blue line)

2) Speech Recognition: In our experiment, speech recognition was less reliable than laser-based sensing. We used Carnegie Mellon University's Sphinx system for speech recognition with a Shure 503-BG microphone for audio recording. The microphone was designed for speech recognition in noisy environments; it has a frequency response that favors human speech and is directional so that, if aimed correctly, it effectively picks up the targeted person's speech and little else. The accuracy of speech recognition in our work was offset by the limited ability to listen to multiple users: the microphone worked on one person at a time.

The robot was programmed to aim the microphone in the direction of the targeted user (the closest one, or, if s/he refused the interaction, then the next closes, and so on). When a speech utterance was detected, it was decoded (by Sphinx) into a sentence, and parsed for keywords (from a limited vocabulary shown in Table I) that indicate the user's intentions. When no utterance was detected, the SILENCE condition was selected, which is interpreted to mean refusal to interact with the robot. This choice is effective in the context of our experiment, but is not general, since silence could have other meanings in different domains.

3) Gesture Recognition: To simplify gesture recognition, we used a brightly colored square that could be easily detected and tracked by the robot's Sony PTZ camera. Bright pink color was effective as it was not easily confused with other colors normally present in indoor environments, such as various skin tones and typical wall paint colors. If the system were deployed in a classroom, the teacher could use such an object as a "control wand" for interactive control of the robot.

We used the ActivMedia CTS color-blob finder tuned to bright pink to find the control wand in the image. The

Words	Meaning
LEFT TURN LEFT	REFER-LEFT
RIGHT TURN RIGHT	REFER-RIGHT
COME HERE MOVE FORWARD I LIKE YOU YES	INFORM-YES
GO AWAY MOVE BACKWARD TURN AROUND BUZZ OFF NO $<< SILENCE >>$	INFORM-NO

TABLE I
VOCABULARY AND INFERRED MEANINGS

blobs were found, tracked, and their direction determined and classified in four general categories: UP, DOWN, LEFT, and RIGHT. SILENCE was selected if no blobs were detected, or if the blob was not moving enough (empirically set at 0.2 m/s) to be considered a gesture.

The position of the blob in the image was tracked over time and an average vector of the past 10 frames of movement was calculated, to reduce tracking noise. If the direction of the average vector was within 45^o of the horizontal, the gesture was selected as LEFT or RIGHT. If the direction of the average vector was within 45^o of the vertical, the gesture was either UP or DOWN.

RIGHT and LEFT motions were mapped to REFER-RIGHT and REFER-LEFT, while DOWN and UP motions were mapped to INFORM-YES and INFORM-NO, respectively.

D. Expressive Behaviors

A simple HRI system may only employ the above perceptual capabilities in order to interact with the user and select targets. However, for the more sophisticated policy described above, which involves human-robot team formation, the robot must indicate to potential targets that it is interested in interacting with them. While this may not seem necessary in one-on-one HRI, it is imperative in crowded situations such as classrooms. Thus, we employed artificially generated speech and active vision to provide the user with a notion of the robot's intentions.

1) Active Vision: It is crucial for the targeted person to know that the robot is speaking to him/her rather than any other person present in the vicinity. To that end, we employed a primitive active vision system; we used the bearing of the targeted person as obtained from the laser tracker to direct the camera and the microphone mounted on on top of it.

We also used the range provided by the laser to properly set the tilt and zoom of the camera so that gesture recognition could be performed effectively even at longer distances, between 2.5m and 3.5m.

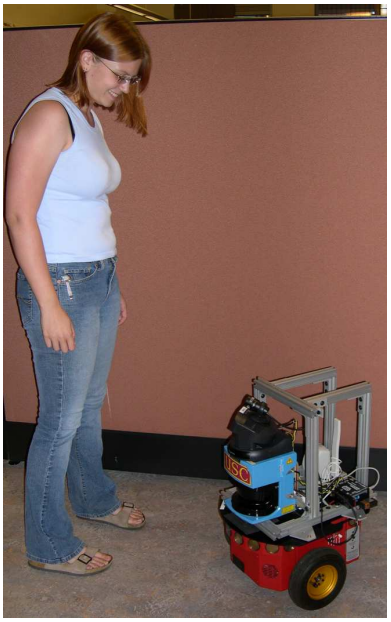

Fig. 3. Robot engaging user by aiming camera and microphone at target

The ability to appropriately adjust the camera's pan, tilt, and zoom toward the user is highly effective in indicating to the person being targeted that the robot is selectively paying attention to him/her [1]. More practically, it also allowed the system to train the camera and microphone on the person for continued tracking.

2) Speech Synthesis: Speech synthesis is important because it is effective in encouraging the human user to speak to the robot and interact with it. For this to be so, the robot's speech must be understandable. We found that traditional artificial speech solutions, such as Festival [5], can be hard for users to understand. In contrast, a pre-recorded human voice has been effective in assistive and interactive robot domains [8]. Some suggest, however, that in order to provide accurate expectations of a robot's capabilities, the robot should sound more mechanical than human [7].

Thus, we used the AT&T text-to-speech system [3] for speech expression, which translates text to a list of phonemes and creates a .wav file by piecing together recorded phonemes in the correct order. The result was an understandable while still mechanical-sounding voice.

3) Robot Movement: Using the robot's capability of movement, "body language", and use of shared space are key yet under-explored components of HRI that separate it from HCI. In our system, the robot communicated its intentions through movement as well as speech. The robot turned toward a targeted person to show interest, it moved toward the person to show commitment to an interaction, it approached the

human when mutual interest in interaction was agreed upon and thus a human-robot team formed.

IV. System Performance

We first describe the performance of each sensor modality involved in HRI individually, then evaluate the system as a whole.

A. Laser Leg Tracking

We evaluated the leg finding and tracking algorithm based on its ability to recognize and track people present in the environment. It performed highly reliably when the people in the scene were more than 1m apart. It was able to correctly identify people, and track their positions 95% of the tracking time. When people became too close to each other, it was too difficult to distinguish between them. As a result, in the majority of our trials, we kept the participants at least 1 meter apart.

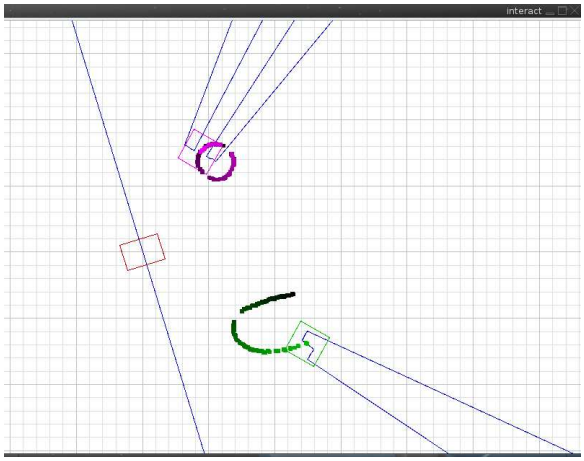

Fig. 4. Tracking of people (boxes) and their histories (trail).

B. Speech Recognition

We tested the speech recognition system by repeating several phrases at varying distances. This sensory modality had the worst performance in our system. It performed well when the microphone was placed near the speaker's mouth (within the range of 0.5-1.5 meters), but since the task also involved interacting with users that were not very close to the robot, those situations resulted in speech recognition system failures, i.e., the inability to recognize the words from the vocabulary (shown in Table I). Quantitatively, we found that the recognition rate was approximately 82% at close ranges (as per above); it dropped to approximately 62.5% at distances over 3 meters.

We did not have human users significantly raise their voices when interacting with the robot and thus have no data on the effect of that strategy for dealing with speech recognition limitations at a distance.

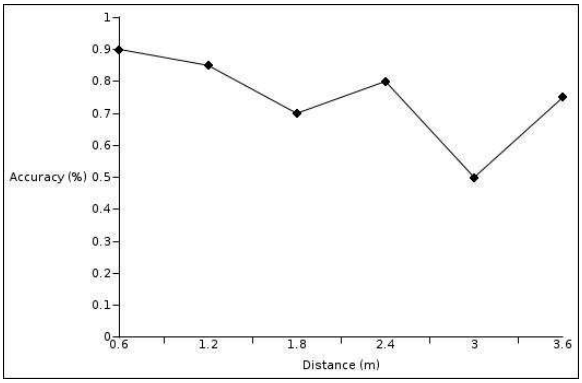

Fig. 5. Speech recognition accuracy vs. distance

It should be noted that Lang et al. [12] successfully used sound-source localization for calling a robot in a crowded room. This suggests that while accurate speech processing at a distance may not be effective, sound-source localization is very promising in this type of HRI context.

C. Gesture Recognition

Gesture recognition alone was as reliable as speech recognition. Because of the camera's ability to zoom, user distance was not a significant issue. In a repeated test of gestures, the system averaged to correctly recognize 32 out of 40 gestures (80%). At ranges beyond 2.5m, the recognition rate dropped by 20%.

D. Evaluation of the Selective Interaction Policy

To evaluate the system as a whole, we assessed how frequently the robot selected the most appropriate action when faced with varying situations. The most appropriate action was defined as approaching the person who took an active interest in interacting with the robot. We evaluated the system by having a person attempt to "guide" the robot to the most willing target, if any. The people involved in the experiment agreed *a priori* who would be willing to interact and who would not, and steered the robot to the willing individual through the use of gestures and speech. In some cases there was no willing individual.

When the robot was tested with only one person in the scene, it selected the correct interaction behavior 80% of the time (8 out of 10 trials). The correct behavior refers to approaching the appropriate (willing to interact) person and not approaching an inappropriate (unwilling to interact) one. The trials were divided into blocks, half of which involved a user willing to interact, and half of which involved no such willing user.

When the robot was tested with multiple people in the environment, the accuracy was reduced to approximately 67% (20 out of 30 trials). Most of the mistakes (60%) were due to speech recognition errors. Additionally, in four of the

trials (13% of the time, 25% of the errors), the robot's person tracker confused two of the people in the scene, resulting in tracking or approaching the wrong person.

We also assessed how often the robot selected people vs. false positives (i.e., objects in the environment). Since objects never gave any indication of a desire to interact (no movement, gesture, or speech), the robot never approached them.

V. Future Work

Pragmatic improvements to the system involve the use of a more robust tracking algorithm and an alternative to speech recognition at a distance. For the former, algorithms such as [10], [19], [17] are suitable candidates. The ability to track robustly will only grow in importance in real-world settings, such as classrooms where children are likely to swarm around the robot rather than approach it gradually and in an organized fashion. For the latter, since speech recognition at a distance is an open problem, sound-source localization [12] is likely to be more effective for the assistive interactive robotics applications we are interested in.

In [9], socially assistive robotics was discussed as human-robot interaction with the goal of assisting people. Such goal-driven human-robot interaction requires a strong model to support it. Joint Intention Theory can be used to model other tasks more complex than the one described in this paper. Since the theory directly addressed interaction between two agents for the sake of forming a team to achieve a goal [15], it is very well suited for use in socially assistive robotics.

For example, Breazeal [2] used joint intention theory to form a policy for interaction between a robot and a teacher trying to show the robot how to manipulate a device correctly. The roles of teacher and student could be reversed so that the robot is teaching a person. An educational robot would be of great use in a classroom, as well as in skill (re)learning post-stroke.

VI. Conclusion

We have presented an HRI system that uses multiple sensory modalities to determine if a user is interested in interacting with the robot. We made use of a selective interaction policy to approach targets that are believed, by the robot, to be most appropriate for interaction.

Since the key distinguishing feature of robots from non-embodied devices is their physical mobility, the use of mobility and embodiment as explicit tools for interaction validates key principles of HRI.

VII. Acknowledgments

This work was supported by the USC Provost's Center for Interdisciplinary Research, the Annenberg Foundation, and the Okawa Foundation. We are grateful to Andrew Howard and Nate Koenig for their help in developing the leg finder used in our experiments.

References

[1] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision for sociable robots. *IEEE Transactions on Man, Cybernetics and Systems*, 31(5), September 2001.

[2] C. Breazeal, G. Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 3, pages 1030–1037, New York, NY, July 2004.

[3] M. Bulut, S. S. Narayanan, and A. K. Syrdal. Expressive speech synthesis using a concatenative synthesizer. In *International Conference on Spoken Language Processing*, Denver, CO, September 2002.

[4] W. Burgard, A. Cremers, D. Fox, D. HŁhnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. Technical Report CMU-CS-98-139, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 1998.

[5] R. A.J. Clark, K. Richmond, and S. King. Festival 2 - build your own general purpose unit selection speech synthesiser. In *ISCA workshop on speech synthesis*, pages 1047–1056, Pittsburgh, PA, June 2004.

[6] P. R. Cohen and H. J. Levesque. *Intentions in Communication*, chapter Rational Interaction as the basis for communication, pages 221–256. System Development Foundation Benchmark Series. MIT Press, Cambridge, MA, 1990.

[7] B. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3):177–190, March 2003.

[8] J. Eriksson, M. Matarić, and C. Winstein. Hands-off assistive robotics for post-stroke arm rehabilitation. In *Proceedings of the International Conference on Rehabilitation Robotics*, Chicago, Il, Jun-Jul 2005.

[9] David Feil-Seifer and Maja J Matarić. Defi ning socially assistive robotics. In *Proceedings of the International Conference on Rehabilitation Robotics*, Chicago, Il, Jul 2005.

[10] A. Fod, A. Howard, and M. Matarić. Laser-based people tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3024–3029, Washington, DC, USA, May 2002.

[11] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Person identifi cation and interaction of social robots by using wireless tags. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2003)*, pages 1657–1664, Las Vegas, NV, October 2003.

[12] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–35, Vancouver, Canada, November 2003. ACM.

[13] M. Montemerlo, J. Prieau, S. Thrun, and V. Varma. Experiences with a mobile robotics guide for the elderly. In *Procedings of the AAAI National Conference on Artificial Intelligence*, pages 587–592, Edmunton, Alberta, August 2002.

[14] Nicholas Roy, Joelle Pineau, and Sebastian Thrun. Spoken dialog management for robots. In *Proceedings of the Association for Computational Linguistics*, Hong Kong, Japan, October 2000.

[15] I. A. Smith, P. R. Cohen, J. M. Bradshaw, M. Greaves, and H. Holmback. Designing conversation policies using joint intention theory. In *proceedings of the Conference on Multi Agent Systems*, pages 269–276, Washington, DC, USA, July 1998.

[16] Y Song, L Goncalves, and P Perona. Unsupervised learning of human motion models. In *Proceedings of the Neural Information Processing Systems Conference*, pages 814–827, Vancouver, British Columbia, Canada, December 2001.

[17] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Brooks/Cole Publishing Company, Pacifi c Grove, CA, 2nd edition, 1999.

[18] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. MINERVA: A second-generation museum tour-guide robot. In *Proceedings: IEEE International Conference on Robotics and Automation (ICRA '99)*, pages 1999–2005, Detroit, Michigan, May 1999.

[19] H. Yan and M. J. Matarić. General spatial features for analysis of multi-robot and human activities from raw position data. In *Proceedings of the IEEE International Conference on Robotics and Systems*, pages 2770–2775, EPFL, Switzerland, Sep 30 - Oct 4 2002.