Available online at www.sciencedirect.com

ScienceDirect

Robotics and Autonomous Systems 54 (2006) 721–728

Robotics and Autonomous Systems

www.elsevier.com/locate/robot

Multi-modal sensor fusion using a probabilistic aggregation scheme for people detection and tracking[☆]

C. Martin*, E. Schaffernicht, A. Scheidig, H.-M. Gross

Department of Neuroinformatics and Cognitive Robotics, Ilmenau Technical University, 98694 Ilmenau, Germany

Received 19 December 2005; received in revised form 5 March 2006; accepted 6 April 2006
Available online 14 July 2006

Abstract

Efficient and robust techniques for people detection and tracking are basic prerequisites when dealing with Human–Robot Interaction (HRI) in real-world scenarios. In this paper, we introduce a new approach for the integration of several sensor modalities and present a multi-modal, probability-based people detection and tracking system and its application using the different sensory systems of our mobile interaction robot HOROS. These include a laser range-finder, a sonar system, and a fisheye-based omni-directional camera. For each of these sensory systems, separate and specific Gaussian probability distributions are generated to model the belief in observing one or several persons. These probability distributions are further merged into a robot-centered map by means of a flexible probabilistic aggregation scheme based on Covariance Intersection (CI). The main advantages of this approach are the simple extensibility by the integration of further sensory channels, even with different update frequencies, and the usability in real-world HRI tasks. Finally, the first promising experimental results achieved for people detection and tracking in a real-world environment (our institute building) are presented.

Keywords: Sensor fusion; People detection; People tracking

1. Introduction

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

Projects such as EMBASSI [1], which aim to detect only the users' faces, usually in front of a static 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 that are not always perceptible.

Other approaches, e.g. TOURBOT [2] or GRACE [3], which try to perceive the whole person rather than only the face, use laser range-finders to detect people as moving objects. Drawbacks of these approaches occur, for instance, in situations where a person stands near a wall and cannot be distinguished from the background, in scenarios with objects yielding leg-like scans (such as table or chair legs), or if the laser range-finder does not cover the whole 360°.

In [4], 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 [5]. Another skin-color-based approach was presented in [6], where a multi-target tracker was realized by using multiple instances of a simple condensation tracker [7]. The major problem of skin-color-based approaches is that, in a natural environment, typically many skin-color-like objects exist that are not humans.

For real-world scenarios, more promising approaches combine more than one sensory channel, such as visual

[☆] This work is partially supported by TMWFK Grant #B509-03007 to H.-M. Gross and HWP Grant to A. Scheidig.

* Corresponding author.
E-mail address: christian.martin@tu-ilmenau.de (C. Martin).
URL: http://www.tu-ilmenau.de/neurob (C. Martin).

0921-8890/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2006.04.012

cues and the laser range-finder scan. An example for these approaches is the SIG robot [8], 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 realizing face-to-face interaction. Further examples are the EXPO-ROBOTS [9], where people are first detected as moving objects by a laser range-finder (resulting from differences from a given static environment map). After that, these hypotheses are verified by visual cues. Other projects, such as BIRON [10], detect people by using the laser range-finder for detecting leg profiles and combine this information with visual and auditory cues. The essential drawback of most of these approaches is the sequential integration of the sensory cues. People are detected by laser information only and are subsequently verified by visual or auditory cues. These approaches typically fail, if the laser range-finder yields no information, for instance, in situations when only the face of a person is perceptible because of leg occlusion.

Therefore, we propose a multi-modal approach, which can be characterized by the fact that all used sensory cues are concurrently processed and integrated into a robot-centered map using a probabilistic aggregation scheme. The overall computational complexity of our approach scales very well with the number of sensors and modalities. This allows a simple extension by integrating further sensory channels, such as sound sources.

As sensory channels, we use the different sensory modalities of our experimental platform 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). Experimental results will be shown in Section 4, followed by a short summary and outlook in Section 5.

2. The interaction-oriented robot system HOROS

To investigate the respective detection and aggregation methods, we use the mobile interaction robot HOROS (HOme RObot System) 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. Therefore, HOROS has to realize a user-based interaction, where the robot has to analyze its user, the gender, the age, the facial expression, the pose and the distance of the person to himself, and subsequently adjust its dialog strategies and presentation modes to adapt to the current user. To realize such an interaction-oriented personal robot, stable methods for people detection and tracking are basic prerequisites.

HOROS' hardware platform is an extended Pioneer-based robot from ActiveMedia. It integrates an on-board PC (Pentium M, 1.6 GHz, 512 MB) 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 Fig. 1). This includes a tablet PC for touch-based interaction, speech recognition and speech generation. The robot was further extended by a robot face that 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 robust multi-modal people detection and tracking.

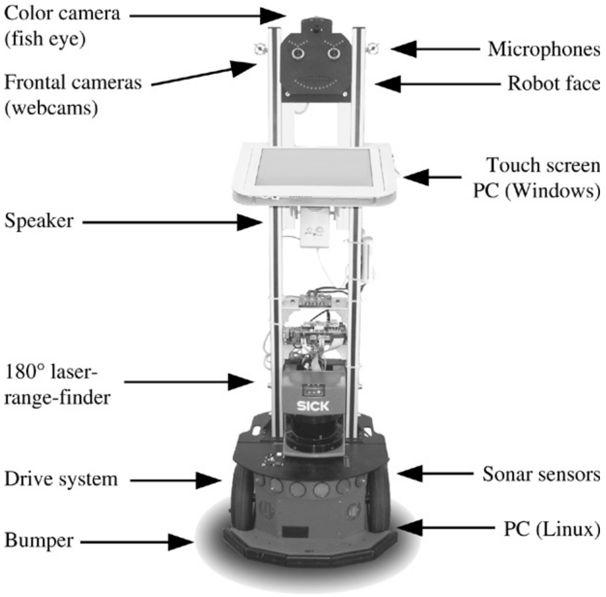

Fig. 1. Sensory and motor modalities of the mobile interaction-oriented robot HOROS (HOme RObot System). The laser range-finder, the sonar sensors, and the omnidirectional camera are used here for people detection and tracking.

2.1. Laser-based information

The laser range-finder is a very precise sensor with a resolution of 1°, perceiving the frontal 180° field of HOROS (see Fig. 2 left). It is fixed on the robot approximately 30 cm above the ground. Therefore it can only perceive the legs of people.

Based on the approach presented in [11], 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 such as width, deviation and other conditions that are characteristic 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 to the robot is extracted. This approach yields very good results for the distances of people standing less than 3 m away. For a greater distance, legs are relatively often 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 detection of table legs or

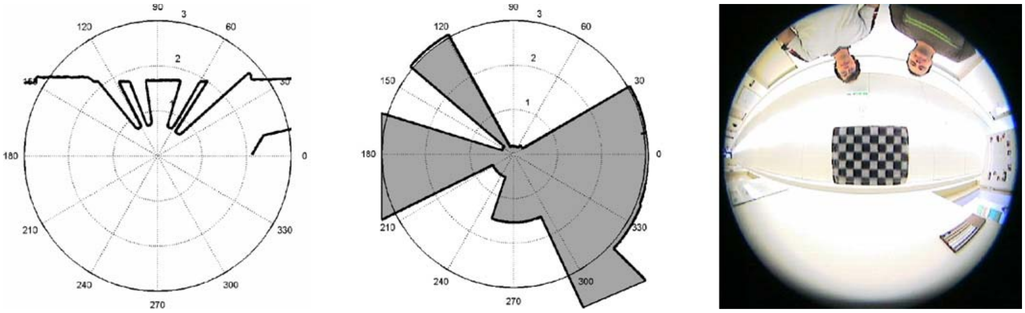

Fig. 2. Exemplary sensory inputs from the laser range-finder (left), the sonar (middle), and the fisheye camera (right) in a typically situation, where two people are standing in front of the robot.

chair legs and also other narrow objects similar to legs. People standing sideways on to the robot or wearing long skirts do not yield appropriate values of the laser range sensor to detect their legs, resulting in a relatively high false-negative rate.

2.2. *Sonar-based information*

Furthermore, HOROS has 16 sonar sensors arranged on the Pioneer platform approximately 20 cm above the ground. Because of this, people detection using the sonar sensors only works by analyzing the sonar scan for leg profiles (see Fig. 2). 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 object's material, the direction of the reflecting surface, crosstalk effects when using several sonar sensors, and the absorption of the broadcast sound. Because of these disadvantages, only distances of less than 2 m can be considered for people detection using these sonar sensors. This means that the sonar sensors yield pretty unreliable and inaccurate values, a fact that has to be considered in the generation of hypotheses for people detection. For the purpose of very simple people detection using this sensory modality, we assume that all measurements less than 2 m could be hypotheses for a person. These hypotheses could be further refined by comparing the position of each hypothesis with a given local map of the environment. If the hypothesis corresponds to an obstacle in the map, it could be neglected. The disadvantage of this refinement strategy is that people standing near to an obstacle often are not considered as valid person hypotheses.

2.3. *Fisheye camera*

As a third sensory cue, we use an omnidirectional camera with a fisheye lens yielding a 360° view around the robot. An example of an image resulting from this camera is depicted in Fig. 2 (right).

To detect people in the omnidirectional image, our skin-color-based multi-target tracker [6] is used. This tracking system is based on the condensation algorithm [7]. It has been extended to allow the visual tracking of multiple objects at the same time. This way, particle clouds used to estimate the probability of people in the omnidirectional image can concentrate on several skin-colored objects. A typical problem of this simple feature extraction for observation is the possible tracking of a large spectrum of skin-color-based but non-human objects, such as wooden shelves, etc. We used this straightforward approach for visual people detection because it is much faster than subsampling the whole image, trying to find regions of interest, and is resistant to minor interferences due to the observation skin-color-model that is used [6].

People detection using omnidirectional camera images yields only hypotheses about the direction of a person but not about his/her distance. Assuming a mean size of a human face and a mean body height, it would be possible to give a rough estimate of the distance of a person to the robot based on the size of the skin colored region and the distortion parameters of the omnidirectional camera. Due to sensor fusion with the other modalities, such a rough distance estimation is expendable. Therefore, it was expected that the fusion of the hypotheses from the camera with the hypotheses of the laser range-finder and the sonar sensors would result in a more powerful people detection system. Subsequently, the method developed for the aggregation of several sensory systems will be discussed in detail.

3. Generation and tracking of object hypotheses

3.1. *Generation of sensor-specific position hypotheses*

For the purpose of tracking, the sensor-specific information about detected humans is converted into Gaussian distributions $\phi(\mu, C)$. The mean μ equals the position of the detection in robot-centric 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. Furthermore, the sensors have different error rates of misdetections that have to be taken into account. All computation is performed in the robot-centric r, φ space. Examples for the resulting distributions are shown in Fig. 3.

3.1.1. *Laser-based information*

The laser range-finder yields very precise data, hence the corresponding covariances are small and the distribution is narrow (see Fig. 3, bottom 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 by more than the

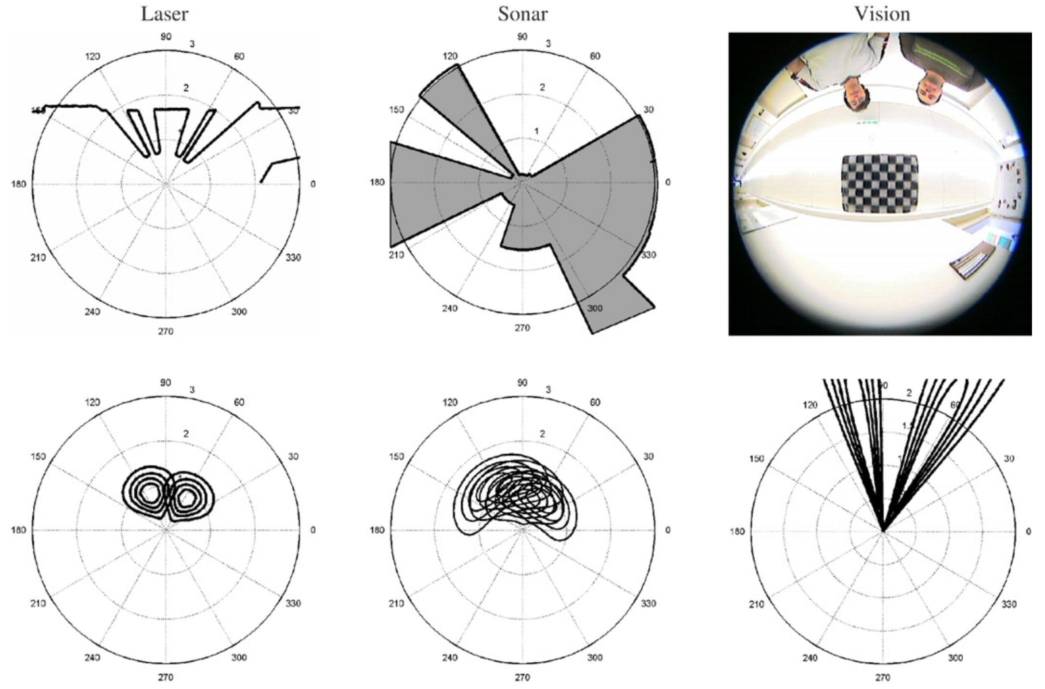

Fig. 3. Examples for generated hypotheses in a situation, where two people were standing in front of the robot. The top row shows the input of the different sensory systems and the bottom row shows the generated hypothesis.

same movement at a distance of 2 m. The smaller the distance of the hypotheses, the larger its variance has to be. Despite the insufficencies (see Section 2), 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 its arrangement, so people behind the robot are ignored when processing laser-based information.

3.1.2. 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 lower (see Fig. 3, bottom middle). Nevertheless, the sonar is included to support people-tracking behind the robot. With that, we are at least able to form an estimate of the distance for a vision-based hypothesis.

3.1.3. Fisheye camera

In contrast to the other detectors, the camera can only provide information about the angle of a detection, but not about the distance of a person (see Section 2). Therefore, for the radial variance of the distance coordinate, a very large value was selected, with a fixed mean value (see Fig. 3, bottom right). The angular variance is determined directly from the angular variance of the particle distribution generated by the skin-color based multi-person tracker, yielding the visual detection hypotheses (see Section 2.3).

The importance of the detection hypotheses is determined by the position of the hypotheses. In front of the robot, the influence is lower, because the laser sensor is considered to be a more reliable sensor. Behind the robot, the image is the only source to obtain information about the presence of a person; the sonar has only a supporting character. Thus, the relative weight of a visual hypothesis should be higher behind the robot.

The modeling and integration of additional sensory cues, such as human voice localization or other features from the camera image (like face structure or movement), can be performed in a similar way to that described here.

3.2. Multi-hypotheses aggregation and tracking

Tracking based on probabilistic methods attempts to improve the estimate x_t of the position of people at time t. These estimates x_t are part of a local map or model M that contains all hypotheses around the robot. Fig. 4 shows the principle architecture of the tracking system, which will be explained in detail in the following section. The local map M is used to aggregate the sensor hypotheses. Therefore, the movements of the robot $\{u_1, \dots, u_t\}$ and the observations about humans $\{z_1, \dots, z_t\}$ have to be taken into account. In other words, the posterior $p(x_t|u_1, z_1, \dots, u_t, z_t)$ 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.

In the map or model, 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 i is the estimate for one person.

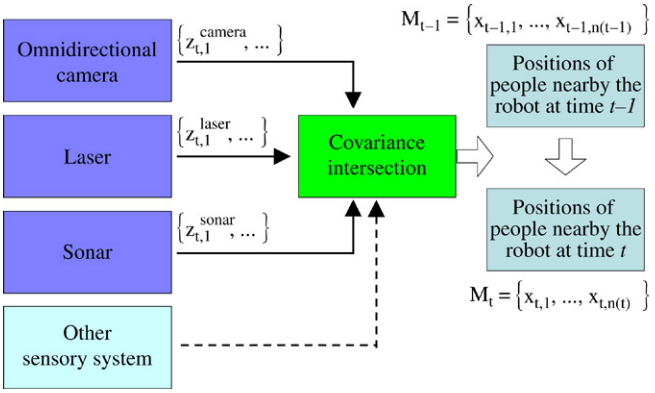
Omnidirectional camera
Laser
Sonar
Other sensory system
$\{z_{t,1}^{camera}, \ldots\}$
$\{z_{t,1}^{laser}, \ldots\}$
$\{z_{t,1}^{sonar}, \ldots\}$
Covariance intersection
$M_{t-1} = \{x_{t-1,1}, \ldots, x_{t-1,n(t-1)}\}$
Positions of people nearby the robot at time t–1
Positions of people nearby the robot at time t
$M_t = \{x_{t,1}, \ldots, x_{t,n(t)}\}$

Fig. 4. The architecture of the tracking system: The observations $z_{t,i}^{\text{camera}}$, $z_{t,i}^{\text{laser}}$ and $z_{t,i}^{\text{sonar}}$ (see Section 3.1) of the different sensory cues are combined in a local map M_t that contains a time-varying number $n(t)$ of estimates $x_{t,j}$ around the robot using the *Covariance Intersection* rule [12].

$\phi_i(\mu_i, C_i)$ is a Gaussian centered at μ_i with the covariance matrix C_i. The weight w_i ($0 < w_i \leq 1$) contains information about the contribution of the corresponding Gaussian to the current model.

Next, the current sensor-specific hypotheses z_t have to be integrated, after they have been preprocessed as described above. If M does not contain any element at time t, all generated hypotheses from z_t are copied to M. Otherwise data association has to be performed to determine which elements from z_t and M refer to the same hypothesis. For that purpose, the Mahalanobis distance d_m and the Euclidian distance d_e between the respective Gaussians $\phi_i \in z_t$ and $\phi_j \in M$ are used as association criteria. In a series of experimental investigations, it turned out that the Euclidian distance leads to better tracking results:

$$\begin{array}{rclcrcl} d_m & = & \mu C^{-1} \mu^T & \text{with} & \mu & = & \mu_i - \mu_j \\ d_e & = & |\mu| & & C & = & C_i + C_j. \end{array} \tag{2}$$

This determined distance is compared to a threshold. As long as there are distances lower than the threshold, the sensor hypothesis i and the map hypothesis j are merged. This is done by means of the *Covariance Intersection* rule [12]. This technique does not need any information about the correlation between the hypotheses, unlike a *Kalman filter*. The covariances C_i (sensor hypotheses) and C_j (map hypotheses) 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 back to the original space. The new mean is computed with respect to the *Information Space*:

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

The purpose of the weight ω is to minimize the resulting determinant by preferring the sharper distribution in the intersection process. With that, a very unreliable sensor input will have only a minimal influence on the resulting hypothesis.

Sensor readings that do not match any hypothesis of M are introduced as new hypothesis in M. The weight w_i represents the certainty of the corresponding map hypothesis, coded as a 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. To get a temporal smoothing, the weight is increased recursively as follows:

$$w_i(t+1) = w_i(t) + \alpha(1 - w_i(t)), \tag{4}$$

if that map hypothesis could be matched to a sensor hypothesis. The constant $\alpha \in [0, 1]$ is chosen with respect to the current sensor in order to integrate the availability and reliability of the sensor system into the aggregation framework (see Section 3.1). The more reliable the sensor, the higher that the α-weight is. These values were determined before by means of experiments. In the case of an unmatching hypothesis, the weight is decreased:

$$w_i(t+1) = w_i(t) - (1-\theta)\frac{t_{\text{new}} - t_{\text{old}}}{t_v}. \tag{5}$$

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

4. Experiment, investigation and discussion

The system that is presented is in practical use on our robot HOROS in a real-world environment (our institute building). The fact of a change in illumination in different rooms and hallways and numerous distractions in the form of chairs and tables is quite challenging.

Fig. 5 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, such as table legs and skin-colored objects. No sensor modality alone was able to detect people correctly. Only the aggregation over several sensor modalities and temporal integration led to the proper result.

To measure the detection rate of our system, an image sequence of the scene was recorded by a camera on the ceiling. The test sequence consists of 300 images over 30 s. These *ground-truth* data were labeled by hand to determine the position of the people in the scence. The proposed tracking system was able to track multiple people correctly with a detection rate of 93% in the experiment. This means that 93% of the labeled *ground-truth* data were detected by the tracker correctly.

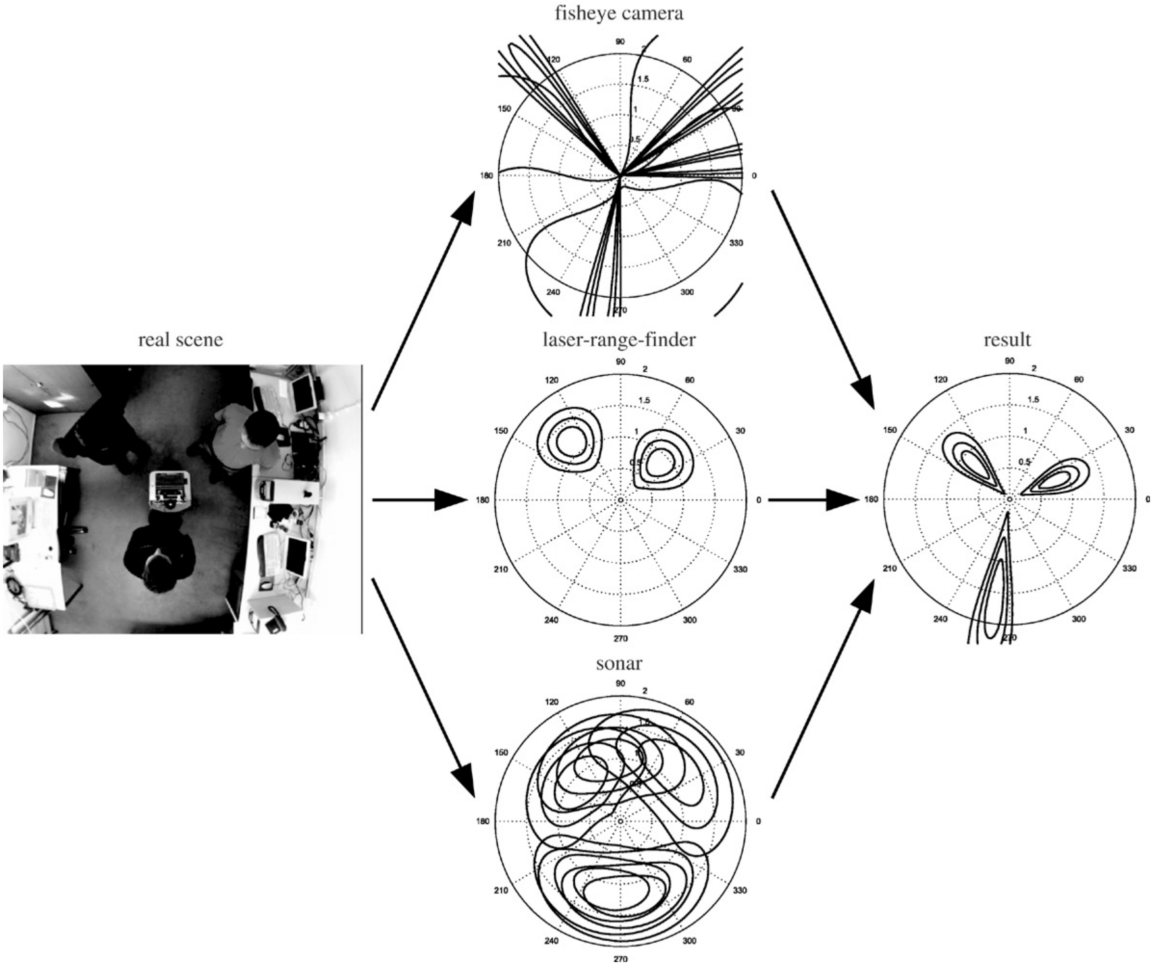

Fig. 5. Aggregation example. The left picture shows the real office scene from a bird's eye view. Three people are surrounding the robot, who stands in the middle. The three figures in the middle row show the current hypotheses generated by fisheye camera, laser-range-finder, and sonar from top to bottom. No sensor on its own can represent the situation correctly. The final picture on the right displays the aggregated result from the sensors and the previous timestep. This is a correct and sharpened representation of the current situation.

In most cases, false-negative detections occured behind the robot. The rate of false-positive detections is higher—about every fourth 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 too many people than finding too many. But there are ways to reduce the amount of false-positive detections. Most misdetections occur from static objects in the environment, so, based on the movement trajectories created by the tracker, they can be separated from correct hypotheses.

In our experiments, the sonar and the laser sensor worked at 10 Hz. The omnidirectional camera produced hypotheses with an update rate of about 7 Hz. As described in Section 3.2, all different sensors are handled independently of each other. Therefore, the resulting map was updated about 27 times per seconds. Overall, the whole tracking system (including the recording of the sensor information and the preprocessing processes) uses about 40–50% CPU load on a 1.6 GHz Pentium M.

The system that is presented improved the detection performance for the area behind the robot only slightly compared to a simple skin-color tracker. This is because the sonar-based sensors do not provide much 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 the prevention of precipitate extinction of hypotheses in case of sudden changes in illumination. In this case, the skin-color tracker will presumedly 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 front of the robot, the multi-modal 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 much more precise hypotheses. 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. 5. This leads to the assumption that the inclusion of additional sensory systems generating hypotheses about interacting people (e.g. sound-source hypotheses of speaking

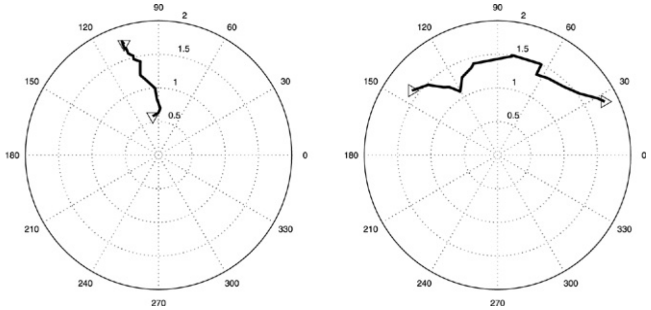

Fig. 6. Left: A trajectory showing a person coming straight towards the robot. Right: The person is crossing from left to the right. In doing so, the robot is avoided. The varying time intervals between the movements and the associated weights are not visible in the figure.

people [13]) will further improve the performance of this multi-modal tracking system.

Our system was tested practically in the context of a survey task. Horos stood in a hallway in our institute building. Its task was to attract the attention of people walking past it. As soon as the system recognized a person in the defined interaction area within a radius of 3 m, the robot addressed the visitor to come nearer. It then offered to participate in a survey about the desired future functionality of Horos. The people-tracking module was used to detect break offs. Thus, if the user was leaving the robot before finishing the survey, the robot tried to fetch them back and finalize the survey. After successful 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 performed in the absence of any visible staff members, so that people could interact in a more unbiased manner.

These efforts are repeated from time to time to gather more information, and there is a second intention that is not obvious. The tracking module was used to observe 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 person as a potential interaction partner. 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 typical different movement trajectories are shown in Fig. 6. The most challenging aspects for a classification of walking trajectories are, in our opinion, the varying speed of the people and the search for typical movement schemes describing the current interest of the potential users in interacting with a robot. This is a typical recognition problem in human–human interaction too.

5. Summary and outlook

We presented a flexible multi-modal probability-based approach for detecting and tracking people. It is implemented on our mobile interaction-oriented robot Horos and works in real-time. Because of the sensor fusion and the probabilistic aggregation, its results are significantly improved compared to a single sensor tracking system. In our future work, we will extend the system with additional cues to further increase the robustness and reliability for real-world environments (especially in areas where the current sensory cues are insufficient). Currently, we are working on the integration of voice-based speaker localization [13]. 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.

References

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

[2] D. Schulz, W. Burgard, D. Fox, A. Cremers, Tracking multiple moving objects with a mobile robot, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR, 2001, pp. 371–377.

[3] 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, B. Maxwell, Grace: An autonomous robot for AAAI robot challenge, AAAI Magazine 24 (2) (2003) 51–72.

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

[5] H. Tao, H.S. Sawhney, R. Kumar, A sampling algorithm for tracking multiple objects, in: Workshop on Vision Algorithms, 1999, pp. 53–68.

[6] T. Wilhelm, H.-J. Boehme, H.-M. Gross, A multi-modal system for tracking and analyzing faces on a mobile robot, Robotics and Autonomous Systems 48 (2004) 31–40.

[7] M. Isard, A. Blake, Condensation — conditional density propagation for visual tracking, International Journal on Computer Vision 29 (1998) 5–28.

[8] K. Nakadai, H. Okuno, H. Kitano, Auditory fovea based speech separation and its application to dialog system, in: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, vol. 2, 2002, pp. 1320–1325.

[9] 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, N. Tomatis, Robox at expo.02: A large scale installation of personal robots, Robotics and Autonomous Systems 42 (3–4) (2003) 203–222.

[10] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. Fink, G. Sagerer, Audiovisual person tracking with a mobile robot, in: Proc. Int. Conf. on Intelligent Autonomous Systems, IAS Press, 2004, pp. 898–906.

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

[12] S. Julier, J. Uhlmann, A nondivergent estimation algorithm in the presence of unknown correlations, in: Proc. American Control Conference, vol. 4, IEEE, 1997, pp. 2369–2373.

[13] R. Brueckmann, A. Scheidig, C. Martin, H.-M. Gross, Integration of a sound source detection into a probabilistic-based multimodal approach for person detection and tracking, in: Proc. Autonome Mobile Systeme, AMS 2005, Springer, 2005, pp. 131–137.

Christian Martin has been a Ph.D. student at the Department of Neuroinformatics and Cognitive Robotics at Ilmenau Technical University since 2004. He received his Diploma degree in Computer Science from the Ilmenau Technical University in 2003. His Ph.D. research is concerned with multi-modal human–robot interaction, especially the development of modeling concepts for human–robot interaction and autonomous mobile robot control architectures.

Erik Schaffernicht has worked at the company MetraLabs GmbH Germany since 2006. He received his Diploma degree in Computer Science from Ilmenau Technical University in 2006. He works in the field of mobile robotics. He is especially interrested in human–robot interaction and methods for autonomous outdoor navigation.

Dr. Andrea Scheidig is a Postdoc at Ilmenau Technical University, Faculty of Computer Science and Automation, Department of Neuroinformatics. She received her Diploma degree in Computer Science in 1996 and her Doctorate degree in Neuroinformatics in 2003. Her main research interests are user-adaptive human–robot interaction, behavior-based systems and reinforcement learning.

Dr. Horst-Michael Gross has been a full professor of Neuroinformatics at the Ilmenau Technical University, Faculty of Computer Science and Automation, and has headed the Department of Neuroinformatics since 1993. He received his Diploma degree in Technical and Biomedical Cybernetics in 1985 and his Doctorate degree in Neuroinformatics in 1989. Among his main research interests are neural computing, autonomous robots, reinforcement learning, and vision-based human–robot interaction. He is a member of INNS and ENNS.