Generating Persons Movement Trajectories on a Mobile Robot

Andrea Scheidig, Steffen Mueller, Christian Martin, and Horst-Michael Gross

Abstract— **For socially interactive robots it is essential to be able to estimate the interest of people to interact with them. Based on this estimation the robot can adapt its dialog strategy to the different people's behaviors. Consequently, 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 an imposed approach for integration of several sensor modalities and present a multimodal, probability-based people detection and tracking system and its application using the different sensory systems of our mobile interaction robot HOROS. For each of these sensory cues, separate and specific Gaussian distributed hypotheses are generated and 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 integration of further sensory channels, even with different update frequencies, and the usability in real-world HRI tasks. Finally, promising experimental results achieved for people tracking in a real-world environment, an university building, are presented.**

I. INTRODUCTION

Human robot interaction (HRI) in real-world environments typically requires that a robot has to interact with people moving around. Prerequisite is a stable people detection and tracking. Depending on the specific robot application that integrates a people 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 voice 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.

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 next to wall and cannot be distinguished from the background, in scenarios with objects yielding leg-like scans, like table- or chair-legs, or if the laser-range-finder does not cover the whole 360 degrees.

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]. An other 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 which are not humans.

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 [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 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 [10] detect people by using the laser-range-finder for detecting leg-profiles and combine these 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 perceivable because of leg occlusion.

Therefore, we propose here a multimodal 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 makes it easy to extent the tracker by further sensory channels, like sound sources.

As sensory channels we use the different sensory modalities of our experimental platform HOROS: the omnidirectional and the frontal camera, the sonar sensors, and the laser-range-finder (see section II). 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 III). Experimental results will be shown in section IV.

II. THE INTERACTION-ORIENTED ROBOT SYSTEM HOROS

For our experiments we use the mobile interaction robot HOROS (HOme RObot System). HOROS' hardware platform

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

A.Scheidig, S.Mueller, Ch.Martin, and H.-M.Gross are with Institute of Neuroinformatics and Cognitive Robotics, Technical University of Ilmenau, Germany andrea.scheidig@tu-ilmenau.de

is an extended Pioneer 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). These include a tablet PC for touch-based interaction, speech recognition and speech generation. The robot was further extended by a robot face which integrates an omnidirectional fisheye camera, two microphones, and two frontal cameras for person detection and analysis of the user's features. Subsequently, the laser-range-finder, the sonar sensors, the omnidirectional and the frontal camera are discussed in the context of a robust multimodal people detection and tracking.

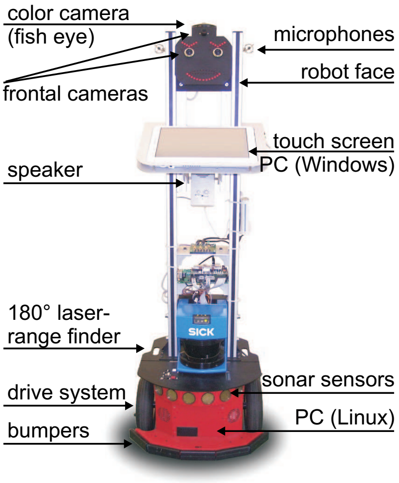

Fig. 1. Sensory and motor modalities of the mobile interaction-oriented robot HOROS (HOme RObot System). The laser-range-finder, the sonar sensors, and all three cameras are used here for people detection and tracking.

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 left, top row). 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 like width, deviation and others 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 distances of people which stand less than 3 meters away. In a greater distance legs are relatively often missed due to the limited resolution of the laser-range-finder (The gaps between single rays become to large.). The strongest disadvantage of this approach is its false-positive classification detecting table- or chair-legs and also other narrow objects as legs. People standing sideways 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.

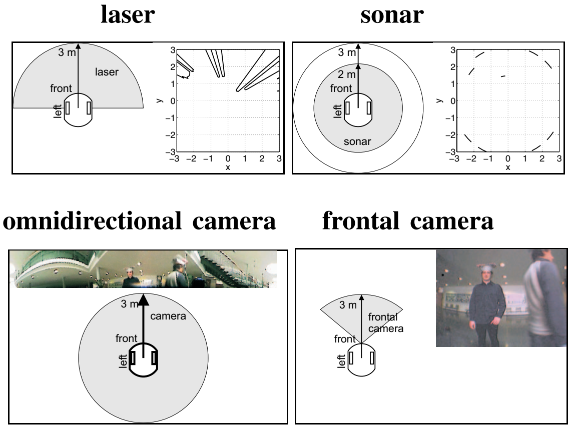

Fig. 2. Exemplary sensory inputs from the laser-range-finder and the sonar depicted in the top row and from the fisheye camera and the frontal camera depicted in the lower row for a typical situation, where two people standing in front of the robot. In the pictures of the laser and the sonar scan, the robot is located at the 0,0 coordinate straightened ahead. Together with the exemplary inputs the principle range of each sensory cue to detect people is depicted. As can be seen, each sensor covers a specific area around the robot. Consequently best tracking results are to be expected if all sensory cues are used concurrently.

Sonar-based information: Furthermore, HOROS has 16 sonar sensors arranged at the Pioneer platform approximately 20 cm above the ground. 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 broadcasted sound. Because of these drawbacks, only distances of less than 2 meters can be considered for people detection. Different to the laser scan it is not possible to detect leg-pairs using sonar sensors. Therefore, each measurement of less than 2 meters will be used as a hypothesis for a person (see Figure 2 right, top row). If the robot is localized in its operational area, these hypotheses could be further refined by comparing the position of each hypothesis with a given local map of the environment. If the hypothesis would correspond to a static obstacle in the map, it could be neglected. The disadvantage of this refinement strategy is, that people standing near by an obstacle often are not considered as valid person hypotheses. Nevertheless, sonar sensors are indispensable for the observation area behind the robot, because no other distance measurement sensors are available there.

Fisheye camera: As third sensory cue we use an omnidirectional camera with a fisheye lens yielding a 360 degree view around the robot. An example of an image resulting from this camera is depicted in Figure 2, left bottom.

To detect people in the omnidirectional image, a 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 to possibly track a large spectrum of skin-colored but non-human objects, like wooden shelves. 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 because it is resistant to minor interferences due to the used skin-color-model [6].

A people detection using omnidirectional camera images only yields hypotheses about the direction of a person but not about his/her distance. Therefore by the usage of hypotheses from the camera together with the hypotheses from the laser-range-finder and the sonar sensors the missing distance information from the camera can be compensated. Nevertheless, in Section III-C we discuss alternative approaches for the combination of visual and distance information.

Frontal camera: As fourth sensory cue, a frontal camera yielding an approximately 90 degree frontal view is utilized (see Figure 2, bottom right). To detect people in the image of the frontal camera, we use the face detection algorithm from VIOLA and JONES [12] to find frontal view faces. Using only the image from the frontal camera would result in the detection of only few people in front of the robot. Similar to the omnidirectional camera, information about the distance of people is not available directly. In Section III-C, we also discuss methods to integrate the information from the frontal camera with distance information from laser and sonars.

Subsequently, the general idea of the developed architecture for the aggregation of the several sensory systems is presented.

III. GENERATION AND TRACKING OF OBJECT HYPOTHESES

A. *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 Cartesian 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 II. Furthermore, the sensors have different error rates of misdetections that have to be taken into account. This is realized by means of different sensor specific weights during aggregation. All computation is done in the robot-centric x, y space, which makes the motion update easier. Examples for the resulting distributions are shown in Fig. 3.

Laser-based information: The laser-range-finder yields very precise data, hence the corresponding variances are small and the distribution is narrow (see Figure 3, top). The mean value of the Gaussian depends on the distance of the detected leg-pair and both variances are fixed with approximately 0.4 meters. Despite the insufficiencies of this sensor like the limited perception space (see section II), the probability of a misdetection is the lowest of the used sensors, and so the used weight to represent the certainty of a sensor hypotheses is the highest.

Sonar information: Information from the sonar tends to be very noisy, imprecise and unreliable. Therefore, the variances are large and the impact on the certainty of a hypothesis is lower. Nevertheless, the sonar is indispensable 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.

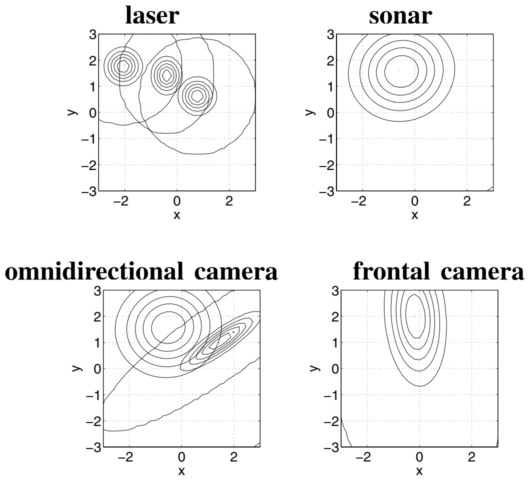

Fig. 3. Exemplary hypotheses, representing the situation depicted in the images of Figure 2, where two people are standing in front of the robot. In each picture the robot is located at the 0,0 coordinate straightened ahead. The top row shows the hypotheses generated from the laser and the sonars. There can be seen one hypothesis for each person and additionally a hypothesis generated by the pillar. The bottom row shows the generated hypotheses using the camera systems. Using the omnidirectional camera (left) two hypotheses can be obtained, which contain no direct distance information. Because of the limited perception range of the frontal camera (right), only one hypothesis can ge generated.

Fisheye and frontal camera: In contrast to distance measuring sensors, the cameras can only provide exact information about the angle of a detection, but not about the distance of a person (see section II).

For both cameras, the means of the Gaussians in angular direction are determined directly by the sensor detections. Furthermore, the Gaussians have fixed values in the distance direction, with 1.5 meters for the fisheye camera and 2 meters for the frontal camera (the frontal camera can perceive people in greater distances).

For the variance of the distance, a large value of 1 meter was selected for the Gaussians of both cameras. The variance in angular direction was also chosen as fixed for the frontal camera with a value of 0.6 meters. For the fisheye camera, this variance is directly determined by the angular variance of the particle distribution generated by the skin-color based multi-person-tracker yielding the visual detection hypotheses (see Section II).

The modelling and integration of additional sensory cues, like human voice localization or other features from the camera image (like movement), can be done in a similar way as described here.

B. Multi-hypotheses aggregation and tracking

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 or model M that contains all hypotheses around the robot. This map is used to aggregate the several sensor hypotheses as described in section III-A. 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 movement 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 (see Figure 4).

In the map or model, a set of weighted Gaussians $M = \{\mu_i, C_i, w_i | i \in [1, n]\}$ is used to represent the positions of people, where each Gaussian i is the estimation for one person. $\phi_i(\mu_i, C_i)$ 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 probability to represent a person by the corresponding Gaussian.

Next, the current sensor specific hypotheses z_t have to be integrated, after they have been preprocessed as described in section III-A. 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 done to determine which elements from z_t and M refer to the same hypothesis. For that purpose, different distance measures between the respective Gaussians $\phi_i \in z_t$ and $\phi_j \in M$ were investigated as association criterion. In a series of experimental investigations it turned out, that the simple Euclidian distance d_e leads to the best tracking results.

The 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 [13].

By applying this rule, the resulting determinant is minimized by preferring the sharper distribution in the intersection process. With that, a very unreliable sensor input will have only minimal influence on the resulting hypothesis.

Sensor readings that do not match with any hypothesis of M are introduced as new hypothesis into M. The weight w_i is representing the certainty of the corresponding map hypothesis. The more sensors support this hypothesis, the higher this weight should be. If the weight passes a threshold ρ, the corresponding hypothesis is considered to be a person.

In the case of no corresponding sensor hypothesis, the weight of the map hypothesis is decreased (see Equation 2). There Δt is the duration since the last sensor update took place. d defines the time that has to pass to consider a person as lost in the map while no sensor has made a new detection that can be associated with this hypothesis.

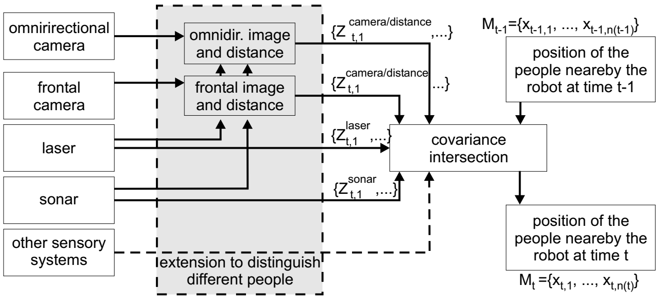
omnirirectional camera
frontal camera
laser
sonar
other sensory systems
omnidir. image and distance
frontal image and distance
extension to distinguish different people
{Z t,1 camera/distance ,...}
{Z t,1 camera/distance ...}
{Z t,1 laser ,...}
{Z t,1 sonar ,...}
covariance intersection
M t-1={x t-1,1, ..., x t-1,n(t-1)}
position of the people neareby the robot at time t-1
position of the people neareby the robot at time t
M t ={x t,1, ..., x t,n(t)}

Fig. 4. The architecture of the tracking system: The observations $z_{t,i}^{omni\ camera}$, $z_{t,i}^{frontal\ camera}$, $z_{t,i}^{laser}$ and $z_{t,i}^{sonar}$ (see Section III-A) 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 [13]. The grey depicted elements are the extensions to the original tracking system. By using these extensions, information from both camera systems is combined with distance information from the laser and the sonar system.

$$w_i(t+1) = w_i(t) - \Delta t \cdot d \tag{2}$$

The weight is increased with each observation of any sensor using a specific value c_s while f_s is the sensors observation frequency.

$$w_i(t+1) = w_i(t) + \frac{c_s}{f_s} \tag{3}$$

By means of choosing each c_s less than d we can reach that a hypothesis must be supported by at least two sensors so that false positives will be reduced.

C. Extensions to distinguish different people

Using the discussed sensory cues purely results in a problem when merging map hypotheses with small variances with a hypothesis from the vision system, which has a large variance in distance direction (see Figure 5 left). In the consequence, the mean values of the Gaussian with the small variances are shifted to the fixed mean value of the Gaussian of the visual cue. Second, by merging the covariances of the map hypothesis with the hypothesis from the vision system, specific distance information represented in the small variances of the map hypothesis get also lost (see Figure 5 right). In the result, after a few observations people in the same direction at different distances can not be distinguished any longer.

There are different ways to handle these problems. First, if a sensor hypothesis will matches more than one map hypothesis, it is merged to the closest one only. Another way is the combination of the distance independent visual hypotheses with distance dependent information from the laser and sonar system before the observations are merged in the map model (see the grey depicted extensions in the middle of Figure 4). This integration of distance information into visual data can be achieved by implicit assumptions about the distance of a person using the camera images. Assuming a mean body height, it would be possible to give a rough estimate about the distance of a person to the robot. Due to the known position of the camera, it is possible to distinguish at least a few rough distance classes based on the vertical position of the skin-color region in the omnidirectional camera image. Using the frontal camera and face detection, the size of the face in the image is reciprocally proportional to the distance. In the consequence, the means and the variances for the Gaussian distributions could be approximated either by explicit distance information or by implicit assumptions about mean people's height and mean face sizes.

In our current work, we have investigated all discussed extensions. The obtained results showed, that the accuracy of the tracker could be improved using each discussed approach. However, to demonstrate the pure functionality of our tracker, in Section IV we present the results obtained from the version without using any extension.

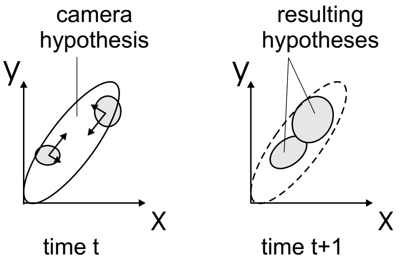

Fig. 5. Merging a hypothesis with small variances with a hypothesis obtained from the vision system (left) results in a hypothesis with wrong distance information (right).

IV. Experiment, Investigation and Discussion

The presented system is in practical use on our robot Horos in a real-world environment, a university building. The fact of a changing illumination in different rooms and hallways and numerous distractions in form of chairs and tables is quite challenging.

To evaluate our multimodal multi-person tracker we obtained data from an experimental setup, where the robot was standing in a foyer and people moved around it. The environment contained several distracting objects, like a pillar and some skin-colored objects. As depicted in the aggregation example in Figure 6, no sensor modality alone was able to detect all the people and their position correctly. Only aggregation over several sensor modalities and temporal integration led to the proper result.

The whole experimental setup was monitored by a further top-down camera mounted above the robot. Because the robot did not move in this experimental setup, we were able to get a reference of the positions of the robot and the people moving around it for about fifteen minutes (see Figure 6 top). To get our ground truth, each person recorded had to wear a red colored hat, which could be tracked easily in the top down images. Because of the distortion in the top-down camera and the perspective projection, varying heights of the persons and other sources of noise the accuracy of our baseline was limited to about 20 cm.

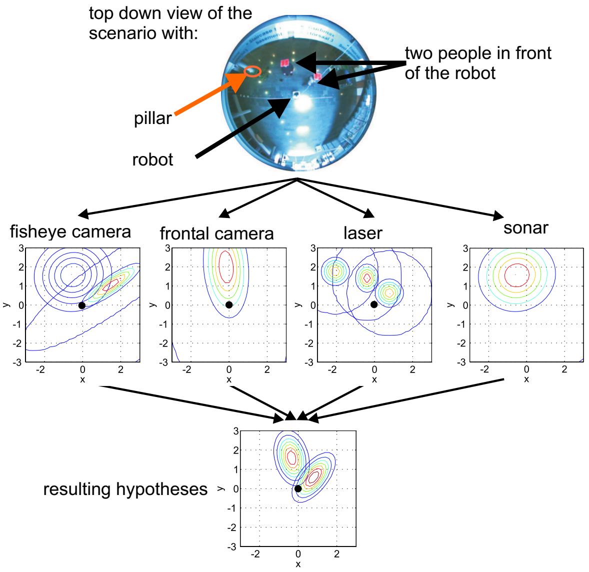

Fig. 6. **Aggregation example.** The upper picture shows the real scene from a bird's eye view. Two people are standing in front of the robot. The four figures in the middle row show the current hypotheses generated by fisheye camera, frontal camera, laser-range-finder, and sonars. No sensor on its own can represent the situation correctly. The final picture on the bottom displays the aggregated result from the sensors and the previous timestep. This is a correct and sharpened representation of the current situation.

To compare the quality of the tracker, first the detection rate has been evaluated by searching for a tracker hypothesis for each known person position in a top-down image. Taking into account the noise in the baseline a person is counted as a detection if the distance between tracker hypothesis and top-down position is below 50 cm. To get an impression up to what range the tracker is able to find people, the detection rate has been evaluated for different distances of people to the robot.

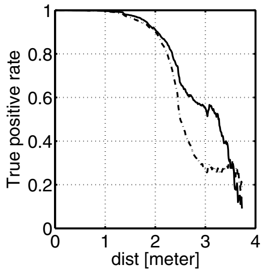

Fig. 7. This picture shows the detection rates obtained from the multimodal multi-person tracker for different distances to the robot. Above 2 m the benefit of using multiple sensors (solid line) over a laser only tracker (dotted line) becomes clearly visible.

Fig. 7 shows this result: Up to a distance of 2.5 m nearly 80% of the persons in the top-down image have been detected, taking into account that the maximum range of the used sensors to detect people is 3 m.

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 tracking results only have any relevance if they are stable for several seconds, what false positives can not achieve.

In a further experiment, the average position error of the trajectories was evaluated. Some exemplary plots of estimated trajectories with a length above a minimum threshold and the respective top-down trajectories are shown in Fig. 8. The error of position for these examples is below 0.5 m, which is sufficient for the intended purpose of classification the movement trajectories with respect to the person's interest to interact with the robot.

In all our experiments, the sonar and the laser sensor worked at 10 Hz. The omnidirectional and the frontal camera produced hypotheses with an update rate of about 7 Hz.

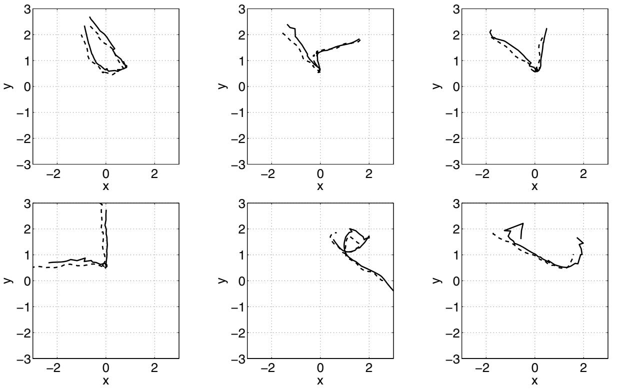

Fig. 8. These pictures show an exemplary comparison of the tracker results (solid line) to the baseline from top-down view (dashed line). The robot was standing in the middle, facing upwards.

V. SUMMARY AND OUTLOOK

We presented a flexible, multimodal, probability-based approach for detecting and tracking people. It is implemented on our mobile interaction-oriented robot HOROS and is working in real-time. Because of the sensor fusion and the probabilistic aggregation, its results are significantly improved compared to a single sensor tracking system. Further, we presented different possible extensions to combine visual and distance information, where each approach will further improve the tracker results.

In our future work, we will extend the system with additional cues to further increase robustness and reliability for real-world environments (especially in areas, where the current sensory cues are insufficient). Next, we are working on the integration of an voice-based speaker localization [14], a head-shoulder tracker and also on the integration of a detector for side view faces using the face detection algorithm from VIOLA and JONES [12]. A further aspect addresses the evaluation of tracker results using a moving robot, where we bared on preliminary investigations expect similar results as we obtained already.

Further, based on the generated movement trajectories, we are estimating the interest of people to interact with the robot. Thereby, different movement trajectories typically represent different kinds of interest to interact with the robot, and thus require specific robot behaviors. For this classification, we utilize different methods to represent and classify the obtained movement trajectories, especially Time Delay Neural Networks and Hidden-Markov Models.

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] Tao, H., Sawhney, H.S., Kumar, R., 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, in: Robotics and Autonomous Systems, Vol. 48, 2004, pp. 31–40.

[7] M. Isard, A. Blake, Condensation - conditional density propagation for visual tracking, Int. 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 42 (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, Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems 43 (2-3) (2003) 133–147.

[12] P. Viola, M. Jones, Robust real-time object detection, in: Proc. of IEEE Workshop on Statistical and Computational Theories of Vision, 2001.

[13] 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.

[14] 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.