Towards a Humanoid Museum Guide Robot that Interacts with Multiple Persons

Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke

University of Freiburg
Computer Science Institute
D-79110 Freiburg, Germany
{*maren, faber, joho, schreibe, behnke*}*@informatik.uni-freiburg.de*

***Abstract*— The purpose of our research is to develop a humanoid museum guide robot that performs intuitive, multimodal interaction with multiple persons. In this paper, we present a robotic system that makes use of visual perception, sound source localization, and speech recognition to detect, track, and involve multiple persons into interaction. Depending on the audio-visual input, our robot shifts its attention between different persons. In order to direct the attention of its communication partners towards exhibits, our robot performs gestures with its eyes and arms. As we demonstrate in practical experiments, our robot is able to interact with multiple persons in a multimodal way and to shift its attention between different people. Furthermore, we discuss experiences made during a two-day public demonstration of our robot.**

I. Introduction

Our goal is to develop a museum guide robot that acts human-like. The robot should perform intuitive, multimodal interaction, i.e., it should use speech, eye-gaze, and gestures to converse with the visitors. Furthermore, the robot should be able to distinguish between different persons and to interact with multiple persons simultaneously. Compared to previous museum tour-guide projects [17], [24], [28], which mainly focused on the autonomy of the (non-humanoid) robots and did not emphasize the interaction part so much, we want to build a robot that behaves human-like during the interaction.

Much research has already been conducted in the area of non-verbal communication between a robot and a human, such as facial expression, eye-gaze, and gesture commands [4], [9], [20], [25], [29]. However, only little research has been done in the area of developing a robotic system that is able to interact with *multiple* persons appropriately. This was also stated by Thrun [27] as one of the open questions in the field of human-robot interaction.

In contrast to previous approaches to human-robot interaction using multimodal sensing [8], [12], [19], our goal is that the robot involves multiple persons into interaction and does not focus its attention on only one single person. It should neither simply look to the person who is currently speaking. Depending on the input of the audio-visual sensors, our robot shifts its attention between different people.

In order to direct the attention of the visitors towards the exhibits, our robot performs gestures with its eyes and arms. To make the interaction even more human-like, we use a head with an animated mouth and eyebrows and show facial expressions corresponding to the robot's mood. As a result, the users get feedback how the robot is affected by the different external events. This is important because expressing emotions helps to indicate the robot's state or its intention. Figure 1 shows our robot Alpha interacting with people during a two-day demonstration in public.

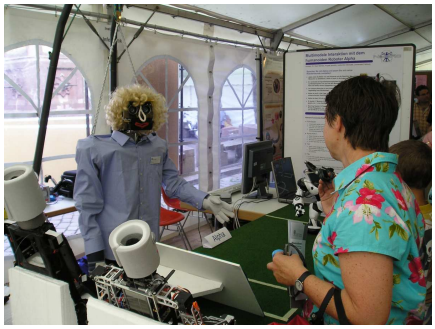

Fig. 1. Our robot Alpha interacting with people during a public demonstration.

This paper is organized as follows. The next section gives an overview over related work, and Section III introduces the hardware of our robot. In Section IV, we present our technique to detect and keep track of people using vision data and a speaker localization system. In Section V, we explain our strategy on how to determine the gaze direction of the robot and how to decide which person gets its attention. In Section VI, we describe the pointing gestures our robot performs, and in Section VII, we illustrate how the robot changes its facial expression depending on external events. Finally, in Section VIII, we show experimental results and discuss the experiences we made during the two-day demonstration of our robot in public.

II. Related Work

Over the last few years, much research has been carried out in the area of multimodal interaction. Several systems exist that use different types of perception to sense and track people during an interaction and that use a strategy to decide which person gets the attention of the robot.

Lang et al. [8] apply an attention system in which only the person that is currently speaking is the person of interest. While the robot is focusing on this person, it does not look to another person to involve it into the conversation. Only if the speaking person stops talking for more than two seconds, the

0-7803-9320-1/05/$20.00 ©2005 IEEE

robot will show attention to another person. Okuno et al. [19] also follow the strategy to focus the attention on the person who is speaking. They apply two different modes. In the first mode, the robot always turns to a new speaker, and in the second mode, the robot keeps its attention exclusively on one conversational partner. The system developed by Matsusaka et al. [12] is able to determine the one who is being addressed to in the conversation. Compared to our application scenario (museum guide), in which the robot is assumed to be the main speaker or actively involved in a conversation, in their scenario the robot acts as an observer. It looks at the person who is speaking and decides when to contribute to a conversation between two people.

The model developed by Thorisson [26] focuses on turn-taking in one-to-one conversations. This model has been applied to a virtual character. Since we focus on how to decide which person in the surroundings of the robot gets its focus of attention, a combination of both techniques is possible. Kopp and Wachsmuth [6] developed a virtual conversational agent which uses coordinated speech and gestures to interact with humans in a multimodal way.

In the following, we summarize the approaches to human-like interaction behavior of previous museum tour-guide projects. Bischoff and Graefe [3] presented a robotic system with a humanoid torso that is able to interact with people using its arms. This robot also acted as a museum tour-guide. However, the robot does not distinguish between different persons and does not have an animated face. Several (non-humanoid) museum tour-guide robots that make use of facial expressions to show emotions have already been developed. Schulte et al. [22] used four basic moods for a museum tour-guide robot to show the robot's emotional state during traveling. They defined a simple finite state machine to switch between the different moods, depending on how long people were blocking the robot's way. Their aim was to enhance the robot's believability during navigation in order to achieve the intended goals. Similarly, Nourbakhsh et al. [16] designed a fuzzy state machine with five moods for a robotic tour-guide. Transitions in this state machine occur depending on external events, like people standing in the robot's way. Their intention was to achieve a better interaction between the users and the robot. Mayor et al. [13] used a face with two eyes, eyelids and eyebrows (but no mouth) to express the robot's mood using seven basic expressions. The robot's internal state is affected by several events during a tour (e.g., a blocked path or no interest in the robot).

Most of the existing approaches do not allow continuous changes in the facial expression. Our approach, in contrast, uses a bilinear interpolation technique in a two-dimensional state space [21] to smoothly change the robot's facial expression.

III. The Design of our Robot

The body (without the head) of our robot Alpha has currently 17 degrees of freedom (four in each leg, three in each arm, and three in the trunk; see left image of Figure 2). The

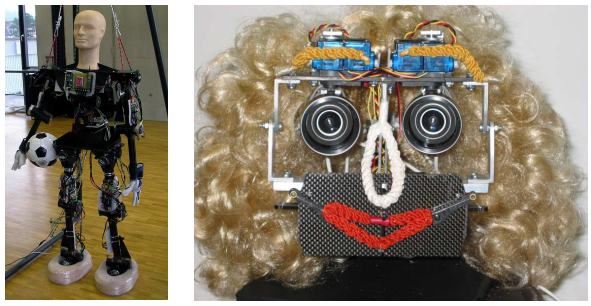

Fig. 2. The left image shows the body of our robot Alpha. The image on the right depicts the head of Alpha in a happy mood.

joints of the robot are driven by Faulhaber DC-motors of different sizes. The robot's total height is about $155cm$. The skeleton of the robot is constructed from carbon composite materials to achieve a low weight of about $30kg$. The head (see right image of Figure 2) consists of 16 degrees of freedom, which are driven by servo motors. Three of these servos move two cameras and allow a combined movement in the vertical and an independent movement in the horizontal direction. Furthermore, three servos constitute the neck joint and move the entire head, six servos animate the mouth, and four the eyebrows. Using such a design, we can control the neck and the cameras to perform rapid saccades, which are quick jumps, or slow, smooth pursuit movements (to keep eye-contact with a user). We take into account the estimated distance to a target in order to compute eye vergence movements. These vergence movements ensure that the target remains in the center of the focus of both cameras. Thus, if a target comes closer, the eyes are turned toward each other (see also [4]).

The cameras are one of the main sensors to obtain information about the surroundings of the robot. Furthermore, we use the stereo signal of two microphones to perform speech recognition as well as sound source localization.

For the behavior control of our robot, we use a framework developed by Behnke and Rojas [1] that supports a hierarchy of reactive behaviors. In this framework, behaviors are arranged in layers that work on different time scales.

IV. Keeping Track of People

To sense people in the environment of our robot, we use the data delivered by the two cameras and the information of our speaker localization system. In order to keep track of persons even when they are temporarily outside the robot's field of view, the robot maintains a probabilistic belief about the people in its surroundings.

A. Visual Detection and Tracking of People

Figure 3 illustrates how the update of the robot's belief works. To find people in the current pair of images, we first run a face detector. Then, we apply a mechanism to associate the detections to faces already stored in the belief and finally, we update the belief according to the new observations. In the following, we explain the individual steps in more detail.

Our face detection system is based on the AdaBoost algorithm and uses a boosted cascade of Haar-like features [10].

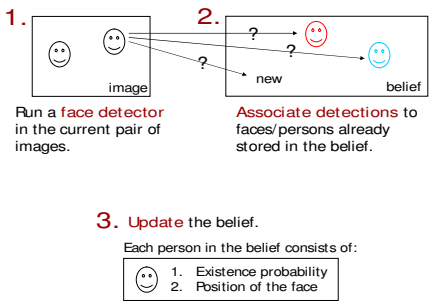

Fig. 3. The three steps carried out to update the belief of the robot about the people in its surroundings based on vision data.

Each feature is computed by the sum of all pixels in rectangular regions, which can be computed very efficiently using integral images. The idea is to detect the relative darkness between different regions like the region of the eyes and the cheeks. Originally, this idea was developed by Viola and Jones [30] to reliably detect faces without requiring a skin color model. This method works quickly and yields high detection rates.

After the face detection process, we must determine which detected face in the current images belongs to which person that already exist in the belief and which face belongs to a new face. To solve this data association problem, we apply the Hungarian Method [7]. The Hungarian Method is a general method to determine the optimal assignment of jobs to machines, using a given cost function in the context of job-shop scheduling problems. Since we currently do not have a mechanism to identify people, we use a distance-based cost function to determine the mapping from current observations to faces already existing in the belief.

To deal with false classifications of face/non-face regions and association failures, we apply a probabilistic technique. We use a recursive Bayesian update scheme [14] to compute the existence probability of a face (details can be found in [2]). In this way, the robot can also keep track of the probability that a person outside the current field of view is still there.

Figure 4 shows three snapshots during face tracking. As indicated by the differently colored boxes, all faces are tracked correctly.

B. Speaker Localization

Additionally, we implemented a system to localize a speaker in the environment. We apply the Cross-Power Spectrum Phase Analysis [5] to calculate the spectral correlation measure between the left and the right microphone channel. By doing so, we can determine the delay between the left and the right channel. As we can use this delay, the relative angle between a speaker and the microphones can be calculated under two assumptions [8]: 1. The speaker and the microphones are at the same height, and 2. the distance of the speaker to the microphones is larger than the distance between the microphones themselves.

We assign the information that the person has spoken to the person in the robot's belief that has the minimum distance to the sound source. If the angular distance between the speaker and the person is greater than a certain threshold, we assume the speaker to be a new person, who just entered the scene.

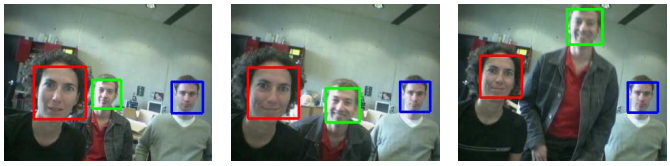

Fig. 4. Tracking three faces.

V. Gaze Control and Focus of Attention

For each person in the belief, we compute an importance value. This importance value triggers the focus of attention of the robot. It currently depends on the time when the person has last spoken, on the distance of the person to the robot (estimated using the size of the bounding box of its face), and on its position relative to the front of the robot. People who have recently spoken get a higher importance than others. The same applies to people who stand directly in front of the robot and to people who are close to the robot. The resulting importance value is a weighted sum of these three factors.

The robot focuses its attention always on the person who has the highest importance, which means that it keeps eye-contact with this person. If at some point in time another person is considered to be more important than the previously most important one, the robot shifts its attention to the other person. For example, this can be the case when a person steps closer to the robot or when a person starts speaking. Note that one can also consider further information to determine the importance of a person. If our robot, for example, could detect that a person is waving with his/her hands to get the robot's attention, this could easily be integrated as well.

If a person that is outside the current field of view and not stored in the belief so far starts to speak, the robot reacts to this by turning towards the corresponding direction. In this way, the robot shows attentiveness and is able to update its belief.

Since the field of view of the robot is constrained (it is approximately 90 degrees), it is important that the cameras move from time in order to time to explore the environment so that the robot is able to update its belief about surrounding people. Thus, the robot regularly changes its gaze direction and looks in the direction of other faces, not only to the most important one. Our idea is that the robot shows interest in multiple persons in its vicinity so that they feel involved into the conversation. Like humans, our robot does not stare at one conversational partner all the time.

VI. Pointing Gestures

As already investigated by Sidner et al. [23] who used a robotic penguin, humans tend to be more engaged in an interaction when a robot uses gestures to refer to objects of interest. The attention of the communication partners is drawn towards the objects the robot is pointing to. Thus, we let the robot perform pointing gestures to an exhibit when it starts to present one. In this way, the visitors are attracted more,

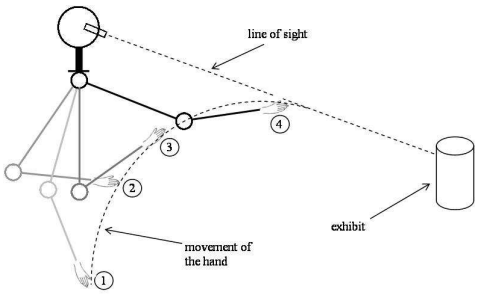

Fig. 5. Side view of the arm movement during a pointing gesture.

follow the gaze direction of the robot, and are able to easily infer which of the exhibits (if there are several nearby) is the one of interest.

While analyzing arm gestures performed by humans, Nickel et al. [15] found out that most people use the line of sight between head and hand when pointing to an object. Compared to this line of sight, the direction of the forearm was not so expressive. People usually move the arm in such a way that in the hold phase, the hand is in one line with the head and the object of interest. We use this result to compute the position of the hand of our robot during the hold phase.

Our robot has arms with three degrees of freedom: two in the shoulder and one in the elbow. To specify an arm movement, we use the *x* (left and right) and *y* (back and forth) direction of the shoulder joint and an abstract parameter that specifies the arm extension. The arm extension is a value which specifies the distance between hand and shoulder relative to the maximum possible distance when the arm is outstretched. Using this extension value, the position of the elbow joint is computed. The *x* component of the shoulder joint accepts values between $-12°$ and $52°$ and the *y* component values between $-38°$ and $66°$.

When the robot starts to explain an exhibit, it simultaneously moves the head and the eyes in the direction of the exhibit, and it points in the direction with the corresponding arm. We first compute the point where the (almost) outstretched arm would meet the line of sight. This is the point where the robot's hand rests during the hold phase. Figure 5 illustrates the movement of the arm during a gesture. To model the arm gesture, we use an individual sine curve for each joint. We optimized the movement so that it appears human-like.

Figure 6 (from (a) to (d)) shows an example scenario from the visitor's perspective. Initially, the robot and the person were looking at each other while talking. Then, the person asked the robot to present an exhibit. Thus, the robot started to explain the exhibit and simultaneously looked in the direction of the corresponding object. Immediately afterwards, it started the arm gesture.

VII. Facial Expressions

Showing emotions plays an important role in inter-human communication because, for example, the recognition of the mood of a conversational partner helps to understand his/her behavior and intention. Thus, to make the interaction more human-like, we use a face with animated mouth and eyebrows

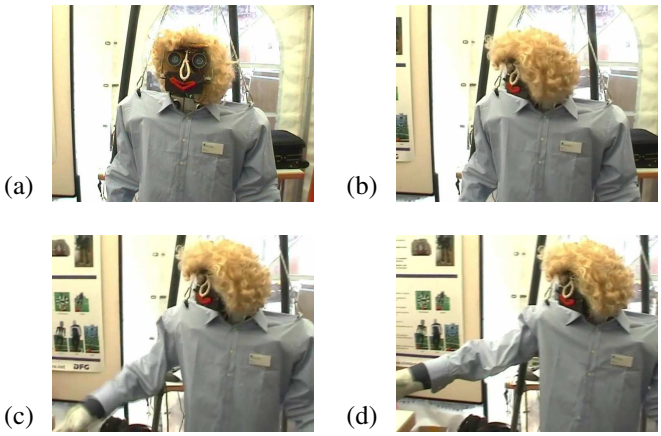

Fig. 6. Alpha performing a pointing gesture (from (a) to (d)). Initially, the robot faces the person. Then, it looks in the direction of the exhibit and starts the arm gesture.

to display facial expressions corresponding to the robot's mood. As a result, the users get feedback how the robot is affected by the different external events.

The robot's facial expression is computed in a two-dimensional space, using six basic emotional expressions (joy, surprise, fear, sadness, anger, and disgust). Here, we follow the notion of the Emotion Disc developed by Ruttkay et al. [21]. The design of the Emotion Disc is based on the observation that the six basic emotional expressions can be arranged on the perimeter of a circle (see Figure 7), with the neutral expression in the center. The Emotion Disc can be used to control the expression of any facial model once the neutral and the six basic expressions are designed. Figure 7 shows the six basic facial expressions of our robot.

The parameters P' for the face corresponding to a certain point P in the two-dimensional space are calculated by linear interpolation between the parameters E'_i and E'_{i+1} of the neighboring basic expressions:

$$P' = l(p) \cdot (\alpha(p) \cdot E'_i + (1 - \alpha(p)) \cdot E'_{i+1}). \quad (1)$$

Here, $l(p)$ is the length of the vector p that leads from the origin (corresponding to the neutral expression) to P, and $\alpha(p)$ denotes the normalized angular distance between p and the vectors corresponding to the two neighboring basic expressions. This technique allows continuous changes of the facial expression.

To influence the emotional state of our robot, we use behaviors that react to certain events. For example, if no one is interested in the robot, it is getting more and more sad, if someone then talks to it, the robot's mood changes to a mixture of surprise and happiness. Each behavior submits its request in which direction and with which intensity it wants to change the robot's emotional state. After all behaviors submitted their requests, the resulting vector is computed by the sum of the individual requests. We allow any movement within the circle described by the Emotion Disc.

VIII. Experimental Results

To evaluate our approach to control the gaze direction of the robot and to determine the person who gets the focus

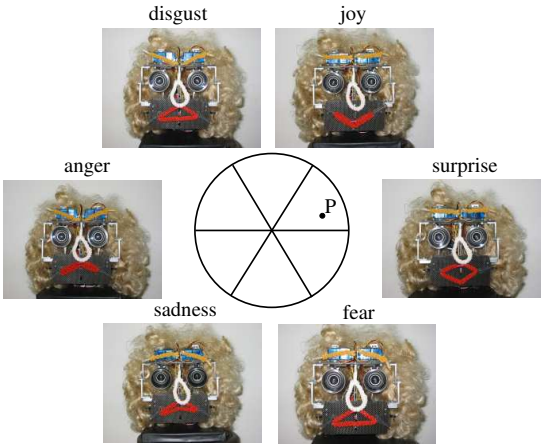

Fig. 7. The two-dimensional space in which we compute the robot's facial expression.

of its attention, we performed several experiments in our laboratory. One of them is presented here. Furthermore, we report experiences we made during public demonstration.

A. Shifting Attention

This experiment was designed to show how the robot shifts its attention from one person to another if it considers the second one to be more important. In the situation considered here, two persons were in the surroundings of Alpha. Person 1 was only listening and person 2 was talking to the robot. Thus, the robot initially focused its attention on person 2 since it had the highest importance. The images from (a) to (d) in Figure 8 illustrate the setup of this experiment and show how the robot changes its gaze direction. The lower image in Figure 8 shows the evolution of the importance value of the two persons. At time steps 10 and 21, the robot looked to person 1 to signal awareness and to involve him/her into the conversation. When looking to person 1 at time step 21, the robot suddenly noticed that this person had come very close. Accordingly, person 1 got a higher importance value, and the robot shifted its attention to this person. As this experiment demonstrates, our robot does not focus its attention exclusively on the person that is speaking. Further experimental results are presented in [2]. We provide videos of our robot Alpha on our webpage[1].

B. Presenting Alpha to the Public

During a two-day science fair of Freiburg University in June 2005, we exhibited our robot. Alpha had simple conversations with the people and presented its robotic friends. Figure 9 shows Alpha in action. For speech recognition, we currently use a commercial software (GPMSC developed by Novotech [18]) and for speech synthesis, the Loquendo TTS software [11], which is also commercial. Our dialogue system is realized as a finite state machine (see [2] for details).

We asked several people who interacted with the robot to fill out questionnaires to get feedback. Almost all people found the eye-gazes, gestures, and the facial expression human-like and felt that Alpha was aware of them. The people were

[1] http://www.nimbro.net/media.html

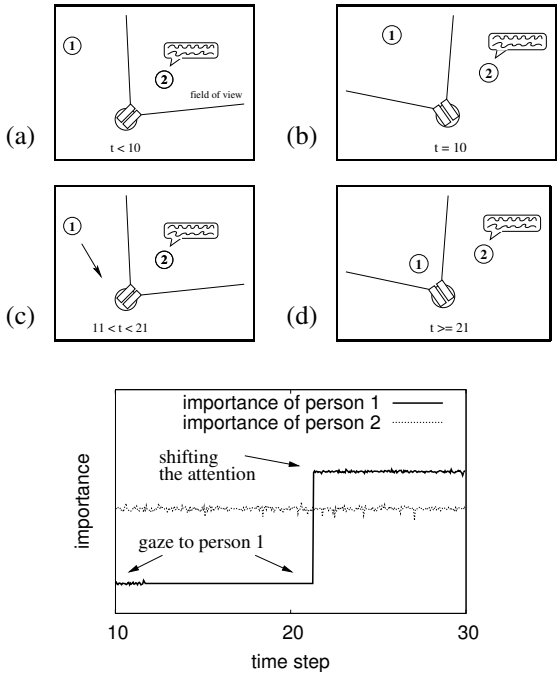

Fig. 8. The images (a) to (d) illustrate the setup in this experiment. The lower image shows the evolution of the importance values of two people. During this experiment, person 2 is talking to the robot. Thus, it has initially a higher importance than person 1. The robot focuses its attention on person 2 but also looks to person 1 at time steps 10 and 21 to demonstrate that it is aware of person 1. At time step 21 the robot notices that person 1 has come very close and thus it shifts its attention to person 1, which has a higher importance now.

mostly attracted and impressed by the vivid human-like eye movements. Most of the people interacted with the robot for more than three minutes. This is a good result because it was rather crowded around our stand. Some toddlers were afraid of Alpha and hid behind their parents. Apparently, they were not sure what an creature the robot is.

One limitation of our current system is that the speech recognition does not work sufficiently well in extremely noisy environments. In the exhibition hall, even the humans had to talk rather loud to understand each other. Thus, the visitors had to use close-talking microphones in order to talk to the robot. Obviously, there were several recognition failures.

To evaluate the expressiveness of the gestures, we performed an experiment in which we asked the people (which were not familiar with robots) to guess which exhibit Alpha was pointing to. In this experiment, Alpha randomly pointed to one of the robots. We had two robots exhibited on each side of a table and, as can be seen from Figure 9, the robots on the same side were sitting quite close to each other. 91% of the gestures were correctly interpreted. Each subject guessed the target of four pointing gestures. One interesting observation was that the people automatically looked into the robot's eyes in order to determine the object of interest. Thus, they noticed that the arm was not the only source of directional information. Another observation was that the people did not verbalize the names of the referenced robots (they were clearly marked), instead they adopted a pointing behavior as well. Further experiments in our laboratory with the aim to evaluate the dereferencability of pointing gestures yielded similar results.

Fig. 9. Alpha presenting its friends.

IX. Conclusions

In this paper, we presented an approach to enable a humanoid robot to interact with multiple persons in a multimodal way. Using visual perception and sound source localization, the robot applies an intelligent strategy to change its focus of attention. In this way, it can attract multiple persons and include them into an interaction. In order to direct the attention of its communication partners towards objects of interest, our robot performs pointing gestures with its eyes and arms. To express the robot's approval or disapproval to external events, we use a technique to change its facial expression.

In practical experiments, we demonstrated our technique to control the robot's gaze direction and to determine the person who gets its attention. Furthermore, we discussed the experiences we made during a public demonstration of our robot.

Acknowledgment

This project is supported by the DFG (Deutsche Forschungsgemeinschaft), grant BE 2556/2-1.

References

[1] S. Behnke and R. Rojas. A hierarchy of reactive behaviors handles complexity. In M. Hannebauer, J. Wendler, and E. Pagello, editors, *Balancing Reactivity and Social Deliberation in Multi-Agent Systems*, pages 125–136. Springer Verlag, 2001.

[2] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Multimodal conversation between a humanoid robot and multiple persons. In *Proc. of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI)*, 2005.

[3] R. Bischoff and V. Graefe. Dependable multimodal communication and interaction with robotic assistants. In *Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002.

[4] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision systems for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31(5):443–453, 2001.

[5] D. Giuliani, M. Omologo, and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis. In *Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1243–1246, 1994.

[6] S. Kopp and I. Wachsmuth. Model-based animation of coverbal gesture. In *Proc. of Computer Animation*, 2002.

[7] H.W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1):83–97, 1955.

[8] 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 *Proc. of the Int. Confonference on Multimodal Interfaces*, 2003.

[9] S. Li, M. Kleinehagenbrock, J. Fritsch, B. Wrede, and G. Sagerer. "BIRON, let me show you something": Evaluating the interaction with a robot companion. In *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 2004.

[10] R. Lienhard and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2002.

[11] Loquendo. Loquendo Text-to-Speech (TTS). http://www.loquendo.com/en/technology/TTS.htm, 2005.

[12] Y. Matsusaka, S. Fujie, and T. Kobayashi. Modeling of conversational strategy for the robot participating in the group conversation. In *Proc. of the European Conf. on Speech Communication and Technology*, 2001.

[13] L. Mayor, B. Jensen, A. Lorotte, and R. Siegwart. Improving the expressiveness of mobile robots. In *Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*.

[14] H.P. Moravec and A.E. Elfes. High resolution maps from wide angle sonar. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1985.

[15] K. Nickel, E. Seemann, and R. Stiefelhagen. 3D-Tracking of heads and hands for pointing gesture recognition in a human-robot interaction scenario. In *International Conference on Face and Gesture Recognition (FG)*, 2004.

[16] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto. An affective mobile robot educator with a full-time job. *Artificial Intelligence*, 114(1-2):95–124, 1999.

[17] I. Nourbakhsh, C. Kunz, and T. Willeke. The Mobot museum robot installations: A five year experiment. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2003.

[18] Novotech. GPMSC (General Purpose Machines' Speech Control). http://www.novotech-gmbh.de/speech_control.htm, 2005.

[19] H. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. of the Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, 2002.

[20] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann. Using gesture and speech control for commanding a robot assistant. In *Proc. of IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002.

[21] Z. Ruttkay, H. Noot, and P. ten Hagen. Emotion Disc and Emotion Squares: Tools to explore the facial expression space. *Computer Graphics Forum*, 22(1):49–53, 2003.

[22] J. Schulte, C. Rosenberg, and S. Thrun. Spontaneous short-term interaction with mobile robots in public places. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1999.

[23] C.L. Sidner, C.D. Kidd, C. Lee, and N. Lesh. Where to look: A study of human-robot engagement. In *ACM Int. Conf. on Intelligent User Interfaces (IUI)*, 2004.

[24] R. Siegwart, K.O. Arras, S. Bouabdallah, D. Burnier, G. Froidevaux, X. Greppin, B. Jensen, A. Lorotte, L. Mayor, M. Meisser, R. Philippsen, R. Piguet, G. Ramel, G. Terrien, and N. Tomatis. Robox at Expo.02: A large-scale installation of personal robots. *Robotics & Autonomous Systems*, 42(3-4):203–222, 2003.

[25] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2004.

[26] K. R. Thórisson. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In B. Granström, D. House, and I. Karlsson, editors, *Multimodality in Language and Speech Systems*, pages 173–207. Kluwer Academic Publishers, 2002.

[27] S. Thrun. Towards a framework for human-robot interaction. *Human Computer Interaction*, 2003. Forthcoming.

[28] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, J. Schulte, and D. Schulz. Probabilistic algorithms and the interactive museum tour-guide robot Minerva. *Int. Journal of Robotics Research (IJRR)*, 19(11):972–999, 2000.

[29] T. Tojo, Y. Matsusaka, T. Ishii, and T. Kobayashi. A conversational robot utilizing facial and body expressions. In *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 2000.

[30] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.