In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2005.

Integrating Vision and Speech for Conversations 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*— An essential capability for a robot designed to interact with humans is to show attention to the people in its surroundings. To enable a robot to involve multiple persons into interaction requires the maintenance of an accurate belief about the people in the environment. In this paper, we use a probabilistic technique to update the knowledge of the robot based on sensory input. In this way, the robot is able to reason about the uncertainty in its belief about people in the vicinity and is able to shift its attention between different persons. Even people who are not the primary conversational partners are included into the interaction. In practical experiments with a humanoid robot, we demonstrate the effectiveness of our approach.**

I. Introduction

Our goal is to develop a humanoid robot that performs intuitive multi-modal interaction with multiple persons simultaneously. One application in this context is an interactive museum tour-guide. Compared to previous museum tour-guide projects [6, 19, 22, 24], which focused on the autonomy of the robots and did not emphasize the interaction part that much, we want to build a robot that behaves and acts like a human. Over the last few years, humanoid robots have become very popular as a research tool. One goal of building robots with human-like bodies and behavior is that people can easily understand their gestures and know intuitively how to interact with such a system.

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 gestures [2, 4, 5, 14, 23, 25]. Only little research has been done in the area of developing a robotic system that really behaves as a conversational partner and acts human-like when *multiple* persons are involved. A prerequisite for this task is that the robot detects people in its surroundings, keeps track of them, and remembers them even if they are currently outside its limited field of view. In this paper, we present a system that makes use of visual perception and speech recognition to detect, track, and involve people into interaction. In contrast to previous approaches [13, 16, 21], our goal is that the robot interacts with multiple persons and does not focus its attention on only one single person. It should also not 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. Furthermore, we developed a strategy that makes the robot look at the persons to establish short eye-contact and to signal attentiveness. Eye movements play an important role during a conversation (compare to Breazeal et al. [3]). Vivid human-like eye-movements that signal attentiveness to people make them feel involved. Fig. 1 shows our robot Alpha shifting its attention from one person to the other during a conversation.

Fig. 1. A conversation of our robot Alpha with two people. As can be seen, the robot shifts its attention from one person to the other to include both into the conversation.

To detect people in the environment of our robot, we use the data delivered by a pair of cameras. To keep track of people over time, we maintain a probabilistic belief and update it based on sensory input. First, we run a face detection system to find faces in the current pair of images. Then, we apply a mechanism to associate the detected faces to people already stored in the belief and update those according to the observations. Since the field of view of the robot is constrained, it moves the cameras from time to time to explore the environment and to get new information about people. Our approach maintains a probabilistic belief about people in the surroundings even if they are currently not in the field of view of the robot.

This paper is organized as follows. The next section gives an overview over related work. Section III presents the hardware of our robot and introduces the basic concepts of our behavior control architecture. In Section IV, we describe how we detect and keep track of people using vision information. In Section V, we present our speech processing and dialogue system. In Section VI, we describe our strategy on how to determine the gaze direction of the robot and how to decide which person gets the attention. Finally, in Section VII, we show experimental results.

II. Related Work

Over the last few years, much research has been carried out in the area of multi-modal interaction. Lang et al. [13] presented a system that combines several sources of information (laser, vision, and sound data) to track people. Since their sensor field of view is much larger than that of our robot, they are not forced to make the robot execute observation actions to get new information about surrounding people. They apply an attention system in which only the person who 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 robot will show attention to another person. Okuno et al. [21] also apply audio-visual tracking and 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. [16] is able to determine the one who is being addressed to in the conversation. In contrast to our application scenario (museum tour-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 attention system presented by Breazeal et al. [3] only keeps track of objects that are located in the field of view of the cameras. In contrast to this, we keep track of people over time and maintain a probabilistic belief about detected faces even if they are currently not observable. Many vision-based approaches exist that aim to reliably track a target in real-time [7, 8, 17, 28]. However, those techniques focus on a single face that is tracked and do not intend to maintain a belief about more than one detected face. Like our approach, the technique presented by Fröba and Küblbeck [9] analyzes the whole image to find all faces and uses a set of Kalman filters to track the faces independently. They apply a greedy nearest neighbor assignment to solve the data association problem. Furthermore, their approach requires a sequence of K positive observations to initialize a tracker. In contrast to this, our work applies a probabilistic update technique to compute the existence probability of a face directly from the beginning and does not distinguish between an initialization and a tracking phase. Additionally, we use a different data association technique.

III. The Design of our Robot

The body (without the head) of our robot Alpha currently has 21 degrees of freedom (six in each leg, three in each arm, and three in the trunk; see left image of Fig. 2). Its total height is about $155cm$. The skeleton of the robot is constructed from carbon composite materials to achieve a low weight of about $30kg$.

To perform the experiments presented in this paper, we used only the head of our robot which is depicted in Fig. 2 (right image). The head consists of 16 degrees of freedom that are

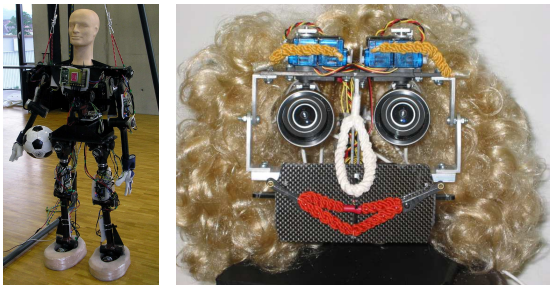

Fig. 2. The left image shows the body of our robot Alpha. The image on the right depicts the head of Alpha in a happy mood.

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). Furthermore, we take into account the estimated distance to a target to compute eye vergence movements. These vergence movements ensure that the target maintains in the center of the field of view of both cameras. Thus, if a target comes closer, we turn the eyes toward each other. For controlling the eye movements, we follow a similar approach to the one presented by Breazeal et al. [3].

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.

We animate the mouth of the robot while it is speaking. Based on the ten servos for the mouth and the eyebrows, we are also able to animate different facial expressions. To enrich human-robot interaction and to express how the robot changes its mood, in the future, we plan to apply a technique to change its facial expression.

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. Detecting and Tracking People

To sense people in the environment of our robot, we use the data delivered by the two cameras. Our robot maintains a probabilistic belief about people in its surroundings to deal with multiple persons appropriately. To find people, we first run a face detector in the current pair of images. Then, we apply a mechanism to associate the detections to faces already stored in the belief and update it according to these observations.

Our face detection system is based on the AdaBoost algorithm and uses a boosted cascade of Haar-like features [15]. 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, for example, the region of the eyes and the cheeks. Originally, this idea was developed by Viola and Jones [27] to reliably detect faces without requiring a skin color model. This method works quickly and yields high detection rates. However, since false classifications are possible, we apply a probabilistic technique to deal with the uncertainty in the detection process.

Maintaining a belief about faces in the surroundings of the robot over time is similar to the map building problem with noisy sensors in mobile robotics. A classical way to update a belief upon sensory input is to apply a recursive Bayesian scheme like the one proposed by Moravec and Elfes [18]. In our case, this update scheme determines the probability of the existence of a face (i.e. of a person) given a sequence of positive and/or negative observations:

$$P(f \mid z_{1:t}) = \left[1 + \frac{1 - P(f \mid z_t)}{P(f \mid z_t)} \cdot \frac{P(f)}{1 - P(f)} \cdot \frac{1 - P(f \mid z_{1:t-1})}{P(f \mid z_{1:t-1})}\right]^{-1} \quad (1)$$

Here f denotes the existence of a face, z_t is the observation (face detected/not detected) at time t, and $z_{1:t}$ refers to the observation sequence up to time t.

As typically assumed in mobile robot map building, we set the prior probability (here $P(f)$) to 0.5. Therefore, the second term in the product in Eq. (1) becomes 1 and can be neglected. Further values that have to be specified are the probability $P(f \mid z = det)$ that a face exists if it is detected in the image and the probability $P(f \mid z = \neg det)$ that a face exists if it is not detected. In our experiments, it turned out that adequate values for those parameters are 0.9 and 0.2, respectively. Using the update rule in Eq. (1), the probability of the existence of a face is increased if positive observations occur and is decreased otherwise.

To track the position of a face over time, we use a Kalman filter [11]. Applying such a filter leads to a smoothing of the estimated trajectories. Each face is tracked independently, and its state vector contains the position and the velocities. Before we can update the Kalman filters and the probabilities of the faces using observations, we must first solve the data association problem, i.e., we must determine which observation corresponds to which face of our belief and which observation belongs to a new face. Since we currently do not have a mechanism to identify people, we use a distance-based cost function and apply the Hungarian method [12] to determine the mapping from observations to faces. 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. In our case, the Hungarian method computes the optimal assignment of detected faces in the current camera images to faces already existing in the belief, given a cost function that takes into account the distances between new observations and existing faces. Note that the cost function within the Hungarian method can also be used to integrate a similarity measure between different faces.

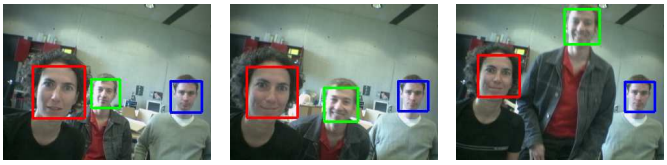

Fig. 3. Tracking three faces with independent Kalman filters. To solve the data association problem we apply the Hungarian method.

To account for the fact that an observation can belong to a face not stored in the belief so far, we add "dummy faces" to the input of the Hungarian method. These dummy faces imply high costs and are therefore only chosen if an observation cannot be assigned to an already existing face in the belief. If we have an observation that is assigned to a dummy face, we initialize a new Kalman filter to track the corresponding face. The update formula in Eq. (1) is used to compute the probability whenever an observation occurs. If the probability of a face drops below a certain threshold, the corresponding Kalman filter is deleted. Either the face was a false positive detection, or the person corresponding to the face moved away. To reduce the probability of false positive detections, we run the face detector in both images. The data association between faces in both images is also solved using the Hungarian method.

In our experiments, we found out that our method works reliably in sparsely populated environments. However, it may fail in crowded situations, also due to the lack of a face recognition system. Fig. 3 shows three snapshots during face tracking using independent Kalman filters and applying the Hungarian method to solve the data association problem. As indicated by the differently colored boxes, all faces are tracked correctly.

Since the field of view of our robot is constrained due to the opening angle of the cameras, we also have to keep track of people whose faces cannot currently be observed. In these cases, we set the velocities in the state vector to zero since we do not know how people move when they are outside the field view. To compute the corresponding probabilities of the people outside the field of view, we also use the update formula in Eq. (1). In this case, we set $P(f \mid z)$ in that equation to a value close to 0.5. This models the fact that the probabilities of people who are assumed to be in the vicinity of the robot but outside its field of view decrease only slowly over time. As explained in Section VI, the robot changes its gaze into the direction of a person to check whether it can be detected whenever its uncertainty exceeds a certain threshold.

V. Speech Processing and Dialogue Management

For speech recognition, we currently use a commercial software [20]. This recognition software has the advantages that it is speaker independent and yields high recognition rates even in noisy environments, which is essential for the environments in which we deploy the robot. The disadvantage, however, is that no sentence grammar can be specified. Instead, a whole list of keywords/phrases that should be recognized needs to be defined. For speech synthesis, we use a freely

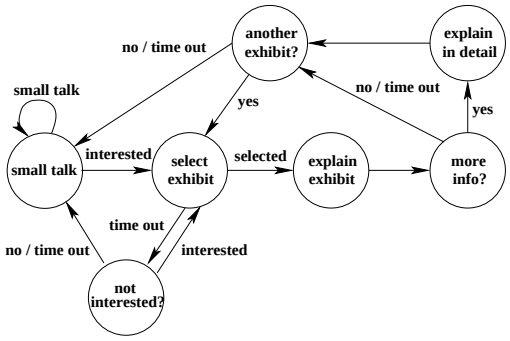

Fig. 4. The finite state machine that models typical dialogues between our robot whose task is to act as a museum tour-guide and visitors. State transitions occur when an utterance is correctly recognized or when no utterance is recognized after a certain period of time.

available system [26] that generates synthesized speech based on strings online.

Our dialogue system is realized as a finite state machine. State transitions in this automaton occur when an utterance is correctly recognized, or when no utterance is recognized after a certain period of time. With each state, a different list of keywords/phrases is associated. This list is sent to the speech recognition system whenever the state of the automaton changes. Fig. 4 depicts the basic structure of the finite state machine of our dialogue system for the situation in which the robot acts as a museum tour-guide. During such a task, this automaton models typical dialogues with visitors in a museum.

Initially, the system is in the state "small talk". In this state, the robot tries to attract visitors and to involve them into a conversation that consists of simple questions and answers. Whenever a user shows interest in exhibits, the robot changes its internal state and explains the exhibits. Possible courses of dialogues can be deduced from Fig. 4. For different tasks carried out by the robot, we apply different finite state machines to model a dialogue.

We also implemented a technique for sound source localization. We apply the Cross-Power Spectrum Phase Analysis [10] to calculate the spectral correlation measure $C_{lr}(t,\tau)$ between the left and the right microphone channel:

$$C_{lr}(t,\tau) \quad = \quad FT^{-1}\frac{\widehat{S}_l(t,w)\widehat{S}_r^*(t,w)}{|\widehat{S}_l(t,w)||\widehat{S}_r(t,w)|}. \qquad (2)$$

Here $\widehat{S}_l(t,w)$ and $\widehat{S}_r(t,w)$ are the short-term power spectra of the left and right channel and $\widehat{S}_r^*(t,w)$ is the complex conjugate. $\widehat{S}_l(t,w)$ and $\widehat{S}_r(t,w)$ are computed through Fourier transforms, applied to windowed segments centered around time t. FT^{-1} denotes the inverse Fourier transform.

Assuming only a single sound source, the argument τ that maximizes $C_{lr}(t,\tau)$ yields the delay δ between the left and the right channel. Once δ is determined, the relative angle between the speaker and the microphones can be calculated under two assumptions [13]: 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. In the experiments, we demonstrate that this technique allows an accurate localization of a speaker.

This information can then be used for example to align the microphones with the speaker, to apply beam-forming, or to shift the attention of the robot to the speaker.

VI. Gaze Direction and Focus of Attention

As explained so far, our robot maintains a belief about the positions of faces as well as the corresponding probabilities. Additionally, it computes for each person an importance value that currently depends 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 stand directly in front of the robot have a high importance. The same applies to people who are close to the robot. The resulting importance value is a weighted sum of those factors.

The behavior system controls the robot in such a way that it focuses its attention on the person who has the highest importance. Thus, the robot follows the movements of the corresponding face and looks the user in the eyes. 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.

Note that one can also consider further information to determine the importance of a person. For example, we plan to use our sound source localization system in this context. As a result, our robot shows human-like behavior since humans usually focus their attention to people standing in front of them, to people who come very close, or to people who speak to them.

Since the field of view of the robot is constrained, it is important that the cameras move from time to time to explore the environment to update its belief about people which are currently not in the field of view. Thus, we additionally implemented a behavior that forces the robot to regularly change its gaze direction and to look in the direction of other detected 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.

Furthermore, if the robot gets too uncertain about whether or not a person who is outside the field of view is still there, it should look around to reduce its uncertainty. The uncertainty of a belief can be determined by the entropy. The entropy H of a discrete posterior $p(x)$ is computed by

$$H(p(x)) \quad = \quad -\sum_{x_i} p(x_i)\cdot\log p(x_i). \qquad (3)$$

Here x_i are the possible values of a discrete belief. In our case, Eq. (3) simplifies to

$$H(p(f)) = -p(f)\log p(f) - (1-p(f))\log(1-p(f)). \quad (4)$$

The entropy of a posterior is maximum in case of a uniform distribution, and is zero in case the robot is absolutely certain about the existence of a face. As soon as the entropy in its belief about a person exceeds a certain threshold, the robot

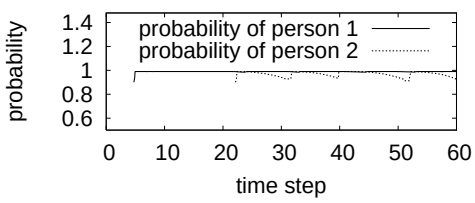

Fig. 5. Evolution of the probabilities of two people. While the robot is interacting with person 1, it also shows interest in person 2 by establishing short eye-contact and updates its belief at time steps 22, 32, 40, and 52. Note that person 2 is outside the field of view while the robot is concentrating on person 1.

considers to perform an observation action and to look to the predicted position of the corresponding face to reduce the uncertainty.

VII. Experimental Results

In order to evaluate our approach to control the gaze direction of the robot and to determine the person who gets the focus of its attention, we performed several experiments in our laboratory. Furthermore, we present experimental results demonstrating the accuracy of our speaker localization system. Besides the experiments presented in this section, we provide videos of our robot Alpha on our webpage[1].

All experiments were performed on a Pentium IV with $2.8GHz$. Using a camera resolution of 320×240 pixels, the face detection algorithm detects faces in a distance of approximately $30 - 200cm$. To speed up the computation of the image processing, we search the whole images for faces only twice in a second. In the time between, we only consider regions in the images. The sizes and locations of these search windows are determined based on the predicted states of the corresponding Kalman filters. Depending on the sizes of the extracted search windows, we operate at a rate of $15 - 25Hz$.

A. Signaling Attentiveness

The first experiment was designed to demonstrate how the robot establishes short eye-contact to a person in order to signal attentiveness. The evolution of the probabilities of two people over time is depicted in Fig. 5. When the robot detected person 1 at time step 4, it started to interact with it. After a gaze to explore the environment and to not stare into the eyes of person 1 all the time, the robot detected person 2 (time step 22). Since person 2 had a lower importance value than person 1, the robot continued its conversation with person 1. Thus, the probability of person 2 decreased in the following time steps since it was outside the robot's field of view again. However, to involve person 2 into the conversation as well the robot regularly looked to person 2 and established short eye-contact. As can be seen from Fig. 5, at time steps 32, 40, and 52 the robot looked to person 2 and also updated its belief correctly. Note that we do not evaluate the camera images during the rapid saccades to avoid false positive or negative detections. During a saccade, the belief therefore stays constant for a short period of time.

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

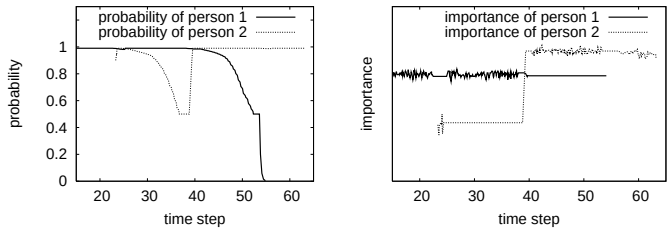

Fig. 6. Evolution of the probabilities of two people (left image) and the corresponding importance values (right image). In the beginning, the robot is interacting with person 1. At time step 22, after an exploring gaze, the robot detects person 2, which was outside its field of view before. Since person 2 has a lower importance, the robot continues concentrating on person 1. Thus, person 2 is outside its field of view again. After looking to person 2 at time step 40, the robot shifts its attention to this person since the robot noticed that it had come very close and is now considered as more important. When the robot looks back to person 1 (time step 53), the person cannot be detected anymore and the robot updates its belief accordingly.

B. Shifting Attention

The following 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. The left image of Fig. 6 shows the evolution of the probabilities of two people during this experiment. In the beginning, the robot was interacting with person 1. At time step 22, the robot performed an exploring gaze. As can be seen from the figure, the robot detected the face of person 2, which was outside its field of view before. Since person 1 had a much higher importance value (it was closer; see the right image of Fig. 6), the robot continued its dialogue with person 1. During the following time steps, the probability of person 2 decreased because it was not in the field of view of the robot anymore. However, the robot kept person 2 in its belief. At time step 40, the robot decided to look to person 2 to reduce its uncertainty about the presence of this person. The robot detected person 2 and noticed that it tried to attract the robot's attention by coming much closer. As a result, person 2 got a higher importance value than person 1 and the robot shifted its attention to person 2. At the same time, the probability of person 1 decreased since it was not in the field of view anymore. At time step 53, the robot looked back into the direction of person 1 to see whether it was still there. However, the person had gone and the robot updated its belief accordingly.

C. Speaker Localization

In the last experiment, we demonstrated the accuracy of our speaker localization. We calculated the short-term power spectra of the left and right microphone channel within a $42.57ms$ window of a signal at $48kHz$ and performed for different angles 50 localizations. The ground truth versus the estimated angle is plotted in Fig. 7. The error bars indicate the 0.95 confidence interval. As can be seen, the localization of the speaker in the vicinity of the robot is quite accurate.

VIII. Conclusions

In this paper, we presented an approach to enable a humanoid robot to converse with multiple persons. We use a

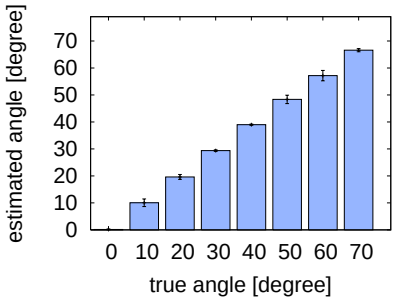

Fig. 7. Ground truth and estimated angle of a speaker (standing in a distance of 1 m) determined with our sound source localization technique.

probabilistic technique to update a belief about people in its surroundings based on vision data. The robot is able to maintain estimations about the positions of people even if they are temporarily outside its field of view. On the one hand, this technique enables the robot to move its cameras to actively search for people as soon as the uncertainty in its belief gets too high. On the other hand, we can apply an intelligent strategy to change the focus of attention, and in this way can attract multiple persons and involve them into a conversation. As a result, we obtain a human-like interaction behavior that shows attentiveness to multiple persons. In practical experiments, we demonstrated our technique to reliably update the belief of our robot and to control its gaze direction. Additionally, we evaluated our speaker localization system which will be integrated into our attention system as well.

In the near future, we will combine the head and the body, in order to enable the robot to perform human-like gestures and movements. Furthermore, we will present the robot to the public soon to see how people interact with the system and to get new insights on how to improve the system.

ACKNOWLEDGMENT

This project is supported by the DFG (Deutsche Forschungsgemeinschaft), grant BE 2556/2-1. We would like to thank Novotech for providing us with an evaluation license for the speech recognition software.

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] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda. Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*, 2004. Submitted for publication.

[3] 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.

[4] L. Brèthes, P. Menezes, F. Lerasle, and J. Hayet. Face tracking and hand gesture recognition for human-robot interaction. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2004.

[5] A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2002.

[6] W. Burgard, A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2):3–55, 2000.

[7] M. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, 2000.

[8] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Learning to identify and track faces in image sequences. In *Proc. of the Int. Conf. on Face and Gesture Recognition*, 1998.

[9] B. Fröba and C. Küblbeck. Face tracking by means of continuous detection. In *Proc. of the CVPR Workshop on Face Processing in Video (FPIV)*, 2004.

[10] 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.

[11] R.E. Kalman. A new approach to linear filtering and prediction problems. *ASME-Journal of Basic Engineering*, 82(March):35–45, 1960.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] A. Mojaev and A. Zell. Real-time face tracking using discriminator technique on standard PC hardware. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2004.

[18] 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.

[19] 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.

[20] Novotech. GPMSC (General Purpose Machines' Speech Control). http://www.novotech-gmbh.de/speech_control.htm, 2004.

[21] 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.

[22] 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 and Autonomous Systems*, 42(3-4):203–222, 2003.

[23] 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.

[24] 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*, 19(11):972–999, 2000.

[25] 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.

[26] Institut für Kommunikationsforschung und Phonetik University of Bonn. Txt2pho – German TTS front end for the MBROLA synthesizer. http://www.ikp.uni-bonn.de/dt/forsch/phonetik/hadifix/HADIFIXforMBROLA.html, 2000.

[27] 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.

[28] J. Yang and A. Waibel. A real-time face tracker. In *Proc. of the IEEE Workshop on Applications of Computer Vision (WACV)*, 1996.