Multimodal Conversation between a Humanoid Robot and Multiple Persons

Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke
University of Freiburg
Computer Science Institute
79110 Freiburg, Germany
{maren,faber,joho,schreibe,behnke}@informatik.uni-freiburg.de

Abstract

Attracting people and involving multiple persons into an interaction is an essential capability for a humanoid robot. A prerequisite for such a behavior is that the robot is able to sense people in its vicinity and to know where they are located. In this paper, we propose an approach that maintains a probabilistic belief about people in the surroundings of the robot. Using this belief, the robot is able to memorize people even if they are currently outside its limited field of view. Furthermore, we use a technique to localize a speaker in the environment. In this way, even people who are currently not the primary conversational partners or who are not stored in the robot's belief can attract its attention. To enrich human-robot interaction and to express how the robot changes its mood, we apply a technique to change its facial expressions. As we demonstrate in practical experiments, by integrating the presented techniques into its control architecture, our robot is able to interact with multiple persons in a multimodal way and to shift its attention between different people.

Introduction

Our goal is to develop a humanoid robot that performs intuitive multimodal interaction with multiple persons simultaneously. One application in this context is an interactive museum tour-guide. Compared to previous museum tour-guide projects (Thrun *et al.* 2000; Siegwart *et al.* 2003), 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 (Breazeal *et al.* 2001; Brèthes *et al.* 2004; Li *et al.* 2004; Stiefelhagen *et al.* 2004; Tojo *et al.* 2000). 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 fi eld of view.

Figure 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 involve both into the conversation.

In this paper, we present a system that integrates several components into one control architecture. More precisely, our system makes use of visual perception, sound source localization, and speech recognition to detect, track, and involve people into interaction. In contrast to previous approaches (Lang *et al.* 2003; Matsusaka, Fujie, & Kobayashi 2001; Okuno, Nakadai, & Kitano 2002), 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. We believe that eye movements play an important role during a conversation (also compare to Breazeal *et al.* (2001)). Vivid human-like eye-movements that signal attentiveness to people make them feel involved. Figure 1 shows our robot Alpha shifting its attention from one person to the other during a conversation.

We use Alpha also as an emotional display. 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, expressing emotions helps to indicate the robot's state or its intention and to show how the robot is affected by events in its environment (Fong, Nourbakhsh, & Dautenhahn 2003). Our goal is to attract people and engage them in a conversational interaction with our robot. To make the interaction even more human-like, we use a face with animated mouth and eyebrows 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.

Related Work

Over the last few years, much research has been carried out in the area of multimodal interaction. In this section, we first concentrate on systems that use different types of perception to sense and track people during an interaction and that use a strategy how to decide which person gets the attention of the robot. Then we present systems that make use of facial expressions to display some emotion of the robot.

Lang *et al.* (2003) presented an approach that combines several sources of information (laser, vision, and sound data) to track people. Since their sensor field of view is much larger than ours, 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 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 robot will show attention to another person. Okuno, Nakadai, & Kitano (2002) 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, Fujie, & Kobayashi (2001) is able to determine the one who is being addressed to in the conversation. Compared 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.* (2001) 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. The model developed by Thórisson (2002) focuses on turn-taking in one-to-one conversations. In contrast to this, 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.

Several robots that make use of facial expressions to show emotions have already been developed. Schulte, Rosenberg, & Thrun (1999) 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 whether and 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.* (1999) 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. Domínguez Quijada *et al.* (2002) developed a head for a robotic tour-guide. The face can display various facial expressions with random intensities. Bruce, Nourbakhsh, & Simmons (2002) used a three-dimensional rendered face to display facial expressions in order to make people comply simple requests of the robot. Breazeal (2003) presented a robotic head that is able to display a variety of facial expressions. The emotional expressions are computed using interpolation in the three-dimensional space with the dimensions arousal, valence, and stance. The robotic head is used to analyze and learn social interactions between an infant (the robot) and its caregiver. Cañamero & Fredslund (2001) built a LEGO robot that can express six different emotional states at various intensities. The emotions are activated by tactile stimuli that are sensed using binary touch sensors on the feet. The goal was to achieve believable human-robot interaction. Scheeff *et al.* (2000) developed a robot that can express nine different emotional states. The robot is tele-operated in order to be able to analyze human-robot interaction more thoroughly. Esau *et al.* (2003) designed a feedback-loop to control a robot based on emotions. They designed a robotic head that is able to express four different emotional states. Arkin *et al.* (2003) presented a system that learns new objects and associates emotional effects to them. They use six basic emotional states in the three-dimensional space with the dimensions pleasantness, arousal and confidence. Suzuki *et al.* (1998) developed an emotion model that consists of four states. The authors apply a self-organizing map to compute the robot's emotional state based on external events.

Most of the existing approaches do not allow continuous changes of the emotional expression. Our approach, in contrast, uses a bi-linear interpolation technique in a two-dimensional state space (Ruttkay, Noot, & ten Hagen 2003) to smoothly change the robot's facial expression.

The Design of our Robot

The body (without the head) of our robot Alpha has currently of 21 degrees of freedom (six in each leg, three in each arm, and three in the trunk; see left image of Figure 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 focus on the head of our robot, which is shown in Figure 2 (right image). The head consists of 16 degrees of freedom that are driven by servo motors. Three of these servos move a stereo camera system 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 cam-

Figure 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.

eras 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 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.* (2001).

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. The different capabilities are implemented as independent modules that are able to update and to query information stored in the belief of the robot.

For the behavior control of our robot, we use a framework developed by Behnke & Rojas (2001) that supports a hierarchy of reactive behaviors. In this framework, behaviors are arranged in layers that work on different time scales.

Visual Detection and Tracking of People

Our robot maintains a probabilistic belief about people in its surroundings to deal with multiple persons appropriately. In this section, we describe our vision system that senses people in the environment using the data delivered by the two cameras. 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 (Lienhard & Maydt 2002). 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 & Jones (2001) 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. Thus, to maintain a belief about faces in the surroundings of the robot over time, we update the belief based on sensory input by applying the recursive Bayesian scheme proposed by Moravec & Elfes (1985). In our case, this update scheme determines the probability of the existence of a face (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, 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 (anymore). 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 (Kalman 1960). Applying such a filter leads to a smoothing of the estimated trajectories of the faces. 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 (Kuhn 1955) 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 under the given cost function. If we have an observation to which no existing face is assigned, 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. Figure 3 shows three snapshots during face tracking

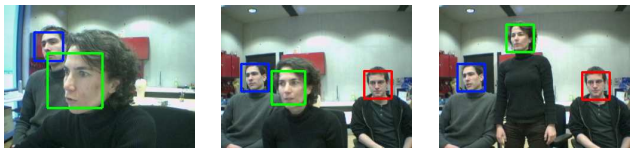

Figure 3: Tracking three faces with independent Kalman filters. To solve the data association problem we apply the Hungarian method.

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.

Speaker Localization

Additionally, we implemented a system that performs sound source localization. We apply the Cross-Power Spectrum Phase Analysis (Giuliani, Omologo, & Svaizer 1994) 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 a speaker and the microphones can be calculated under two assumptions (Lang *et al.* 2003): 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.

Once the sound source localization system has localized the speaker, the information that the person has spoken is assigned to that 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 that just entered the scene.

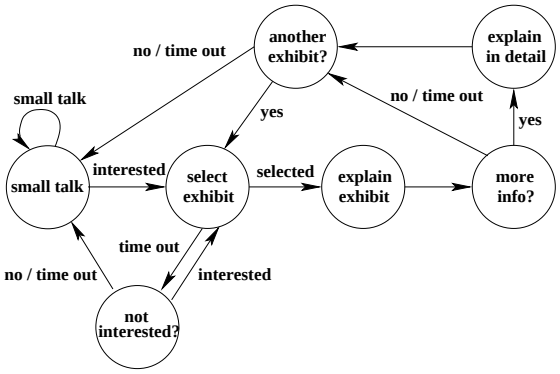

Figure 4: The finite state machine that models typical dialogues between our robot whose task is to act as a museum tour-guide and visitors. State transitions occur when an utterance is correctly recognized or when no utterance is recognized after a certain period of time.

Dialogue Management

For speech recognition, we currently use a commercial software (Novotech 2005). 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 available system (University of Bonn 2000) 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.

Figure 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 Figure 4. For different tasks carried out by the robot, we apply different finite state machines to model a dialogue.

Gaze Control and Focus of Attention

As explained so far, our robot maintains a belief about the positions of faces as well as the corresponding probabilities and the information about when the person has spoken last. Additionally, it computes for each person an impor-

tance value that currently depends on when the person has spoken last, 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 high importance. 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 those three 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. 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 its hands to get the robot's attention, this could be easily integrated as well.

To enable the robot to react to an unknown person outside its fi eld of view who is speaking to it, we implemented a behavior that reacts to salient sounds sources that cannot be assigned to any person already existing in the belief. Thus, the robot looks into the direction of the speaker to signal attentiveness and to update its belief.

Since the fi eld of view of the robot is constrained, it is important that the cameras move from time to time to explore the environment to get new information about other people. 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, the robot is in this way able to update its belief about people outside its current fi eld of view.

As a result, the robot shows human-like behavior since humans usually focus their attention on people who speak to them, on people standing in front of them, and on people who come very close.

Facial Expressions

Our robot is able to express its mood by means of facial expressions, with animated mouth and eyebrows, to make the conversation more human-like and to provide additional feedback to the conversational partners.

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, Noot, & ten Hagen (2003). 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 5), with the neutral expression in the center. The Emotion Disc can be used to control the expression of any

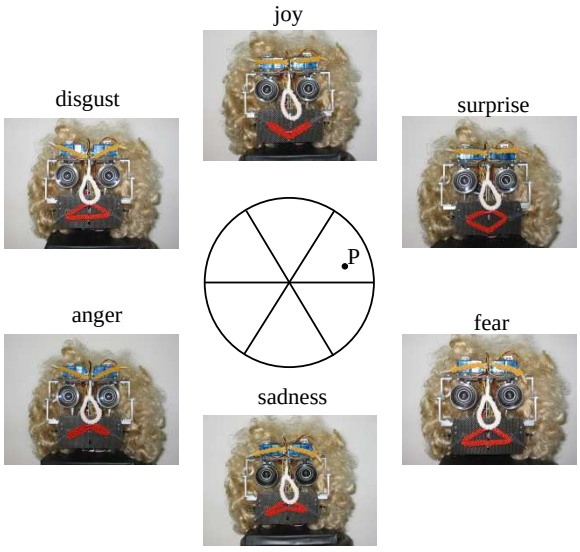

Figure 5: The two-dimensional space in which we compute the robot's facial expression. The expression corresponding to point P is computed according to Eq. (3). The images show the six basic emotional expressions of our robot

facial model once the neutral and the six basic expressions are designed. Figure 5 shows the six basic facial expressions of our robot. In our case, we defi ne them in terms of height of the mouth corners, mouth width, mouth opening angle, and angle and height of the eye-brows. 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:

$$P' = l(p) \cdot (\alpha(p) \cdot E'_i + (1 - \alpha(p)) \cdot E'_{i+1}). \qquad (3)$$

Here, $l(p)$ is the length of the vector p, which 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. 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.

Experimental Results

To evaluate our approach, which controls the gaze direction of the robot and which determines the person who gets the focus of its attention, we performed several experiments in our laboratory. Furthermore, we present experiments in which we utilize changes of the robot's facial expression to enrich human-robot interaction. Using a camera resolution of 320×240 pixels, the face detection algorithm detects

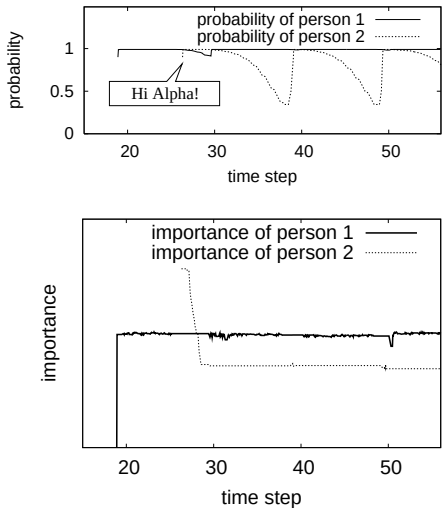

Figure 6: Evolution of the probabilities of two people (top image) and the corresponding importance values (bottom image). While the robot is chatting with person 1, it recognizes the voice of a second person and turns towards it at time step 26. As can be seen, person 2 is detected and the robot updates its belief accordingly. Person 2 does not continue talking and thus its importance decreases during the following time steps. Therefore, the robot concentrates on person 1 again (time step 29) but also shows interest in person 2 by establishing short eye-contact and updates its belief at time steps 38 and 49.

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.

Localizing a Speaker and Signaling Attentiveness

The first experiment was designed to demonstrate how the robot reacts to a person outside its current field of view who is talking to it and how the robot establishes short eye-contact to signal attentiveness. The evolution of the probabilities of two people over time is depicted in the top image of Figure 6. When the robot detected person 1 at time step 18, it started to interact with it. At time step 26 the robot recognized the voice of a second person, who was outside its field of view, and turned towards it. As can be seen, the face of person 2 is detected and the robot updated its belief. The importance value of person 2 decreased during the following time steps (see bottom image of Figure 6) since it was farther away and did not continue talking. Thus, the robot proceeded concentrating on person 1 (time step 29). However, to involve person 2 into the conversation as well and to update its belief, the robot regularly looked to person 2 and established short eye-contact. 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

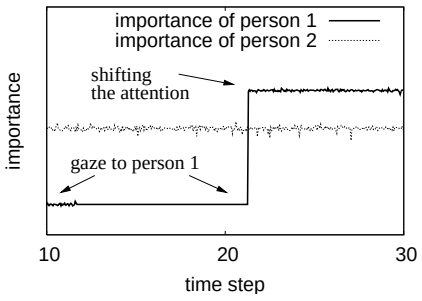

Figure 7: Evolution of the importance values of two people. During this experiment, person 2 is talking to the robot. Thus, it has initially a higher importance than person 1. The robot focuses its attention on person 2 but also looks to person 1 at time steps 10 and 21 to demonstrate that it is aware of person 1. At time step 21 the robot notices that person 1 had come very close and thus it shifts its attention to person 1, which has a higher importance now.

therefore stays constant for a short period of time. As can be seen from Figure 6, at time steps 38 and 49, the robot looked to person 2 and also updated its belief correctly.

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. In this experiment, person 2 was talking to the robot. Thus, the robot initially focused its attention on person 2 because it had the highest importance. However, the robot looked to person 1 at time steps 10 and 21 to signal awareness. When looking to person 1 at time step 21 the robot suddenly noticed that this person had come very close. Accordingly, person 1 got then a higher importance value and the robot shifted its attention to this person. As this experiment shows, our robot does not focus its attention exclusively on the person that is speaking.

Changes of the Facial Expression

The last experiment aims to demonstrate how the robot changes its emotional state according to external events. In the beginning of the experiment, the robot had not detected any person for several minutes and therefore its facial expression was a blending of sadness and fear (see top left image of Figure 8). Afterwards, in the situation shown in the center image of the first row, the robot suddenly detected a person and displayed a mood corresponding to a mixture of surprise and happiness. Since the person started to interact with the robot, the robot got happy as shown in the following images.

Demonstration at RoboCup German Open 2005

We presented Alpha during the RoboCup German Open 2005 in Paderborn. Figure 9 shows the robot in a conversation with three people. We asked people who interacted

Figure 8: This figure shows continuous changes of the facial expression. Initially, the robot is in a mood corresponding to a blending of sadness and fear since it is alone (top left image). Then, the robots suddenly detects a person and changes its facial expression towards surprise (center image in the first row). Afterwards, the robot gets happy because the person starts to interact with it (following images).

Figure 9: Alpha is intercting with three people at the RoboCup German Open 2005 in Paderborn.

with the robot to fill out questionaires to get feedback. The 30 people between the age of 10 and 61 who filled out the questionaire had fun in interacting with Alpha and noticed that the robot was aware of their presence. The people found the eye-gazes and facial expression human-like and could recognize different emotional states. Most of the people interacted with the robot for more than three minutes.

Besides the experiments presented in this paper, we provide videos of our robot Alpha on our webpage[1]. Currently, only conversations in German are possible but we are already working on integrating English speech recognition and synthesis as well. In the videos, we want to demonstrate how the robot performs exploration gazes in the beginning, how it reacts to sound, how it changes its focus of attention, and how it establishes short eye-contact in order to signal attentiveness.

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

Conclusions and Future Work

In this paper, we presented an approach to enable a humanoid robot to interact with multiple persons in a multimodal way. We described all components that our robot control architecture comprises. We use a probabilistic technique to maintain a belief about the presence of people in the surroundings of our robot based on vision data. The robot is able to estimate the positions of people even if they are temporarily outside its field of view. To enable the robot to shift its attention to people who are talking to it, we use a system to localize the direction of speakers. Using vision and sound information, we can apply an intelligent strategy to change the focus of attention, and in this way can attract multiple persons and include them into a conversation. To enrich human-robot interaction and to express the robot's approval or disapproval to external events, we use a technique to change its facial expression.

As a result, we obtain a human-like interaction behavior that shows attentiveness to multiple persons. In practical experiments, we demonstrated that our technique was able to reliably update the belief of our robot, to control its focus of attention and its gaze direction, and to change its facial expression.

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 a broader public soon to see how people interact with the system and to get new insights how to improve the system.

Acknowledgment

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

References

Arkin, R.; Fujita, M.; Takagi, T.; and Hasegawa, R. 2003. An ethological and emotional basis for human-robot interaction. *Robotics & Autonomous Systems, special issue on Socially Interactive Robots* 42(3-4):191–201.

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

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

Breazeal, C. 2003. Emotion and sociable humanoid robots. *Int. Journal of Human Computer Studies* 119–155.

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

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

Cañamero, L., and Fredslund, J. 2001. I show you how I like you: Human-robot interaction through emotional expression and tactile stimulation. *IEEE Transactions on Systems, Man, and Cybernetics, Part A* 31(5):454–459.

Domínguez Quijada, S.; Zalama Casanova, E.; Gómez García-Bermejo, J.; and Perán González, J. 2002. Development of an expressive social robot. In Bradbeer, R.; ; and Billingsley, J., eds., *Mechatronics and Machine Vision 2002: Current Practice*. Research Studies Press. 341–348.

Esau, N.; Kleinjohann, B.; Kleinjohann, L.; and Stichling, D. 2003. MEXI: Machine with Emotionally eXtended Intelligence – A software architecture for behavior based handling of emotions and drives. In *Proc. of Int. Conf. on Hybrid and Intelligent Systems (HIS)*.

Fong, T.; Nourbakhsh, I.; and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics & Autonomous Systems, special issue on Socially Interactive Robots* 42(3-4):143–166.

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

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

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

Lang, S.; Kleinehagenbrock, M.; Hohenner, S.; Fritsch, J.; Fink, G.; and Sagerer, G. 2003. 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*.

Li, S.; Kleinehagenbrock, M.; Fritsch, J.; Wrede, B.; and Sagerer, G. 2004. "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)*.

Lienhard, R., and Maydt, J. 2002. 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)*.

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

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

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

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

Okuno, H.; Nakadai, K.; and Kitano, H. 2002. 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)*.

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

Scheeff, M.; Pinto, J.; Rahardja, K.; Snibbe, S.; and Tow, R. 2000. Experiences with Sparky, a social robot. In *Proc. of the Workshop on Interactive Robotics and Entertainment (WIRE)*.

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

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

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

Suzuki, K.; Camurri, A.; Ferrentino, P.; and Hashimoto, S. 1998. Intelligent agent system for human-robot interaction through artificial emotion. In *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*.

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

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

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

University of Bonn, I. f. K. u. P. 2000. Txt2pho – German TTS front end for the MBROLA synthesizer. http://www.ikp.uni-bonn.de/dt/forsch/phonetik/hadifi x/HADIFIXforMBROLA.html.

Viola, P., and Jones, M. 2001. 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)*.