Proc. of the 3rd European Conference on Mobile Robots (ECMR) 2007, Freiburg, pp. 211-216, Albert-Ludwigs-Universität Freiburg - Universitätsverlag 2007

Are you still following me?

Steffen Müller, Schaffernicht Erik, Andrea Scheidig, Hans-Joachim Böhme, Horst-Michael Gross*
Neuroinformatics and Cognitive Robotics Lab, Technical University Ilmenau, Germany

***Abstract*— In this paper, the people tracking system of a mobile shopping assistant based on a SCITOS-G5 platform is explained in detail. The robot has two tasks, to find people requiring assistance and to guide a user to a target without losing contact. For that purpose, a probabilistic model and a Bayesian update scheme have been developed, where data of various sensory systems is merged asynchronously. Experimental results of the tracking behavior during several guided tours in a home store demonstrate the reliability of our approach.**

***Index Terms*— Person Tracking, Human Robot Interaction, Robot Assistant**

I. Introduction

Robust detection and tracking of people in the surroundings of mobile service robots plays a central role in many applications. Especially for our long-term research project PERSES (PERsonal SErvice System) [3], this knowledge is essential to allow an intuitional interaction. The aim of our project was to develop an interactive mobile shopping assistant that can autonomously guide its user, a customer, to desired articles within a home store. For this purpose, the robot patrols in the home store and offers service to customers looking around. Once a person has started to interact with the robot, this user has to be observed during a guided tour, to recognize when the user stops or wants to continue the tour.

Depending on the specific application scenario that integrates a people detection and tracking task, different approaches are prevalent. Typically the utilization of visual cues for face detection and tracking is a preferred way, but for our application tracking objects in image space is not sufficient. The goal is to track people in the two dimensional world space, allowing to estimate their distance to the robot and to infer about their behavior and movement trajectories.

Therefore, in [4] we introduced a probabilistic approach for tracking people's positions in a robot centered (r, φ)-coordinate system, which realizes an equitable fusion of different sensory systems. The main improvement there was to overcome the disadvantages of single sensor or hierarchical tracking systems, often used on mobile robots. TOURBOT [7] or GRACE [10], for example, use a laser-based system for detecting people in the robot's surroundings. In spite of recent improvements on the classification of laser range scans as proposed by [1], in an extremely dynamic and cluttered environment like the home store, the reliability of exclusively laser-based systems is not sufficient. To increase the reliability, in hierarchical approaches visual information is often used for verification of hypotheses generated by the laser [9]. The combination of laser with visual and additionally auditory cues is presented e.g. in [2]. The essential drawback of most of these hierarchical approaches is the sequential integration of the sensory cues. These systems typically fail if the laser range finder yields no information. Besides, for a mobile robot which has to deal with moving people, faces will not always be perceivable, hence verification could fail too.

*This work is partially supported by TAB Grant #2006FE0154 to H.-M. Gross

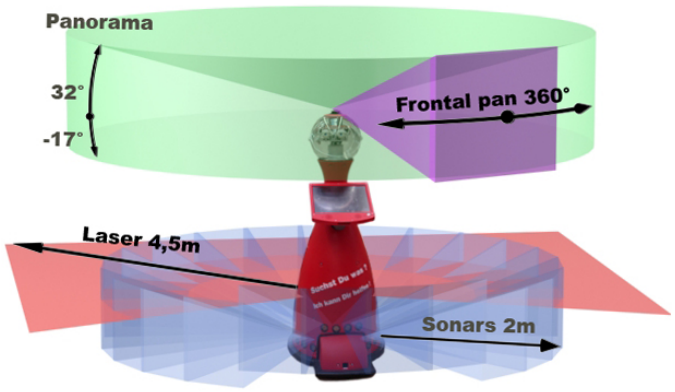

Fig. 1. The interactive robot platform SCITOS A5, produced by MetraLabs GmbH. Shown are the ranges of the different sensor systems: a SonyRPU camera at the top yielding a pano ramic and a frontal image, a laser range scanner, and 24 sonar sensors at the bottom.

Due to these findings, for the application of the shopping assistant, we concentrated on the improvement of the already existing sensor fusion method [4], which additionally has the advantage that the perceivable area is not limited by the range of one sensor. In [5] the algorithm of [4] had been already improved by means of representing the position of persons in an Euclidean world space, allowing to generate movement trajectories and to estimate their velocity. For the application presented here, the model had to be extended by components needed for making decisions on the robot's behavior. Furthermore, the use of a Kalman Filter-like update mechanism instead of the former covariance intersection and the introduction of a motion model for interaction partners did improve the robustness in difficult environments, like the home store.

The rest of the paper is structured as follows: First, our Platform SCITOS A5 is introduced and the mission in the home store is described. Afterwards, the utilized sensory cues will be shown in detail, before the probabilistic model and the Bayesian aggregation scheme is explained. Finally, experimental results demonstrate the effectiveness of our system.

II. Robot system and its application

The robot, which has been developed for the application as a shopping assistant, is a SCITOS A5 (see Fig. 1). For navigation and interaction purposes the system is equipped with different sensors. First, and most important is an omnidirectional camera mounted on the top of the head. Due to the

integrated hardware transformation, we are able to get both a panoramic image (720x150 pixels) and high resolution frontal image (720x362 pixels), which can be panned around 360^o. Besides this main sensor, the robot is equipped with a set of 24 sonar sensors at the bottom, which is used for obstacle detection and localization. Because of their diffuse characteristics, these sensors do not allow to distinguish objects from people, but they cover the whole 360^o around the robot. The last sensor available for person detection is the laser range finder HOKUYO URG-04-LX mounted in front direction at a height of 35cm. Additionally, the robot has a touch display, a sound system, and a 6 DOF head for interaction.

The aim of the robot in the home store is to assist customers during their purchase. Therefore, at first, people, who seem to need help have to be found, while the system is patrolling. Indications for the interest of customers to interact with the robot are given, when a person is standing still or facing the robot for a while. Once a dialog started, the user has the option to be guided to a specific article. During this guided tour, the robot has to observe the user to detect if the person stops or keeps on following the robot.

III. SYSTEM ARCHITECTURE

As mentioned above, different sensory cues are used for estimating the positions of people nearby the robot. The main source of information are the images and the occupancy map of the local environment, integrating all the range information from sonar and laser (see Fig.2). Due to the noisy characteristics of the sonars, the direct utilization of their measurements is not useful.

The centered column of Fig.2 shows the preprocessing modules, which will be explained in more detail in the next sections. Similar to our former approach [4], each of the cues provides Gaussian distributed hypotheses $H_s = (\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{C}_s), w_s)$ of persons' positions each with an individual reliability weight w_s. Here $\mathcal{N}$ is the Gaussian with mean $\boldsymbol{\mu}_s$ and covariance matrix $\boldsymbol{C}_s$. The improved probabilistic user model at the right side of Fig.2 is using these observations to estimate the users' positions and further properties as described below. In extension to the algorithm of [4, 5], a user motion model, further sources of information, and a Kalman Filter-based update rule have been introduced. In our fusion approach, the different sensory systems are absolutely equitable. So, their advantages and shortcomings can complement each other.

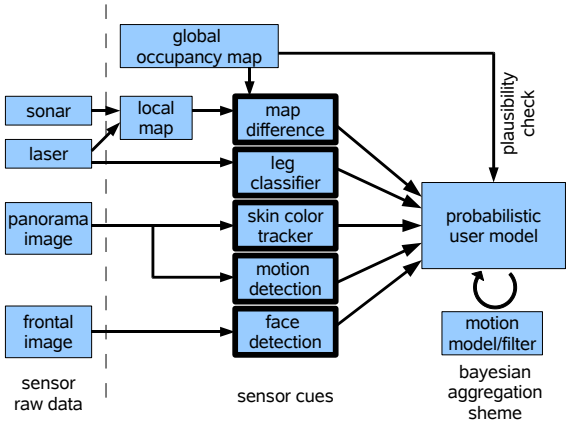

Fig. 2. Architecture of the multimodal tracker, left: the sources of information, middle: preprocessing systems, delivering Gaussian hypotheses, right: probabilistic model and update mechanism.

A. Getting hypotheses from occupancy maps

One disadvantage of visual cues is their weak accuracy in distance information. Therefore, range measuring sensors are essential for tracking in world coordinates. Because of the limited scan sector of the laser range finder and the noisy sonars, we decided to extend our former system by integrating the raw range measurements in an occupancy gridmap, representing the current local surroundings of the robot. Once an occupancy map is estimated, we are able to calculate a virtual local 360^o range scan $R_l(n)$ at the robot's position, which is representing the current situation with an acceptable certainty. This scan is acquired using 72 virtual sensors each scanning a 5^o sector. Objects in that virtual scan then are classified as person or background. For that purpose, a second scan $R_g(n)$ is extracted from a global occupancy map, which has been learned before and is also used for self-localization. By comparing these two scans, hypotheses of non-static moveable objects can be generated. In a dynamic environment, like a home store, the probability for these detections to be a person is relatively weak, such that the hypotheses resulting from this cue have a low weight w_s.

As a result of the mean global localization error of about $25cm$ [6], it is necessary to evaluate the differences in the scans $R_l(n)$ and $R_g(n)$ as follows: If the range in the local scan is shorter than the range in the global scan, it is a hint for a dynamic object. For each direction n with $R_l(n) < R_g(n)$, the minimum distance $d(n)$ of the endpoint of the scanray in the local map to all the endpoints in the global scan is calculated. For that, the ranges $R_l(n)$ and $R_g(n)$ are transformed back into a world position, where Euclidean distances can be derived. The probability $P(n = obj)$ for a direction n to show an object results from a simple ramp function with maximum value P_{max} above a distance threshold d_{th}.

$$P(n=obj) = \begin{cases} 0 & ,if R_l(n) > R_g(n) \vee \\ & \quad d(n) < d_{min} \\ \frac{P_{max}(d(n)-d_{min})}{d_{th}-d_{min}} & ,if R_l(n) < R_g(n) \wedge \\ & \quad d(n) \in [d_{min}, d_{th}] \\ P_{max} & ,if R_l(n) < R_g(n) \wedge \\ & \quad d(n) > d_{th} \end{cases} \tag{1}$$

Each connected sequence of a $P(n = obj) > 0$ afterwards is transformed into a Gaussian hypothesis for the probabilistic tracker by estimating the center and variances of the points defined by $R_l(n \in sector)$. The weight w_s of this hypothesis is taken from the average $P(n = obj)$ in the sector. Here, the question arises, why not to compare the maps directly? Due to the fact that the robot can not sense the area behind an object, the local map is only valid inside the range of a scan taken from robot's center. Using a virtual scan helps suppressing false positive hypotheses resulting from occluded objects.

B. Leg hypotheses based on laser scans

In contrast to the sonar sensors, the laser scan has a high resolution and a better reliability. As mentioned above, different projects, which use leg detection in laser scans, make an effort to get a high precision for the classification as, e.g. proposed in [1]. Due to the integration of further sensors, we do not depend on very robust leg detection. Therefore, a simple heuristic is used for generating hypotheses. First, steps in the range scan $R_{laser}(\varphi)$ of more than $7cm$ are detected to define segments between them. Then each segment has to fulfill different criteria to be accepted as a possible leg.

1) segments must have a length of at least $8cm$ and at most $20cm$,
2) the standard deviation of R_{laser} of all values inside a sector must be limited by $4cm$,

Given the set of possible legs in 2D world coordinates, pairs of legs are found, if the Euclidean distance between any two legs is less than $0.5m$. For each of these pairs a Gaussian observation hypothesis is generated at the average position of the legs. Due to the cluttered environment, the fix reliability w_s of observing a person this way is chosen quite low.

C. Skin color based hypotheses

A popular feature extracted from images is skin color. Many models for skin color are known [14] but the most serious problem of using color in a dynamic environment is the changing illumination. For the purpose of white balancing the used camera (see Fig. 1) is equipped with a white reference, which is the best known method to cope with changing light.

For classification of color, an empirical tabular model in normalized rg-color space has been built up by labeling faces in several hundred images captured in the home store. By binning the rg-plane into 64x64 cells the ratio of face pixels to nonface pixels, falling into a cell, gives a probability $P(rg = skin)$.

To find and track clusters of skin color in the panoramic image $I(\varphi, z)$ with a vertical image size Z, a set of particle filters is used. The concept introduced in [13] allows to track multi-modal distributions of skin color in the image resulting in a set of disjoint hypotheses for people's skin regions.

The color tracker works as follows: Each of the individual particle filters is representing the distribution of the position of a skin color blob by a set of 100 particles $p_n = (\varphi_n, z_n, \pi_n)$. Then, on each image the following cycle is iterated ten times.

1) Weight sampling: Using the tabular skin color model, new weights π_n result from rg color at the particle position in the image.

$$\pi_n = P(I_{rg}(\varphi_n, z_n) = skin) \tag{2}$$

2) Resampling: A set of next generation particles is chosen from the old one according to the distribution of π_n.
3) Motion model: A random walk model is used. For each particle a postion offset $\Delta\varphi_n \propto \mathcal{N}\left(0, (18^o)^2\right)$ and $\Delta z_n \propto \mathcal{N}\left(0, (0.04\,Z)^2\right)$ is drawn.

After that, each of the particle filters should have converged to one cluster of skin color. If two particle sets concentrate on the same cluster, one set is initialized equally distributed over the whole image in order to find other possible spots of skin color. To generate observation hypotheses in 2D world coordinates, the average direction $\bar{\varphi}$ and a typical distance to the robot $r_{fix} = 1.5m$ are used to specify the mean of the Gaussian hypotheses. Furthermore, the variance in radial direction is chosen fix to $(1.5m)^2$ and the variance in tangential direction to the robot is estimated from the particle's variance in φ. The average weight $\bar{\pi}$ of the particles of the respective filter defines the weight w_s for the new hypothesis.

D. Motion based hypotheses

A reliable feature of living objects is motion, which therefore is chosen as a further cue. Because proper motion on a mobile robot is making motion classification difficult, this is only useful, while the robot is not moving. To overcome the time consuming computation of optical flow, which is often used for motion detection, a difference image based approach is realized. Further speedup is reached by summing up the pixels in an image column to get a reduced one dimensional signal $\bar{I_{rgb}}(\varphi) = \sum_{z\in[0.3\,Z, 0.7\,Z]} I_{rgb}(\varphi, z)$. The difference of $\bar{I_{rgb}}(\varphi)$ on two succeeding images gives a reliable hint for motion. Due to the averaging process, pixel noise is of low influence, a threshold of 8 gray values yields to stable classification of moving sections over φ. The set of difference sections is postprocessed by a morphological closing operation to get connected intervals of moving people, which afterwards can be transformed into Gaussian hypotheses, if a minimum size is exceeded.

In contrast to the color hypotheses, the estimation of distance of moving objects is difficult, due to the unknown size and speed of objects. For instance, near a window, a moving car outside may lead to a similar motion detection as a human near to the robot. For handling that problem, the distance needed for the Gaussian hypothesis is taken from the range scan in the local occupancy map. If this distance is above a threshold of $2.5m$ the hypothesis is discarded.

E. Face detection based hypotheses

A promising hint for persons in the surrounding of the robot as well as their direction of sight, is a frontal face in the image. For finding faces, the well known Viola and Jones detector [12] is applied. Unfortunately, faces in the omnidirectional image are pretty small, which restricts detection to the frontal image only. To cover the whole 360^o area, this magnified image is panned around periodically, until a face is found and tracking commences. For getting more information about the user, the face image is analyzed by means of ICA feature projection [13]. Besides the features for identifying people, a likelihood for an image to be a face is estimated, which is useful for validating the face detections and yielding a reliability w_s. For generation of Gaussian observation hypotheses H_s, besides the given direction, a distance is needed again. The fixed size of faces ($15cm$, empirical determined) is utilized to triangulate the distance with a deviation of about $40cm$.

Concluding, we point to the fact, that different sensory systems are chosen to complement each other. We have range

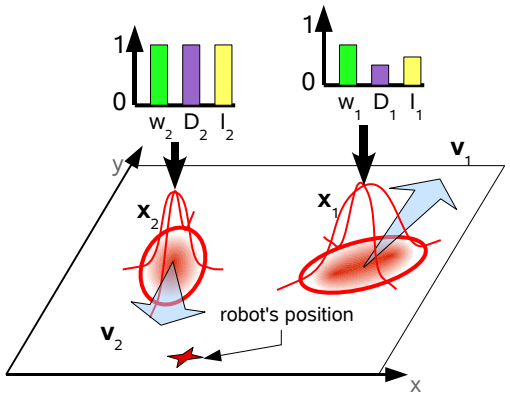

Fig. 3. Two components of the probabilistic user model are shown. Each one has a position $\boldsymbol{x}$ and a velocity $\boldsymbol{v}$ which are modelled by Gaussians in 2D world coordinates. Further, for each hypothesis the probability of being a human w, the need for interaction I, and the probability of being the current dialog partner D are modeled by discrete distributions over $\{0, 1\}$

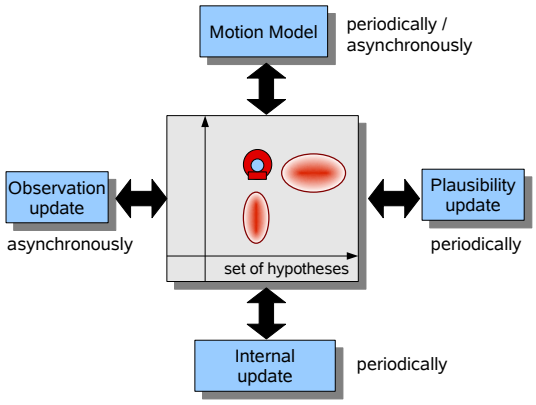

Fig. 4. The probabilistic model consisting of a set of hypotheses, is independently modified by the different, processes (blue boxes)

sensors useful for observing the position of people with a high accuracy, but with a high false positive rate. Therefore these inputs will have a high influence on the position of a hypothesis, but a weak one on the belief of representing a person at all. On the other side, there are the vision-based observations, each with a bad spatial accuracy, but with a better selectivity. Thus, the main task of skin color is to generate hypotheses at a distinct direction to the robot. Later, they can be refined in their distance by the range sensors. Finally, motion and face detections are useful to prove the hypothesis to be a person with a high reliability. Face detections give a further hint, that the person did notice the robot, because only frontal faces are found.

All these observations H_s from the various subsystems are used to keep a probabilistic model up to date, which is explained in the following section.

IV. Probabilistic Tracking

A. Modelling the persons

A probabilistic model for people in the surroundings of the robot is used, as illustrated in Fig. 3. There is a varying number of hypothesis $H_k = (\boldsymbol{x}_k, \boldsymbol{v}_k, I_k, D_k, w_k)$, where the position $\boldsymbol{x}_k \propto \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{C}_k)$ and the velocity $\boldsymbol{v}_k \propto \mathcal{N}(\boldsymbol{\nu}_k, \boldsymbol{V}_k)$ are modeled by Gaussians in 2D world coordinates with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{C}_k$ respectively $\boldsymbol{\nu}_k$ and $\boldsymbol{V}_k$. The probability for the object in the model to be a human at all is described by w_k. Additionally, for each hypothesis k the need for interaction I and the probability D of being the user who is in dialog with the robot are estimated internally. I is used to decide whether to offer service to a detected person while searching for new users. The position of the hypothesis with maximum D and w_k above a threshold, determines whether to wait for the user, or to continue a guiding tour to a given target point in the store.

Given the model, in the following its update and estimation process is shown. As illustrated in Fig.4, there are four processes modifying the model independently.

B. Data association and observation update

The sensory systems and so the preprocessing cues are providing new observations $H_s = (\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{C}_s), w_s)$ asynchronously with a sensor specific update rate t_{sensor}. Thus, on each new observation the estimation of the model can be improved, after predicting the current state with the user motion model which is described in section IV-E.

Before updating the parameters of a single hypothesis the data association problem has to be solved. That means, by which one of the K hypotheses, a new observation is probably generated from? In literature for that difficulty, Joint Probability Data Association Filters [8] have been applied with a high computational requirement. Following our approach from [4], we concentrated on a simple distance based heuristics, which leads to a modified Expectation-Maximization algorithm.

Based on the Mahalanobis-distance $d(H_s, H_k)$ between observation distribution H_s and the model hypothesis H_k, the affiliation $P(k)$ of the observation to a certain component k is defined using a threshold function.

$$P(k) \propto \begin{cases} 0 & , if\ d(H_s, H_k) > d_{max} \\ \frac{1}{d(H_s, H_k)} & , else \end{cases} \tag{3}$$

If no component has a probability larger than zero, a new hypothesis is created at the observed position. In the other case $P(k)$ is used for updating the position $\boldsymbol{x}_k$, after a normalization step. Further on, the person probability w_k (see Fig. 3) is adjusted proportionately.

The update of the position $\boldsymbol{x}_k \propto \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{C}_k)$ for all components with $P(k) > 0$ is done using a Kalman Filter-like update rule (see eqs. 4-6), resulting in the pointwise product of the old position distribution and the observation distribution following the typical scheme of Bayes Filters [11]. Controlling the influence of the new observation, K is called the Kalman gain, while the affiliation $P(k)$ regards the data association problem.

$$\boldsymbol{K} = P(k)\ \boldsymbol{C}_k\ (\boldsymbol{C}_k + \boldsymbol{C}_s)^{-1} \tag{4}$$

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k + \boldsymbol{K}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_k) \tag{5}$$

$$\boldsymbol{C}_k^{new} = (\boldsymbol{I} - \boldsymbol{K})\,\boldsymbol{C}_k \tag{6}$$

The probability w_k of being a human, can hardly be handled in a Bayesian manner, because there are only observations

proofing the presence of a person. Due to the deficit of non-person observations, a simpler method is realized for estimating w_k. During the internal update (see Fig.4), the probability is decreased over time with a decay rate of w_{decay} to realize a disappearance of person hypotheses not observed any longer. To compensate this decay, the sensor specific probability w_s is normalized by the update rate t_s of the respective sensor, and then added to the weight considering the affiliation to component k. If w_k^{new} exceeds 1, it is clipped.

$$w_k^{new} = w_k - w_{decay}\,\Delta t + w_s\,w_{decay}\,t_s\,P(k) \qquad (7)$$

Here Δt is the time elapsed since the last update.

By means of sensor reliabilities $w_s < 1$ it is guaranteed, that only observations supported by two or more sensors can increase their person probability w_k and, therefore, the deficits of single sensors can be compensated.

C. Internal update

As already mentioned, besides the asynchronous observation update, an internal update is periodically applied in order to update the $\boldsymbol{v}_k$, D_k and I_k components of the model. Furthermore, to limit the number of hypotheses contained by the model, hypotheses with a probability w_k reaching zero are deleted.

Given the current position for a person, the velocity can be updated. Because the position may jump within a short time span, which is caused by the different sensors, a longer interval for speed estimation is used. For that, the history of each position, i.e. the former estimations $\boldsymbol{\mu}_k^{t-l}, \ldots, \boldsymbol{\mu}_k^{t}$, are stored in a queue. Using this information, the velocity $\boldsymbol{v}_k \propto \mathcal{N}(\boldsymbol{\nu}_k, \boldsymbol{V}_k)$ can be estimated recursively as follows.

$$\boldsymbol{\nu}_k^t = \alpha\,\boldsymbol{\nu}_k^{t-1} + (1-\alpha)\tfrac{1}{\tau}\textstyle\sum_{\delta t}\left(\boldsymbol{\mu}_k^t - \boldsymbol{\mu}_k^{t-\delta t}\right) \qquad (8)$$

$$\boldsymbol{V}_k^t = \alpha\,\boldsymbol{V}_k^{t-1} + \tfrac{1}{\tau}\textstyle\sum_{\delta t}\left(\boldsymbol{\mu}_k^t - \boldsymbol{\mu}_k^{t-\delta t}\right)\left(\boldsymbol{\mu}_k^t - \boldsymbol{\mu}_k^{t-\delta t}\right)^T \qquad (9)$$

Here δt runs over the history queue from one to two seconds in the past, and τ is the number of elements considered.

Afterwards, the probability for the need for interaction I_k is estimated. A person, standing still for a while and facing the robot, is supposed to require assistance. Another indication for interest in interaction is given, if a person is actively approaching to the robot and its touch display on the back (see Fig. 1). Following these rules, the interaction probability I_k will be increased i) if a face detection has been observed or ii) if a person is approaching the area in the back of the robot. In the same manner as updating the person probability w_k, a constant decay I_{decay} is subtracted from I_k periodically. The value is increased with each face detection, indicating that the person faces the robot. Because face detection is not as reliable as expected, the main increment for I_k must be generated by the second variant, the approaching behavior. Therefore, people standing nearby the robot, will get an increment depending on their distance d to the interaction side, which is a point in front of the touch display.

As a last component of the model, the probability D_k of a person to be the current dialog partner is estimated. It would be promising to identify the person, may be using the face analysis results, but experiments showed that this feature is not yet as robust as needed. The user has to face the robot and a proper model for his appearance has to exist. Both are points of failure, thus the decision on D_k up to now is made on position only. Because the user's presence is of chief interest during a guiding tour, a region dependent on the robot's movements is defined. A movable angular sector, limited by the absolute angles α_l and α_r in world coordinates, is defined, where the borders will follow the robot's rotation delayed, which is realizing that if the robot is turning, the possible user's position is more flexible. Inside the cone, between α_l and α_r, the probability of being the current dialog user is increased and outside it is decreased using an identical mechanism as already described for the w_k values. Based on the new I_k and D_k values, the robot now can easily select its behavior appropriately.

D. Plausibility update

Due to the structured environment, sometimes hypotheses will emerge inside the goods shelves. For example, due to skin-colored and leg-like structures in the goods shelves, their probability w_k to be a person, can grow wrongly. Knowing about the structure of the environment, false-positive hypotheses laying inside the obstacles in the global occupancy map can be supposed to be not a person, which yields an enormous improvement compared to the original tracking system [5]. Therefore, periodically the average occupancy value o_k at the position $\boldsymbol{x}_k$ is estimated by sampling the Gaussian distribution of the hypothesis and looking up the occupancy values in the global map. Following the principles described above, the person probability w_k then is updated according to:

$$w_k^{new} = w_k - w_{map}\,o_k\,t_{map} \qquad (10)$$

Whereas w_{map} is a decay factor and t_{map} is the time since the last map check.

E. Motion update

In order to apply the Bayes filter approach completely, before each observation update, the state of each hypothesis (its position $\boldsymbol{x}_k$) has to be predicted using a process model. And even if no new observation updates occur, the person's position is changing and, therefore, a state prediction using a motion model needs to be done periodically.

For a person primarily a random walk model is applied in order to increase the variances of position and speed. Therefore, a time dependent constant is added to covariance matrices $\boldsymbol{C}_k$ and $\boldsymbol{V}_k$. If a person is observed over a longer timespan, additionally their velocity is used to predict the position by simply adding $\boldsymbol{\nu}_k\,\Delta t$ to the mean position $\boldsymbol{\mu}_k$. Here Δt is the timespan since the last motion update took place.

During the guiding tour of the robot, a modified motion model can help to find people following the robot. To prevent the hypotheses in the model to be missed by new observations, what will lead to the introduction of a new hypothesis without the already estimated parameters, the robot's speed is used to predict an expected position for obedient users. That way, the tracker filters those hypotheses, which follow a similar

trajectory to the one of the robot. This new feature helps to decrease the misassociations drastically when the robot is moving fast.

$$\boldsymbol{C}_k^{new} = \boldsymbol{C}_k + s\,\Delta t\,\boldsymbol{I} + f_v\,\boldsymbol{V}_k \quad (11)$$

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k + f_v\,\Delta t\,\boldsymbol{\nu}_k + f_r \boldsymbol{v}_{robot}\,\Delta t \quad (12)$$

$$f_v = 1 - \frac{|\boldsymbol{v}_r|}{v_r^{max}} \quad (13)$$

In the equations, $\boldsymbol{v}_r$ is the velocity vector of the robot and f_r and f_v denote the rate of influence for the robot's and the person's estimated speed, while s denotes the speed of the random walk component.

V. Experimental results

The current form of the model evolved over years. It has been applied on different platforms each with an individual set of sensors. Also a variety of other applications for a tracking system similar to the presented one exist.

Keeping in mind the desired application as a shopping assistant, differing from previous work, where the tracking of the exact positions of people was in the focus of attention, here the ability of the model to provide the robot with crucial information for decision making is essential. Therefore, instead of ground truth tracking results, an application oriented test has been chosen. In order to evaluate the ability to track a customer during a guided tour, the robot had to give a tour, while different people followed it in an individual distance. In Fig. 5 three of these runs are shown. The plots visualize the continuity of the tracking result for different average distances. During the tour with a length of about $50m$, the robot lost contact only two times in five trials, where the people where asked to keep a distance of about $1m$ to $1.5m$, which is a normal displacement for people really doing a shopping tour.

In usability tests, where 155 people have been interviewed, most of the people were pleased with the behavior of the robot. 59 of the 155 users did a guided tour, and only 3 of them criticized, that the robot did stop and lost them too often.

In a second experiment, the number of false positive hypotheses was evaluated. Here, the robot should interrupt its tour if a customer stops following. Therefore, at the marked point S in Fig. 5, the experimenter stopped while the robot continued its guiding tour. In all of these trials, the robot detected the disappearance of the user correctly after at least $5m$.

VI. Conclusions and outlook

Summing up, in this work the current user tracking system of SCITOS as a mobile shopping assistant is described in detail.

The idea of fusion of different sensory cues for people tracking has been taken up and enhanced significantly. The probabilistic model has been extended by a motion model allowing a better prediction and filtering of persons' hypotheses according to their trajectories. Further, abstract components modelling the relevant properties of a user, namely the need for assistance by the robot and the fact of being the user at all, are modeled. Finally, the shopping assistant realizes a fairly

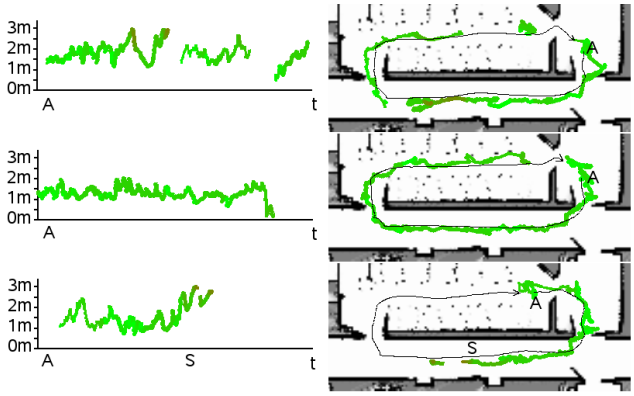

Fig. 5. Exemplary tracking results of a person instructed to follow the robot (black path) in different distances, top: $2m$, middle: $1.5m$, bottom: stopping at point S, color is showing the probability of being the dialog user (green), left: estimated distance of the robot to the user over time; right: map of our office building

natural interaction behavior, that seems very pleasant to its users.

In future work, we try to extend the probabilistic model by personal properties describing the current user like facial features and color histograms, allowing to identify the user if contact was lost to him.

References

[1] K. O. Arras, Ó. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *Proc. of 2007 IEEE International Conference on Robotics and Automation*, Italy, 2007.

[2] J. Fritsch, M. Kleinehagenbrock, S. Lang, G.A. Fink, and G. Sagerer. Audiovisual person tracking with a mobile robot. In *Proc. Int. Conf. on Intelligent Autonomous Systems*, pages 898–906, March 2004.

[3] H.-M. Gross and H.-J. Böhme. Perses - a vision-based interactive mobile shopping assistant. In *Proc. IEEE Int. Conference on Systems, Man and Cybernetics (IEEE-SMC 2000)*, pages 80–85, Nashville, 2000.

[4] C. Martin, E. Schaffernicht, A. Scheidig, and H.-M. Gross. Sensor fusion using a probabilistic aggregation scheme for people detection and tracking. In *Proceedings of 2nd European Conference on Mobile Robots (ECMR 2005)*, pages 176–181, stampalibri, 2005.

[5] A. Scheidig, S. Müller, C. Martin, and H.-M. Gross. Generating person's movement trajectories on a mobile robot. In *Proc. RO-MAN 2006 - 15th IEEE Int. Symposium on Robot and Human Interactive Communication*, pages 747–752, Hatfield (UK), 2006.

[6] Ch. Schröter, A. König, H.-J Böhme, and H.-M. Gross. Multi-sensor monte-carlo-localization combining omnivision and sonar range sensors. In *Proc. of the 2nd European Conference on Mobile Robots (ECMR 2005)*, pages 164–169, Ancona, Italy, 2005.

[7] D. Schulz, W. Burgard, D. Fox, and A. Cremers. Tracking multiple moving objects with a mobile robot. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 371–377, 2001.

[8] D. Schulz, W. Burgard, D. Fox, and A.B. Cremers. People Tracking with Mobile Robots Using Sample-Based Joint Probabilistic Data Association Filters. *The International Journal of Robotics Research*, 22(2):99, 2003.

[9] R. et al. Siegwart. Robox at expo.02: A large scale installation of personal robots. *Robotics and Autonomous Systems*, 42:203–222, 2003.

[10] R. et al. Simmons. Grace: An autonomous robot for AAAI robot challenge. *AAAI Magazine*, 24(2):51–72, Summer 2003.

[11] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.

[12] P. Viola and M. Jones. Fast and robust classication using asymmetric adaboost and a detector cascade. In *NIPS 2001*, pages 1311–1318, 2001.

[13] 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*, volume 48, pages 31–40, 2004.

[14] T. Wilhelm and C. Martin. Vergleich von hautfarbbasierten multi-target-trackern. In *Self-Organization of AdaptiVE Behavior (SOAVE)*, 2004.