in: Proc. of the 2nd European Conference on Mobile Robots (ECMR 2005), Ancona, Italy, pp. 176-181, stampalibri 2005 (ISBN 88-89177-187)

SENSOR FUSION USING A PROBABILISTIC AGGREGATION SCHEME FOR PEOPLE DETECTION AND TRACKING *

Christian Martin, Erik Schaffernicht, Andrea Scheidig, and Horst-Michael Gross

Department of Neuroinformatics and Cognitive Robotics
Ilmenau Technical University
christian.martin@tu-ilmenau.de

ABSTRACT

Stable methods for people detection and tracking are fundamental features when dealing with methods of human-robot interaction in the context of a real mobile robot. In this paper, we discuss a new approach for integrating several sensor modalities and we present a multimodal people detection and tracking system and its application using the different sensory systems of our mobile interaction robot HOROS working in a real office environment. These include a laser-range-finder, a sonar system, and a fisheye-based omnidirectional camera. For each of these sensory information, a separate Gaussian probability distribution is generated to model the belief of the observation of a person. These probability distributions are further combined using a flexible probabilistic aggregation scheme. The main advantages of this approach are a simple integration of further sensory channels, even with different update frequencies and the usability in real-world environments. Finally, promising experimental results achieved in a real office environment will be presented.

1. INTRODUCTION

Dealing with Human-Robot-Interaction (HRI) in real-world environments, one of the general tasks is the realization of a stable people detection and the respective tracking functions. Depending on the specific application that integrates a person detection, different approaches are possible. Typical approaches use visual cues for face detection, a laser-range-finder for detection of moving objects, like legs, or acoustical cues for sound source detection.

Projects like EMBASSI [1], which aim to detect only the users' faces, usually in front of a stationary station like a PC, typically use visual cues (skin-color-based approaches, sometimes in combination with the detection of edge oriented features). Therefore, these approaches cannot be applied for a mobile robot which has to deal with moving people with faces not always perceivable. In [2] a skin-color-based approach for a mobile robot is presented using an extension of particle filters to generate object configurations which represent more then one person in the image.

Other approaches,e.g. TOURBOT [3] or GRACE [4], trying to perceive the whole person rather than only her face use laser-range-finders to detect people as moving objects or directly by their legs.

Drawbacks of these approaches occur, for instance, in situations where a person stands near a wall and cannot be distinguished, in scenarios with objects yielding leg-like scans, like table-legs or chair-legs, or if the laser-range-finder does not cover 360 degrees of the robot space.

For real-world scenarios, more promising approaches combine more than one sensory channel,like visual cues and the scan of the laser-range-finder. An example for these approaches is the SIG robot [5], which combines visual and auditory cues. People are detected by a face detection system and tracked by using stereo vision and sound source detection. This approach is especially useful for scenarios like face-to-face interaction. Further examples are the EXPO-ROBOTS [6], where people are detected as moving objects by a laser-range-finder (resulting from differences from a given static environment map) firstly. After that, these hypotheses are verified by visual cues. Other projects like BIRON [7] detect people by using the laser-range-finder for detecting leg-profiles and combine these information with visual and auditory cues (anchoring). The essential drawback of these approaches is the sequential processing of the sensory cues. People are detected by laser information only and are subsequently verified by visual cues. These approaches fail, if the laser-range-finder yields no information, for instance, in situations when only the face of a person is perceivable because of leg occlusions.

Therefore, we propose a multimodal approach to realize a robust detection and tracking of people. Compared to other approaches, all used sensory cues are concurrently processed using a probabilistic aggregation scheme. The overall computational complexity our approach scales very well with the number of sensors and modalities. This way

*THIS WORK IS PARTIALLY SUPPORTED BY TMWFK-GRANT #B509-03007 TO H.-M. GROSS AND A HWP-GRANT TO A. SCHEIDIG

people are not only detected by only one feature. They can be detected by their legs and their faces or by only one of this features, respectively. The main advantage of our approach is the simple expandability by integrating further sensory channels, like sound sources, because of the used aggregation scheme.

As sensory channels we use the different sensory modalities of HOROS: the omnidirectional camera, the sonar sensors, and the laser-range-finder (see section 2). Using these modalities, we generate specific probability-based hypotheses about detected people and combine these probability distributions by *Covariance Intersection* in the aggregation scheme (see section 3). Respective results will be shown in section 4 followed by a short summary and outlook in section 5.

2. THE ROBOT SYSTEM HOROS

To investigate respective methods, we use the mobile interaction robot HOROS as an information system for employees, students, and guests of our institute. The system's task includes that HOROS autonomously moves in the institute, detects people as possible interaction partners and interacts with them, for example, to answer questions like the current whereabouts of specific people.

HOROS hardware pletform is an extended Pioneer-based robot from ActiveMedia. It integrates an on-board PC (Pentium M, 1.6 GHz, 512MB) and is equipped with a laser-range-finder (SICK) and sonar sensors. For the purpose of HRI, this platform was mounted with different interaction-oriented modalities (see Figure 1).

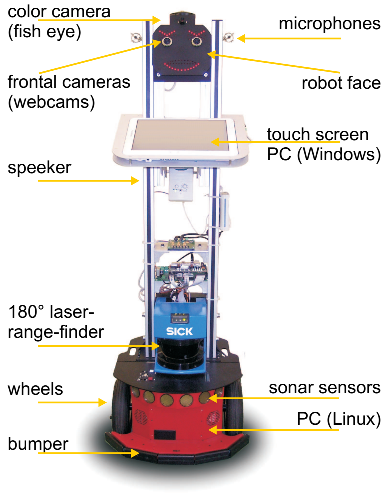

Fig. 1. Sensory and motory modalities of the mobile interaction robot HOROS (HOme RObot System).

This includes a tablet PC for speech recognition, speech generation, and pen-based interaction. It was further extended by a robot face which integrates an omnidirectional fisheye camera, two microphones, and two frontal webcams for the analysis of the user features.

Subsequently, the laser-range-finder, the sonar sensors, and the omnidirectional camera are discussed in the context of a person detection.

Laser-based Information: The laser-range-finder is a very precise sensor with a resolution of one degree, perceiving the frontal 180 degree field of HOROS (see Figure 2 upper left). It is fixed on the robot approximately 30 cm above the ground. Therefore it can only perceive the legs of people (see Figure 2 bottom left).

Based on the approach presented in [8], we also analyze the scan of the laser-range-finder for leg-pairs using a heuristic method. The measurements are segmented into local groups of similar distance values. Then each segment is checked for different conditions like width, deviation and others that are common for legs. The distance between segments classified as legs is pairwise computed to determine whether this could be a human pair of legs. For each pair found, the distance and direction is extracted.

This yields very good results for distances of people which stand less than 3 meters away. In a greater distance legs will be missed due to the limited resolution of the laser-range-finder. The gaps between single rays become larger than the width of legs. The strongest disadvantage of this approach is its false-positive classification detecting table-legs, chair-legs and also waste-paper baskets as legs. Also people standing sideway to the robot or are wearing long skirts do not yield appropriate values of the laser-range sensor to detect their legs.

Sonar-based Information: HOROS has 16 sonar sensors, arranged at the Pioneer platform approximately 20 cm above the ground. Because of this, a person detection using the sonar sensors does only work by analyzing the sonar scan for leg profiles (see Figure 2, middle).

The disadvantage of these sonar sensors is their high inaccuracy. The measurement depends not only on the distance to an object, but also on the objects material, the direction of the reflecting surface, crosstalk effects when using several sonar sensors and the absorption of the broadcasted sound. Because of these disadvantages, only distances of at least 2 meters can be considered for person detection using these sonar sensors (see Figure 2, middle). This means the sonar sensors yield pretty unreliable and inaccurate values, a fact which has to be considered in the generation of a hypothesis of a person detection. For the purpose of a very simple person detection, we assume that all measurements less than 2 meters could be hypotheses for a person. These hypotheses could be further refined by comparing the position of the hypothesis with a map of the environment. If the hypothesis would correspond to an obstacle in the map,

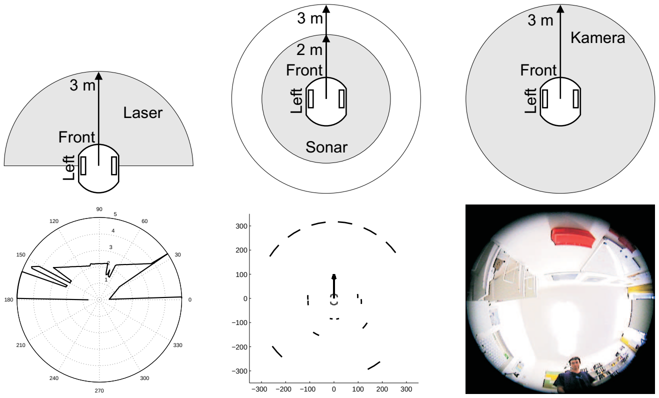

Fig. 2. **Top:** Top view of the schematic HOROS. The sensory range used to detect people is depicted grey. **Bottom:** Exemplary sensory inputs from laser-range-finder, sonar, and fisheye camera.

it could be dismissed. The disadvantage of this refinement is, that people standing near by an obstacle would not be considered as a valid hypothesis.

Fisheye Camera: For HOROS we use an omnidirectional camera with a fisheye lens yielding a 360 degree view around the robot (see Figure 2 upper right). Because of the task of person detection, the usage of such a camera requires that the position of the camera is lower than the position of the faces. An example of an image resulting from the camera is depicted in Figure 2 bottom right.

To detect people in the omnidirectional camera image a skin-color-based multi-target-tracker [9] is used. This tracking system is based on the condensation algorithm [10]. It has been extended to allow the visual tracking of multiple people at the same time. The particle clouds used to estimate the probability of people in the omnidirectional image will concentrate on the different skin-colored objects. A problem is the possible tracking of skin-color-based objects like wooden objects or cork pinboards. The used approach for person detection is much faster than subsampling the whole image trying to find regions of interest and is resistant to minor interferences.

A person detection using omnidirectional camera images yields hypotheses about the direction of a person but not about the distance. Therefore, the integration of the information from the camera with the information from the laser-range-finder and the sonar sensors results in a more powerful person detection system. Subsequently the developed method for the combination of the sensory systems will be discussed.

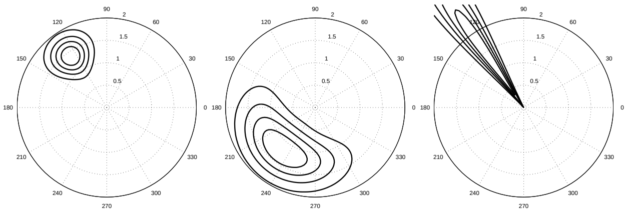

Fig. 3. Examples for generated hypotheses. The center of each plot represents the robot. **Left:** Hypothesis generated by laser showing a person left in front of the robot. **Middle:** Sonar-based information showing a hypothesis behind the robot. **Right:** Camera-based hypothesis without depth information showing the same person as in the left figure.

3. GENERATION OF USER MODELS

3.1. User Modeling Using Different Sensor Information

For the purpose of tracking the information about detected humans is converted into Gaussian distributions $\phi(\mu, C)$. The mean μ equals the position of the detection in polar coordinates and the covariance matrix C represents the uncertainty about this position. The form of the covariance matrix is sensor-dependent due to the different sensor characteristics described in section 2 and is computed based on a heuristic. Furthermore, the sensors have different error rates of misdetections that have to be taken into account. All computation is done in the cartesian r, φ space. Examples for the resulting distributions are shown in Fig. 3.

Laser-based Information: The laser-range-finder is a very precise measure, hence the corresponding covariances are small and the distribution is narrow (see Figure 3 left). The radial variance is fixed for all possible positions, but the variance of the angular coordinate is distance dependent. A sideways step of a person standing directly in front of the robot changes the angle more than the same movement in a distance of 2 meters. The smaller the distance of the detection the larger the variance has to be. The probability of a misdetection is the lowest of the used sensors, but the laser-range-finder only covers the front area of the robot due to sensor arrangement, so hypotheses in the back are ignored when processing laser-based information.

Sonar Information: Information from the sonar tends to be very noisy, imprecise und unreliable. Therefore, the variances are large and the impact on the certainty of a hypothesis is minimal (see Figure 3 middle). Nevertheless, the sonar is included to support people tracking behind the robot. So we are able to form an estimate of the distance in vision-based hypotheses.

Fisheye Camera: In contrast to the other sensors, the camera can only provide information about the angle of a detection, but not about the distance of a person. Therefore, for the radial variance of the distance coordinate a very large value was selected, with a fixed mean value (see Figure 3 right). The angular variance is determined by the angular variance of the particle distribution used in the visual skin-color based person tracker (see Section 2). The information content of a detection in the image of the fisheye camera is controlled by the position of the detection. In the front area of the robot, the influence is lower, because of the available laser as reliable sensor. Behind the robot, the image is the only source to get information about the presence of a person, the sonar has only supporting character. Thus, the relative weight of a visual hypothesis should be higher behind the robot.

The modeling and integration of additional sensorical cues, like sound localization or other features from the camera image, could be done in a similar way as described above.

3.2. Multi-Hypotheses Tracking Using Covariance Intersection

Tracking based on probabilistic methods attempts to improve the estimate x_t of the position of the people at time t. These estimates x_t are part of a local map M that contains all hypotheses around the robot. This map is used to aggregate the sensor hypotheses. Therefore, the movements of the robot $\{u_1, ..., u_t\}$ and the observations about humans $\{z_1, ..., z_t\}$ have to be taken into account. In other words, the posterior $p(x_t|u_1, z_1, ..., u_t, z_t)$ is estimated. The whole process is assumed to be Markovian. So, the probability can be computed from the previous state probability $p(x_{t-1})$, the last executed action u_t and the current observation z_t. The posterior is simplified to $p(x_t|u_t, z_t)$. After applying the Bayes rule, we get

$$p(x_t|u_t, z_t) \propto p(z_t|x_t)p(x_t|u_t) \tag{1}$$

where $p(x_t|u_t)$ can be updated from $p(x_{t-1}|u_{t-1}, z_{t-1})$ using the motion model of the robot and the assumptions about the typical movements of people.

A Gaussian mixture $M = \{\mu_i, C_i, w_i | i \in [1, n]\}$ is used to represent the positions of people, where each Gaussian is the estimate for one person. $\phi_i(\mu_i, C_i)$ is a Gaussian centered at μ_i and the covariance matrix C_i. The weight w_i $(0 < w_i \leq 1)$ contains information about the contribution of the corresponding Gaussian.

Next, the current sensor readings z_t have to be integrated, after they have been preprocessed as described earlier. If M does not contains any element at time t, all generated hypotheses from z_t are copied to M. Otherwise data association has to be done to determine which elements from z_t and M refer to the same hypothesis. The Mahalanobis distance d_m between two Gaussians $\phi_i \in z_t$ and $\phi_j \in M$ is used as association criterion.

$$\begin{aligned} \mu &= \mu_i - \mu_j \\ C &= C_i + C_j \\ d_m &= \mu C^{-1} \mu^T \end{aligned} \tag{2}$$

This distance is compared to a threshold. As long as there are distances lower than the threshold, the sensor reading i and the hypothesis j with the minimum distance are merged. The problem of merging hypotheses in case two people pass near each other has to be tackled seperatly. Update is done using the *Covariance Intersection* rule [11]. This technique does not need any information about the correlation between the hypotheses, unlike a *Kalman filter*. The covariances are transformed into the so called *information space* by computing the respective inverses. Then the matrices are combined using a weighted linear combination and propagated to the original space. The new mean is computed with respect to the *information space*.

$$\begin{aligned} C_{new}^{-1} &= (1-\omega)C_i^{-1} + \omega C_j^{-1} \\ \mu_{new}^{-1} &= C_{new}\left[(1-\omega)C_i^{-1}\mu_i + \omega C_j^{-1}\mu_j\right] \end{aligned} \tag{3}$$

The weight ω is chosen as

$$\omega = \frac{|C_i|}{|C_i| + |C_j|} \tag{4}$$

The criterion for the weight ω is to minimize the resulting determinant by prefering the sharper distribution in the intersection process. A very unreliable sensor input will have only minimal influence on the resulting hypothesis.

Sensor readings not matched with a hypothesis of M are introduced as new hypothesis in M. The weight w_i is representing the certainty of the corresponding Gaussian. The more sensors support this hypothesis, the higher this weight should be. If the weight passes a threshold, the corresponding hypothesis is considered to be a person. The weight is increased as

$$w_i(t+1) = w_i(t) + \alpha(1 - w_i(t)), \tag{5}$$

if that hypothesis has been matched with a sensor reading. The constant $\alpha \in [0, 1]$ is chosen with respect to the current sensor (see section 3.1). The more reliable the sensor, the higher the α-weight is. These values were determined based on experiments. In the case of an unmatched hypothesis the weight is decreased.

$$w_i(t+1) = w_i(t) - (1-\theta)\frac{t_{new} - t_{old}}{t_v} \tag{6}$$

The term t_{new} is the current point of time and t_{old} the moment the last sensory input was processed. A person is considered to be lost if t_v seconds passed and no sensor has made a new detection that can be associated with this hypothesis. This temporal control regime is sensor dependent, too. Hypotheses with a weight lower than the threshold θ are deleted.

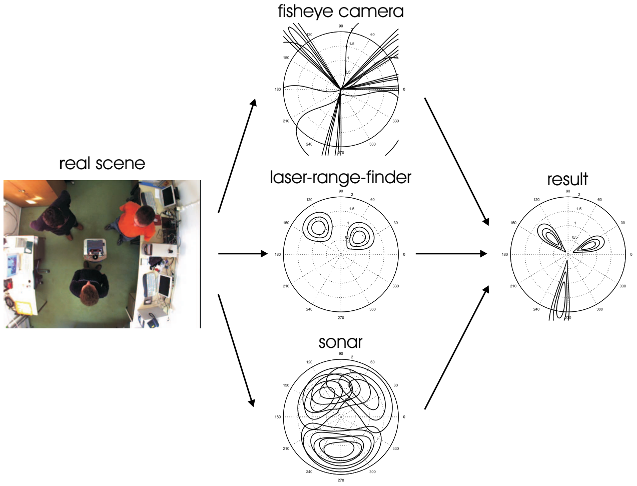

Fig. 4. **Aggregation example.** The left picture shows the real office scene from a bird's eye view. Three people are surrounding the robot, who stands in the middle. The three figures in the middle row show the current hypotheses generated by fisheye camera, laser-range-finder, and sonar from top to bottom. No sensor on its own can represent the scene correctly. The final picture displays the aggregated result from the sensors and the previous timestep. This is a correct and sharpened representation of the current situation.

4. EXPERIMENTS AND RESULTS

The presented system is in use on the HOROS robot in a real-world office environment. The fact of a changing illumination in different rooms and numerous distractions in form of chairs and tables is quite challenging.

Figure 4 shows a typical aggregation example. In this experiment, the robot was standing in the middle of an office room and did not move. Up to three people were moving around the robot. The enviroment contained several distracting objects, like table legs and skin-colored objects. No sensor modality was able to detect the people correctly. Only aggregation over sensors modalities and time led to the proper result.

The system was able to track multiple people correctly with an accuracy of 93% in the experiment. In most cases false negative detections occured behind the robot. The rate of false positive detections is higher, about every forth hypothesis was a misdetection. This is due to the simple cues integrated into the system. But for the intended task of HOROS, the interactive office robot, it is considered to be more important not to miss to many people than finding to many. But there are ways to reduce the amount of false positive detections. Most misdetections are static in the environment, so based on the movement trajectories created by the tracker they can be identified.

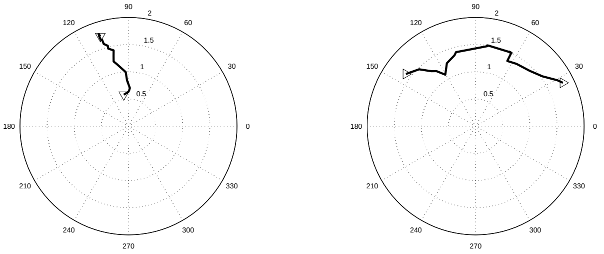

Fig. 5. **Left:** A trajectory showing a person coming straight towards the robot. **Right:** The person is crossing from left to the right. In doing so the robot is avoided. The varying time intervals between the movements and the associated weights aren't visible in the figure.

The presented system improved the performance in the area behind the robot only slightly compared to a simple skin-color tracker. This is, because the sonar-based sensors do not provide many useful information for the tracking task. The main contribution of the sonar sensors is the addition of distance information to existing hypotheses extracted from the fisheye camera and preventing a precipitate extinction of hypotheses in cases of sudden changes in the illumination. In this case, the skin-color tracker will presumably fail, but if the sonar-based information still confirms the presence of the person at the respective position, the hypothesis will not be deleted until the skin-color tracker has recovered.

In the front area of the robot, the system clearly outperforms single sensor-based tracking. Here the influence of the sonar on the result is not observable, because in most cases the laser-range-finder generates hypotheses more precisely. The laser reduces the deficiency of the skin-color tracker, while the skin-color based information compensates the shortcomings of the laser. These results are observable in Fig. 4. This leads to the assumption that the inclusion of additional sensory systems generating hypotheses about people (e.g. sound source hypotheses) will further improve the performance of this tracking system.

The system was practically tested in the context of a survey task. HOROS was standing in a hallway in our institute building. His task was to attract attention of people that came by. As soon as the system recognized a person near him, the robot addressed the visitor to come nearer. He then offered to participate in a survey about desired future functionality of HOROS. The people tracking module was used to detect break offs, thus if the user was leaving before finishing the survey. The robot tried to fetch them back and finalize the survey. After the successfull completion of the interaction or a defined time interval with no person coming back, the cycle began again with HOROS waiting for the next interaction partner. The experiment was made in the absence of any visible staff members, so the people could interact more unbiased.

These efforts are repeated from time to time to gather more information, and there is a second, not obvious, intention. The tracking module was used to generate typical movement trajectories of the users. In our future work, we will attempt to classify the path of movement to gain more knowledge about the potential user. In the context of adaptive robot behavior and user models, it is an important issue to assess the interaction partner. The users' movements and the positions relative to the robot are a fundamental step in this direction. If the robot can distinguish between people with different goals, an appropriate reaction can be learned. The use of a multi-person-tracker is a prerequisite, since the experiments show visitors often appearing in groups of two or more people. Examples for different trajectories are shown in Fig. 5. The most challenging aspects for a classification of trajectories are in our opinion the varying speed of the people and the search for typical movement schemes describing the interest of potential users. Based on the trajectories longtime immovable hypotheses can be discarded with respect to position and interaction status as a false detection.

5. SUMMARY AND OUTLOOK

We presented a flexible multimodal probability-based approach for detecting and tracking people. It is implemented on our mobile office robot HOROS and is working in realtime. Because of the sensor fusion and the probabilistic aggregation, its results are significantly improved compared to a single sensor tracking system. It can be easily extended with other sensors.

In our future work, we will extend the system with additional cues to further increase robustness and reliability for real-world environments. Currently, we are working on the integration of an audio-based speaker localization. In addition, it will be investigated if a face detector could be integrated into the aggregation scheme as an additional cue. Furthermore, we will study the behavior of our system compared to other known approaches and investigate the localization accuracy using labeled data of reference movement trajectories.

6. REFERENCES

[1] B. Froeba and C. Kueblbeck, "Real-time face detection using edge-orientation matching," in *Audio- and Video-based Biometric Person Authentication (AVBPA'2001)*, 2001, pp. 78–83.

[2] C. Martin, H.-J. Boehme, and H.-M. Gross, "Conception and realization of a multi-sensory interactive mobile office guide," in *IEEE Conference on Systems, Man and Cybernetics*, 2004, pp. 5368–5373.

[3] D. Schulz, W. Burgard, D. Fox, and A. Cremers, "Tracking multiple moving objects with a mobile robot," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 371–377.

[4] R. Simmons, D. Goldberg, A. Goode, M. Montemerlo, N. Roy, B. Sellner, C. Urmson, A. Schultz, M. Abramson, W. Adams, A. Atrash, M. Bugajska, M. Coblenz, M. MacMahon, D. Perzanowski, I. Horswill, R. Zubek, D. Kortenkamp, B. Wolfe, T. Milam, and B. Maxwell, "Grace: An autonomouse robot for AAAI robot challenge," *AAAI Magazine*, vol. 24, no. 2, pp. 51–72, Summer 2003.

[5] K. Nakadai, H.G. Okuno, and H. Kitano, "Auditory fovea based speech separation and its application to dialog system," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002)*, 2002, vol. 2, pp. 1320–1325.

[6] 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," *Special issue on Socially Interactive Robots, Robotics and Autonomous Systems*, vol. 42, pp. 203–222, 2003.

[7] J. Fritsch, M. Kleinehagenbrock, S. Lang, G.A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *Int. Conf. on Intelligent Autonomous Systems*, F. Groen, N. Amato, A. Bonarini, E. Yoshida, and B. Krse, Eds. March 2004, pp. 898–906, IOS Press.

[8] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Ploetz, 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*, vol. 43, no. 2-3, pp. 133–147, 2003.

[9] T. Wilhelm, H.-J. Boehme, and H.-M. Gross, "A multi-modal system for tracking and analyzing faces on a mobile robot," in *Robotics and Autonomous Systems*, 2004, vol. 48, pp. 31–40.

[10] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal on Computer Vision*, vol. 29, pp. 5–28, 1998.

[11] S. Julier and J. Uhlmann, "A nondivergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the 1997 American Control Conference*. June 1997, pp. 2369–2373 vol.4, IEEE.