Enabling a Humanoid Robot to Interact with Multiple Persons

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

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

Abstract

Showing attentiveness to people is an essential capability for a robot designed to interact with humans. Involving several persons into an interaction requires that the robot knows where the persons are relatively to its current location. Therefore, we propose an approach that maintains a probabilistic belief about people in the surroundings of the robot which is updated based on sensory information. 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 stored in the belief of the robot can attract its attention. As we show in practical experiments, our humanoid robot is able to shift its attention between different persons. Even people who are not the primary conversational partner are included into the interaction.

1 Introduction

Our goal is to develop a humanoid robot which performs multi-modal interaction with multiple persons simultaneously. This capability plays an important role whenever a robot has to interact with humans. One application in this context is an interactive museum tour-guide. Compared to previous museum tour-guide projects [1, 2], which focused on the autonomy of the robots and did not emphasize the interaction part that much, we want to build a robot which behaves and acts like a human. Over the last few years, humanoid robots have become very popular. The 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 [3, 4, 5, 6, 7]. Only little research has been done in the area of developing a robotic system which 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 which makes use of vision, sound source localization, and speech to detect, track, and involve people into interaction. In contrast to previous approaches [8, 9, 10], our goal is that the robot interacts with multiple persons and does not focus its attention on only one single person or use a strategy to 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. [3]). This results in a more human-like behavior, and the people feel better involved when they notice that the robot shows interest in them. The two images on the right of Figure 1 show our robot Alpha shifting its attention from one person to the other during a conversation.

This paper is organized as follows. The next section gives an overview over related work and Section 3 describes our robot. In Section 4, we present our technique to detect and keep track of people using visual information. In Section 5, we describe our dialogue system and explain how we perform speaker localization. In Section 6, we present our strategy on how to determine the gaze direction of the robot and how to decide which person gets the attention. Finally, in Section 7, we show experimental results.

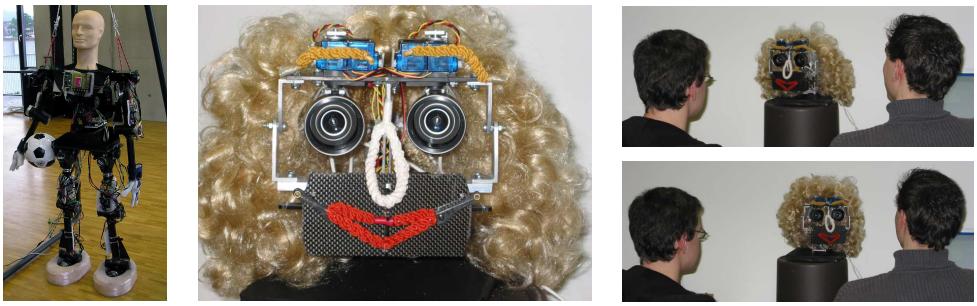

Figure 1: The body of our robot Alpha can be seen in the left image and its head is shown in the center image. The two images on the right show snapshots of a conversation of Alpha with two people. As can be seen, the robot shifts its attention from one person to the other to involve both into the conversation.

2 Related Work

Over the last few years, much research has been carried out in the area of multi-modal interaction. Lang et al. [8] presented an approach which combines several sources of information (laser, vision, and sound data) to track 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 et al. [10] 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. [9] 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. [3] only keeps track of objects which 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.

3 The Design of our Robot

The body (without the head) of our robot Alpha currently has of 21 degrees of freedom (six in each leg, three in each arm, and three in the trunk; see left image of Figure 1). Its total height is about $155cm$. The skeleton of the robot is constructed from carbon composite materials to achieve a low weight of about $38kg$. The hard skeleton is covered by soft materials to protect the robot in the case of a fall. To perform the experiments presented in this paper, we focus on the head of our robot which is shown in Figure 1 (center image). The head consists of 16 degrees of freedom which 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 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 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 [11] that supports a hierarchy of reactive behaviors. In this framework, behaviors are arranged in layers that work on different time scales.

4 Detecting and Tracking People Using Vision Data

Our robot maintains a probabilistic belief about people in its surroundings to deal with multiple persons appropriately. In this section, we describe how to sense 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 [12]. 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 [13] 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 and Elfes [14]. 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:

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

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 which 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 linear Kalman filter [15]. Applying such a filter leads to a smoothing of the estimated movements 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 [16] 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 2 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 6, 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.

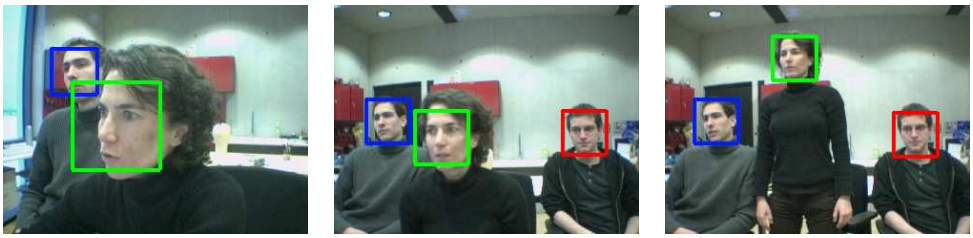

Figure 2: Tracking three faces with independent Kalman filters. To solve the data association problem we apply the Hungarian method.

5 Dialogue Management and Speaker Localization

For speech recognition, we currently use a commercial software [17]. 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/sentences which should be recognized needs to be defined. For speech synthesis, we use a freely available system [18] which 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/sentences is associated. This list is sent to the speech recognition system whenever the state of the automaton changes.

We also implemented a technique for sound source localization. We apply the Cross-Power Spectrum Phase Analysis [19] to calculate the spectral correlation measure $C_{lr}(t,\tau)$ between the left and the right microphone channel:

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

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 which 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 [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. In the experiments, we show how to use this information to shift the attention of the robot to a speaker.

6 Controlling the Gaze Direction and the Focus of Attention

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 which 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 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. 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 and to get new information about other people. Thus, we additionally implemented a behavior which 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 field of view.

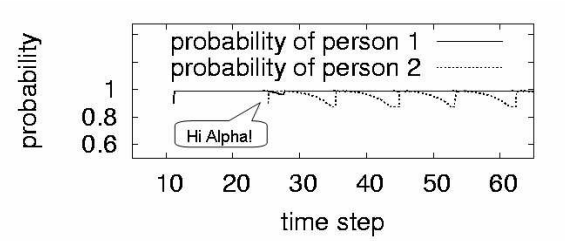

Figure 3: Evolution of the probabilities of two people. While the robot is chatting with person 1, it recognizes the voice of a second person and turns towards it at time step 25. As can be seen, person 2 is detected and the robot updates its belief. Afterwards, it continues talking to person 1 but also shows interest in person 2 by establishing short eye-contact and updates its belief at time steps 35, 45, 53, and 62.

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. To shift the attention of the robot temporarily to a person who is talking to it, we use our speaker localization system. In this way, the robot is also able to involve people outside its field of view who have not been detected before. As a result, the 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.

By looking around, the robot is able to detect faces of people who have not been stored in its belief before. Furthermore, if the robot gets too uncertain about whether or not a person who is outside its current field of view is still there, it looks 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

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

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

$$H(p(f)) = -p(f) \log p(f) - (1 - p(f)) \log(1 - p(f)). \tag{4}$$

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

7 Experimental Results

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. Using a camera resolution of 320×240 pixels, the face detection algorithm detects faces in a distance of approximately $30 - 200\,cm$. 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.

7.1 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 Figure 3. When the robot detected person 1 at time step 11, it started to interact with it. At time step 25 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. Since person 2 had a lower importance value (it was farther away), the robot continued its interaction with person 1. However, to involve person 2 into

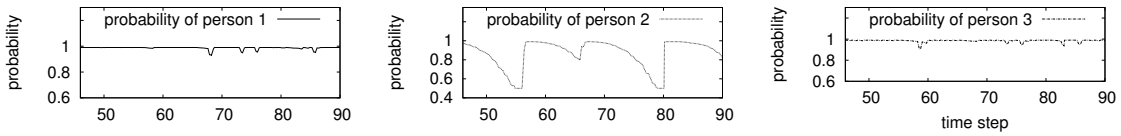

Figure 4: Evolution of the belief about three people. While persons 1 and 3 are initially in the field of view, person 2 is not. The robot concentrates its attention on persons 1 and 3, but to also show interest in person 2 and to update its belief, the robot looks to person 2 at time steps 56, 66, and 80.

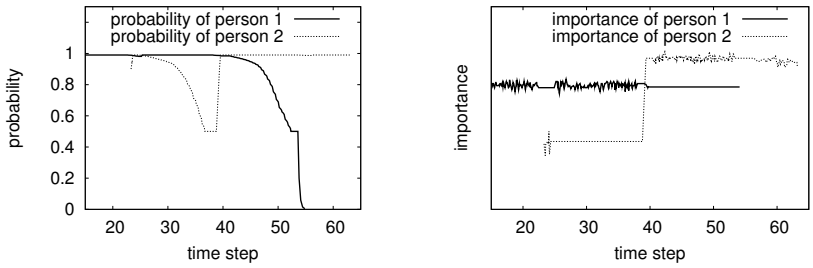

Figure 5: Evolution of the probabilities of two people (left image) and the corresponding importance values (right image). In the beginning, the robot is interacting with person 1. At time step 22, after an exploring gaze, the robot detects person 2. However, it continues concentrating on person 1. After looking to person 2 at time step 40, the robot shifts its attention to person 2 since the robot noticed that it had come very close and is thus 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.

the conversation as well 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 therefore stays constant for a short period of time. As can be seen from Figure 3, at time steps 35, 45, 53, and 62, the robot looked to person 2 and also updated its belief correctly. A similar experiment with three people, in which the robot concentrated its attention on two persons but also looked to a third person from time to time, is shown in Figure 4.

7.2 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 Figure 5 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. Since person 1 had a much higher importance value (it was closer; see the right image of Figure 5), the robot continued concentrating on 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. However, the robot kept person 2 in its belief. At time step 40, the robot decided to look to person 2 to check whether it was still there. The robot detected it, and since person 2 had come much closer, it got a higher importance than person 1. Thus, 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 into the direction of person 1, however, the person had gone and the robot updated its belief accordingly.

8 Conclusions

In this paper, we presented an approach to enable a humanoid robot to converse with multiple persons. We use a probabilistic technique to update a belief about people in the surroundings of our robot 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. 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 involve them into a conversation.

As a result, we obtain a human-like interaction behavior that shows attentiveness to multiple persons. This is valuable for several real world scenarios in which the robot needs to interact with humans. In practical experiments, we demonstrated our technique to reliably update the belief of our robot and to control its gaze direction.

Acknowledgment

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

References

[1] 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.

[2] 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.

[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] 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.

[6] 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.

[7] 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.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[18] 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.

[19] 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.