Audiovisual Person Tracking with a Mobile Robot

J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer
Bielefeld University, Faculty of Technology, Bielefeld, Germany
{jannik, mkleineh, slang, gernot, sagerer}@techfak.uni-bielefeld.de

Abstract. Mobile service robots are recently gaining increased attention from industry as they are envisaged as a future market. Such robots need natural interaction capabilities to allow unexperienced users to make use of these robots in home and office environments. In order to enable the interaction between humans and a robot, the detection and tracking of persons in the vicinity of the robot is necessary. In this paper we present a person tracking system for a mobile robot which enables the robot to track several people simultaneously. Our approach is based on a variety of input cues that are fused using a multi-modal anchoring framework. The sensors providing input data are two microphones for sound source localization, a pan-tilt camera for face and torso recognition, and a laser range finder for leg detection. Through processing camera images to extract the torso position, our robot can follow a person guiding the robot that is not oriented towards the robot and that is not speaking. In this case the torso information is especially important for robust tracking during temporary failure of the leg detection.

1 Introduction

Mobile service robots are fundamentally different from static setups used for research on human-machine-interfaces. In typical static setups, the presence and position of the user is known beforehand as the user either wears a close-talking microphone or stands at a designated position. On a mobile robot that operates in an environment where several people are moving around, it is often difficult for the robot to determine which of the persons in its vicinity wants to interact with it. In order to enable the robot to automatically recognize its instructor, it is necessary to develop techniques that allow to robustly track the persons in the robot's surrounding.

For person tracking a variety of cues can be used with some of them also applicable for the subsequent task of detecting whether the human instructor is currently interacting with the robot or with other persons nearby. In a scenario where one or several humans are in the vicinity of the robot, tracking is often accomplished using data from a laser range finder containing characteristic patterns resulting from the legs of the surrounding humans (see, e.g., [6, 7]). While this information is important for determining the appropriate velocity for following a specific human, it does not allow to infer whether a human is facing the robot or not. Two other cues that can be used to track a person are its face and its voice. Faces can be detected based on image data provided by a pan-tilt camera. Similarly, a stereo microphone allows to localize a sound source, i.e., the voice of a talking person. Both types of information represent additional position information while at the same time they can be used for attention control.

For our mobile robot BIRON – Bielefeld robot companion – we have realized a system integrating the three modalities described to perform not only tracking of persons [7] but also focusing of attention [9]. BIRON has already performed tracking and attention control successfully during several demonstrations, for example at the International Conference on Computer Vision Systems (ICVS) 2003 in Graz (see [9]).

However, the three different types of information that are fused for obtaining the position of a person are not always available. If the robot is required to follow a person guiding the robot, the face is usually not visible and no sound information is available. Therefore, the leg detection based on laser range data is the only available cue for tracking a non-speaking person that is not facing the robot. Consequently, a failure in the leg detection, e.g., due to an obstacle in the field of view of the laser range finder, will result in loosing the person. Therefore, additional cues are required to cope with such situations. In this paper we present a color-based torso recognition approach based on camera images that provides information about the direction of a person relative to the robot. Through adding a perceptual subsystem for creating a color model of the torso and for recognizing the torso in images, persons can be tracked more reliably.

The paper is organized as follows: At first we discuss related work on person tracking in section 2. Then, in section 3 our robot hardware is presented. Next, multi-modal person tracking is outlined in section 4. Our approach for recognizing the torso of a person is described in section 5 and section 6 presents a small evaluation of the overall system performing person tracking. The paper concludes with a short summary in section 7.

2 Related Work

As noted in the previous section, a mobile robot does not act in a closed or even controlled environment. A prototypical application is its use as a tour guide in scientific laboratories or museums (cf. e.g. [4]). All humans approaching or passing the robot have to be tracked in order to enable the robot to focus its attention on one person that intends to interact with the robot. Another application scenario becoming increasingly interesting is showing a robot around in a private home. This 'home tour' scenario is an interaction situation of fundamental importance as a human has to teach all the objects and places relevant for interaction to the robot. Interaction, however, requires the robot to be able to track all the humans in its vicinity.

For person tracking a variety of approaches have been developed which fuse different sensing modalities. Darrell et al. [5] integrate depth information, color segmentation, and face detection results for person tracking. The individual tracks are fused using simple rules. Feyrer and Zell [6] also track persons based on vision and laser range data. Here the two types of sensor data are fused using a potential field representation for the person positions.

Besides parallel fusion of different types of sensor data, some approaches perform sequential processing in a hierarchical architecture. After associating coarse position estimates, a smaller search space is used for processing more precise sensor data. For example, Schlegel et al. [12] propose vision-based person tracking that uses color information to restrict the image area that is processed in order to find the contour of a human. A more sophisticated method to realize a sequential search space reduction is proposed by Vermaak et al. [13]. In their approach sound and vision data are sequentially fused using particle filtering techniques. A related probabilistic fusion approach that applies graphical models for combining sound and vision data is presented by Beal et al. [3].

While all these person tracking approaches either use simple rules or learned/designed probabilistic relations to fuse different types of data, we will present in the following a person tracking approach that relies on a structured framework to track multiple persons simultaneously based on their individual 'components'. The main focus of this paper lies on the incorporation of the position of the human 'torso' through learning and tracking the color of the clothing of the upper body part of a person.

3 Robot Hardware

The hardware platform for BIRON is a Pioneer PeopleBot from ActivMedia (Fig. 1) with an on-board PC (Pentium III, 850 MHz) for controlling the motors and the on-board sensors and for sound processing. An additional PC (Pentium III, 500 MHz) inside the robot is used for image processing.

The two PC's running Linux are linked with a 100 Mbit Ethernet and the controller PC is equipped with wireless Ethernet to enable remote control of the mobile robot. For the interaction with a user a 12" touch screen display is provided on the robot.

A pan-tilt color camera (Sony EVI-D31) is mounted on top of the robot at a height of 141 cm for acquiring images of the upper body part of humans interacting with the robot. Two AKG far-field microphones which are usually used for hands free telephony are located at the front of the upper platform at a height of 106 cm, right below the touch screen display. The distance between the microphones is 28.1 cm. A SICK laser range finder is mounted at the front at a height of approximately 30 cm.

For robot navigation we use the ISR (Intelligent Service Robot) control software developed at the Center for Autonomous Systems, KTH, Stockholm [2].

Figure 1: BIRON.

4 Combining Multiple Modalities for Person Tracking

Person tracking with a mobile robot is a highly dynamic task. The sensory perception of persons is constantly changing as both the persons tracked and the robot itself might be moving. Another difficulty arises from the fact that a complex object like a person usually cannot be captured completely by a single sensor system alone. Therefore, we use the sensors presented in section 3 in order to obtain different percepts of a person:

- The camera is used to recognize faces and torsos. Our detection of faces (in frontal view) is based on the framework proposed by Viola and Jones [14]. This method allows to process images very rapidly and with high detection rates. From the face detection step the distance, direction, and height of the observed person are extracted, while an identification step provides the identity of the person if it is known to the system beforehand. Furthermore, the clothing of the upper body part of a person (the color of its torso) can be used to track this person, especially if it is not oriented towards the robot, i.e., its face is not visible for face detection. The torso recognition is described in detail in section 5.

- The stereo microphones are applied to locate sound sources using a method based on Cross-Powerspectrum Phase Analysis [8]. An extensive evaluation of our sound source localization has shown that the use of only one pair of microphones is sufficient for robust speaker localization within the multi-modal anchoring framework [9].
- The laser range finder is used to detect legs. In range readings pairs of legs of a human result in a characteristic pattern that can be easily detected [7]. From detected legs the distance and direction of the person relative to the robot can be extracted.

The percepts resulting from processing the data of these sensors provide information about the same overall object: the person. Consequently, the information about the individual percepts has to be fused. For combining the percepts from the different sensors we proposed *multi-modal anchoring* [7]. The goal of anchoring is defined as establishing connections between processes that work on the level of abstract representations of objects in the world (symbolic level) and processes that are responsible for the physical observation of these objects (sensory level). These connections, called *anchors*, must be dynamic, since the same symbol must be connected to new percepts every time a new observation of the corresponding object is acquired.

Our multi-modal anchoring framework allows to link the symbolic description of a complex object to different types of percepts, originating from different perceptual systems. It enables distributed anchoring of individual percepts from multiple modalities and copes with different spatio-temporal properties of the individual percepts. Every part of the complex object which is captured by one sensor is anchored by a single *component anchoring process*. The composition of all component anchors is realized by a *composite anchoring process* which establishes the connection between the symbolic description of the complex object and the percepts from the individual sensors. In the domain of person tracking the person itself is the composite object while its components are *face*, *torso*, *speech*, and *legs*, respectively.

The framework for anchoring the composite object *person* is depicted in Figure 2. It is based on anchoring the four components face, torso, speech, and legs. For more details on multi-modal anchoring please refer to [7].

Potentially, more than one person might be present in the vicinity of the robot. In order to track multiple composite objects simultaneously, the anchoring framework is extended by a so-called *supervising module*. This module manages all composite anchoring processes, e.g., it coordinates the assignment of percepts to the individual anchoring processes in order to cope with potential ambiguities. Moreover, the supervising module establishes new anchoring processes if percepts cannot be assigned to the existing ones. Contrary, anchoring processes are removed if no percepts were assigned for a certain period of time. More details and an evaluation of the robot tracking multiple persons in an office environment are provided in [7].

If several persons are tracked by the robot simultaneously, it must be able to recognize which of the persons is the robot's current communication partner. For this purpose, an attention system is provided, which enables the robot to focus its attention on the most interesting person. It is assumed that a communication partner is speaking and at the same time looking at the robot. Therefore, the pan-tilt camera is always turned towards the person which is currently speaking. Then, from the face detection process it can be determined whether the speaker is also looking at the robot and hence is considered the communication partner. For more details on the attention system see [9].

Besides evaluating the robot's performance in the lab, we successfully demonstrated its capabilities in public. In April 2003 our robot was presented at the exhibition part of the In-

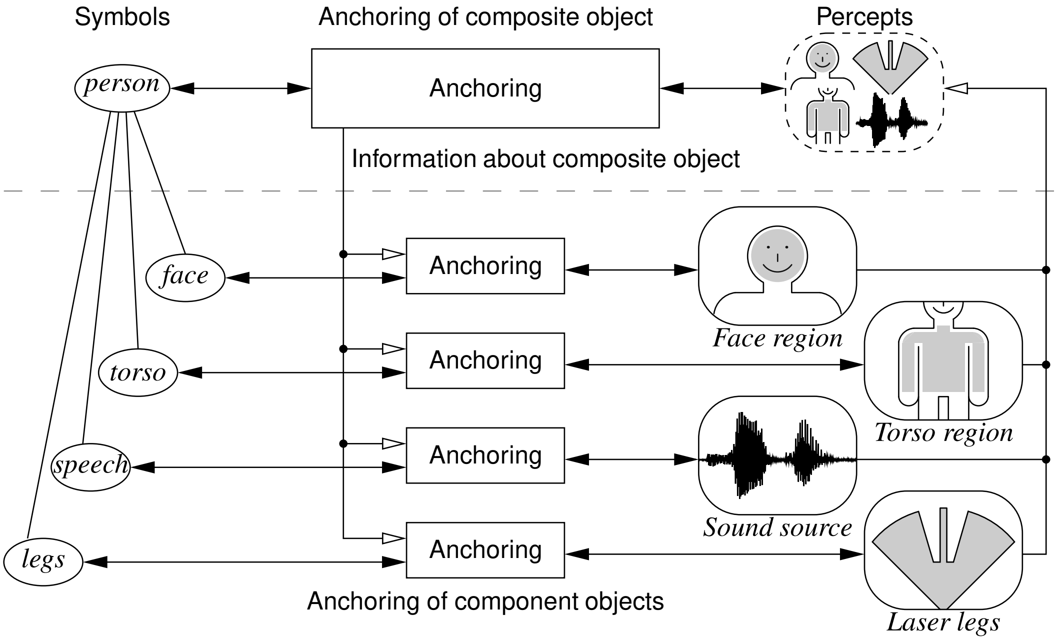

Figure 2: Multi-modal anchoring of persons.

ternational Conference on Computer Vision Systems (ICVS) 2003, Graz. There we demonstrated the robot's capabilities in multi-modal human-robot interaction, i.e., detection and tracking of multiple persons as well as detecting and eventually following communication partners (a video is available at [1]).

5 Color-based Torso Recognition

In order to supply the anchoring framework presented in section 4 with information about the torso position of an observed person, camera images of size 256×192 are analyzed to find an image area that matches a previously learned model of the torso color. Now the direction of a person relative to the robot can be extracted based on the camera orientation and the position of the torso within an image. A mixture of Gaussians is used to represent the torso color as such a parametric model is well-suited for efficient modeling of previously unknown color distributions. Previous applications using a mixture of Gaussians in order to track a coke can [10] or faces [11] under varying lighting conditions have demonstrated the flexibility of applying a parametric model.

For color representation we use the LUV color space [15] as it is perceptually uniform. This allows us to apply a homogeneous distance criterion for calculating the fit between an observed color and the color model. Initialization of the color model is performed after successfully detecting for the first time the face of the person that is being tracked. Using the anchoring framework, the distance of the person is known and a position 35 cm below the face position can be transformed into image coordinates. At this position in the image an elliptical image area is selected for creating the initial mixture model. The parameters of the individual components of the Gaussian mixture are calculated using a *k-means* clustering algorithm. For modeling a typical torso (the color of the clothing), a mixture with three components has been shown to provide good results. The number of mixture components has to be increased appropriately for colorful clothing which exhibits a large variety of different colors.

In order to distinguish the torso color described by the mixture model from the background, either a background model needs to be available or a suitable rejection criterion has to be defined. In our application the constantly varying background is unknown and, therefore, can not be described by some color distribution. Consequently, we use a rejection criterion based on an automatically determined threshold S_{class} on the probability density scores $p(x_i)$ obtained from the mixture model. First, we estimate a discrete distribution of the expected probability density values by calculating a histogram of all scores $p(x_i)$ for color vectors x_i contained in the current training set. The bins of the histogram thus represent a small range of possible scores. We then adjust the threshold S_{class} such that 98 % of all scores observed on the training data lie above S_{class}, i.e., the probability of observing a score greater than S_{class} is equal to 0.98:

$$Pr(Y > S_{\text{class}}) = 0.98, \qquad Y = p(x_i) \tag{1}$$

Choosing a fraction of 98 % of the training pixels has been determined empirically. In this way, outliers contained within the last 2 % of the training set are ignored.

Using this threshold, an image can be classified in torso and non-torso pixels. In order to remove isolated pixels from the resulting label image and to provide a more homogeneous result, a median of size 5×5 is applied to smooth the label image. Next, a connected components analysis is carried out to obtain the region segmentation result. Subsequently, polygonal descriptions of the image regions and region features like compactness, pixel count, center of mass, etc. are calculated. Using additional constraints with respect to the minimal and maximal size of the torso region, the result of this step is the segmented torso region R_{torso}. Based on the center of mass of this region and the current camera position, the direction of the person relative to the robot can be calculated.

As the mobile robot encounters varying lighting conditions while following a person, the torso color model has to be adapted to a changing visual appearance of the torso. For this purpose an image region has to be determined for updating the color model appropriately. Note that the shape of the torso area is not fixed due to, e.g., variations in the orientation of the person's body. Additionally, parts of the torso may be temporarily occluded by the skin-colored arms if the person wears clothes with short sleeves. Therefore, we have chosen to construct the update region R_{update} by enlarging the segmented region R_{torso}. The polygon describing R_{torso} is stretched to describe a region with an area 1.5 times the size of the segmented region. In this way image parts next to the currently segmented torso area are included in the update region R_{update}.

In order to avoid adapting the color model to background areas that are contained in the stretched update region, we use a training threshold S_{train} to select only those pixels from R_{update} that exhibit a color close to the current torso color model. Similar to the classification threshold S_{class} (see Eq. 1), the training threshold is calculated from the histogram of probability density values obtained on the previous training set. The threshold value is chosen such that the scores $p(x_i)$ of 99 % of the training pixels from the previous update step are above the threshold:

$$Pr(Y > S_{\text{train}}) = 0.99, \qquad Y = p(x_i) \tag{2}$$

Only those pixels in the current update area that have a probability value above the training threshold S_{train} are considered for updating the torso color model. In this way, we enforce a smooth adaptation of the color model. Note that this method to detect the torso based on its color is only applicable if no background objects exhibit a color similar to the color of the torso.

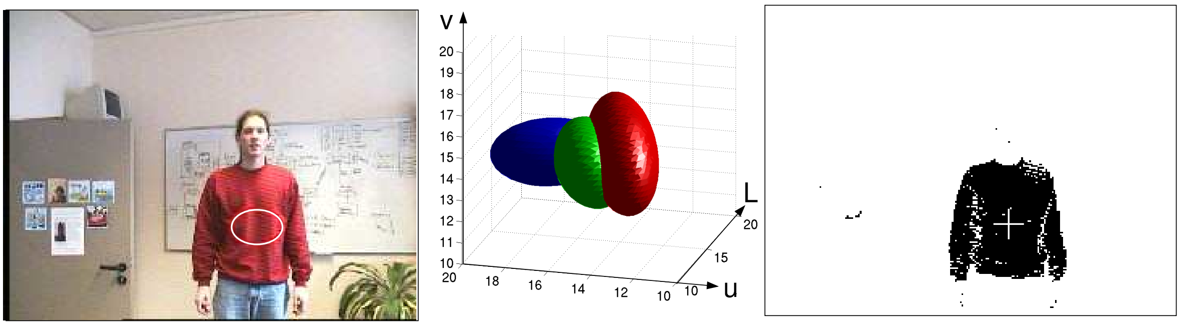

Figure 3: Input image, resulting torso color mixture model, and segmentation result.

An example for the processing of input images for torso recognition is depicted in Fig. 3. Based on the face position, the anchoring framework provides the image position of the torso. An elliptical image area around this position indicated by a white ellipse is used for training the color mixture model. In the center image the resulting mixtures are depicted. The right image shows the resulting segmentation result before median filtering with the center of mass selected later as torso position indicated by a white cross.

The processing of the input images is performed on the 500 MHz Pentium III computer at a frame rate of 5 Hz. In every step the image is analyzed with the face detection algorithm consuming about 20 % of the processing time and the torso color segmentation as well as the updating of the color model are carried out.

6 System Performance

With the integration of the perceptual system for torso recognition, the person tracking has become much more robust. Previously, when a person was guiding the robot to another place without facing the robot, i.e., not walking backwards, the person was only followed based on detected legs. If the legs were not detectable for several processing cycles, the robot would have lost the person. After integrating the torso information, the robot can now track persons more reliably even if they are not facing the robot.

Another more important scenario are situations in which the legs of a person are partially occluded by, e.g., furniture. In order to demonstrate the performance of our system, we have chosen a setup in our office room as shown in Figure 4. The robot (R) observes the door of the room and is allowed to steer the camera and to rotate its base. A person (P) is supposed to enter the room and to verbally instruct the robot to track the person. Then, the person approaches a desk in order to interact with an object (O). When passing the flower (F) which is located on the floor, the legs of the person are not detectable for the robot anymore as the flower and the cupboard next to it occlude the lower body part of the person. Because the person is turned away from the robot when interacting with object O, neither the legs nor the face of the person are observable by the robot. Moreover, the person is assumed not to speak, which prevents the robot from acquiring any speech percepts. Therefore, only the torso percepts are available for tracking. After interacting with the object, the person returns to the position next to the door. If the robot was able to successfully track the person based on torso percepts then the robot is still focusing on the person.

We carried out 25 runs with several subjects wearing different shirts and sweaters. In 80 % of the test cases the person was correctly tracked throughout the interaction with the

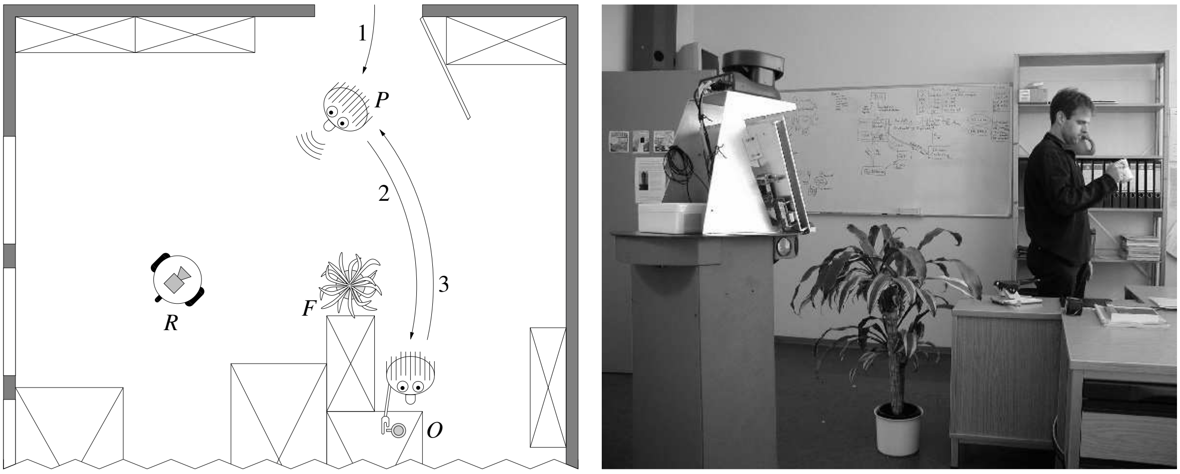

Figure 4: Setup for evaluation. Left: top view. Right: view from behind the robot

object. During the interaction phase lasting five to ten seconds only the *torso* percepts were available as the legs and the face were not visible and the person was not speaking. Without incorporation of the torso information the robot would have lost the person in all cases.

The processing speed, i.e., the rate at which percepts are fused to obtain the overall person position, depends on the rate at which the individual percepts are provided. Face recognition and torso detection are performed on images of a size of 256×192 at a rate of 5.0 Hz. Localization of sound sources runs at a rate of 5.5 Hz. The laser range finder provides new data at a rate of 4.7 Hz while the processing time for the detection of legs is negligible. The anchoring processes of the persons are updated asynchronously as percepts become available.

Although the torso information substantially improves tracking in this scenario, the tracking algorithm relies on just a single cue for a considerable period of time. Therefore, it is desirable to have more perceptual cues available for robust tracking in a wider variety of situational contexts. This can be achieved by complementing the existing system with detectors for, e.g., the human head-shoulder contour or faces in non-frontal view.

7 Summary

In this paper we presented a multi-modal person tracking system for our mobile robot BIRON. The described approach applies audiovisual cues in addition to laser range data. In previous work we demonstrated that the system is able to simultaneously track multiple persons in the vicinity of the robot based on face recognition, speech localization, and leg detection. In this paper we introduced an additional perceptual sub-system for extracting information about the torso of a tracked person. Based on the color of the clothes worn by the human, the torso is tracked with an adaptive color segmentation approach. The torso helps to track a person when other cues are not available, e.g., due to visual occlusion of the legs. Results of a small evaluation in an office environment demonstrate the increase in robustness resulting from the incorporation of the torso detection in our framework for tracking persons.

8 Acknowledgments

This work has been supported by the German Research Foundation within the Collaborative Research Center 'Situated Artificial Communicators' and the Graduate Programs 'Task Oriented Communication' and 'Strategies and Optimization of Behavior'.

References

[1] http://www.techfak.uni-bielefeld.de/ags/ai/projects/BIRON/.

[2] M. Andersson, A. Orebäck, M. Lindstrom, and H. I. Christensen. ISR: An intelligent service robot. In H. I. Christensen, H. Bunke, and H. Noltmeier, editors, *Sensor Based Intelligent Robots; International Workshop Dagstuhl Castle, Germany, September/October 1998, Selected Papers*, volume 1724 of *Lecture Notes in Computer Science*, pages 287–310. Springer, New York, 1999.

[3] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[4] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. The interactive museum tour-guide robot. In *Proc. Nat. Conf. on Artificial Intelligence*, pages 11–18, 1998.

[5] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *Int. Journal of Computer Vision*, 37(2):175–185, 2000.

[6] S. Feyrer and A. Zell. Robust real-time pursuit of persons with a mobile robot using multisensor fusion. In *Proc. Int. Conf. on Intelligent Autonomous Systems*, pages 710–715, Venice, 2000.

[7] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.

[8] D. Giuliani, M. Omologo, and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1243–1246, Yokohama, Japan, 1994.

[9] 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. Int. Conf. on Multimodal Interfaces*, pages 28–35, Vancouver, Canada, November 2003. ACM.

[10] S. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17:225–231, March 1999.

[11] N. Oliver, A. Pentland, and F. Berard. LAFTER: a real-time face and lips tracker with facial expression recognition. *Pattern Recognition*, 33:1369–1382, 2000.

[12] C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Wörz. Vision based person tracking with a mobile robot. In *Proc. British Machine Vision Conference*, pages 418–427, Southampton, UK, 1998.

[13] J. Vermaak, A. Blake, M. Gangnet, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. Int. Conf. on Computer Vision*, volume 1, pages 741–746, 2001.

[14] P. Viola and M. Jones. Robust real-time object detection. In *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001.

[15] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons, New York, 1982.