in: Proc. 2008 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC 2008), Singapore, pp. 3471-78, IEEE 2008

ShopBot: Progress in Developing an Interactive Mobile Shopping Assistant for Everyday Use

H.-M. Gross, H.-J. Boehme, C. Schroeter, S. Mueller, A. Koenig
Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology, 98693 Ilmenau, Germany
Horst-Michael.Gross@tu-ilmenau.de

Ch. Martin, M. Merten, A. Bley
MetraLabs Robotics GmbH
98693 Ilmenau, Germany
http://www.metralabs.com

***Abstract*—The paper describes progress achieved in our long-term research project SHOPBOT, which aims at the development of an intelligent and interactive mobile shopping assistant for everyday use in shopping centers or home improvement stores. It is focusing on recent progress concerning two important methodological aspects: (i) the on-line building of maps of the operation area by means of advanced Rao-Blackwellized SLAM approaches using both sonar-based gridmaps as well as vision-based graph maps as representations, and (ii) a probabilistic approach to multi-modal user detection and tracking during the guidance tour. Experimental results of both the map building characteristics and the person tracking behavior achieved in an ordinary home improvement store demonstrate the reliability of both approaches. Moreover, we present first very encouraging results of long-term field trials which have been executed with three robotic shopping assistants in another home improvement store in Bavaria since March 2008. In this field test, the robots could demonstrate their suitability for this challenging real-world application, as well as the necessary user acceptance.** [1]

I. INTRODUCTION

In the year 2000, we got our long-term research project SHOPBOT (also known as PERSES) started which aims at the development of interactive mobile shopping assistant for everyday use in public environments, like shopping centers or home improvement stores [1]. Such shopping companions are to autonomously contact potential customers, intuitively interact with them, and adequately offer their services. Typical service tasks tackled in this project are to autonomously guide customers to the locations of desired goods (see Fig. 1), and to accompany them during the purchase as personalized mobile companion offering a set of functionalities, like video conferencing to a salesperson, price scanning, infotainment, etc. To accommodate the challenges that arise from the specifics of this interaction-oriented scenario and the characteristics of the operational area, a uniformly structured, maze-like and populated environment, we have placed special emphasis on vision-based and multi-modal methods for both human-robot interaction and robot navigation. Since one of the robot's tasks is to guide customers to the goods they are looking for, it needs to know its current position in the operation area as accurate as possible to give precise information how to find the desired articles. This requires advanced self-localization methods and

[1]The research leading to these results has received funding from the State Thuringia (TAB-Grant #2006-FE-0154) and the AiF/BMWI (ALRob-Project Grant #KF0555101DF).

Fig. 1. Interactive mobile shopping assistant based on a SCITOS A5 platform developed by MetraLabs GmbH Ilmenau, Germany, during a guided tour to a goods location in our test site, an ordinary home improvement store in Erfurt (Germany).

- as a prerequisite for this - robust map building techniques. Therefore, in section III, we first focus on recent progress concerning the on-line building of maps of the operation area using advanced Rao-Blackwellized SLAM (Simultaneous Localization And Mapping) approaches employing both sonar-based gridmaps as well as vision-based graph maps as representations. For an effective human-robot interaction and a user-specific, personalized dialog control, a robust people detection and tracking during the whole shopping process is another prerequisite. Therefore, the developed probabilistic, multi-modal people tracking system of our mobile shopping assistant will be presented in Section IV. Finally, first encouraging results of long-term field trials which have been executed with three shopping companions in a home improvement store in Bavaria since March 2008 will be presented.

II. THE MOBILE SHOPPING ROBOT SCITOS

The robot, which has been developed for the application as shopping assistant, is a SCITOS A5 shown in Fig. 2. For navigation and interaction purposes the system is equipped with different sensors. First, there 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

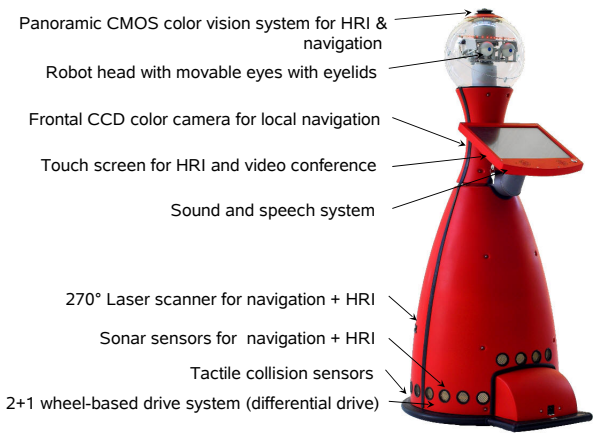

Fig. 2. Interactive mobile shopping assistant SCITOS A5

person detection is the laser range finder SICK S300 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.

III. SLAM in Large-scale, Public Environments

A basic requirement for an autonomous mobile robot is the ability to build a map of the environment and to use this map for self-localization and path planning. Because mapping depends on a good estimate of the robot's pose in the environment, while localization needs a consistent map, the localization and mapping problems are mutually coupled. The term Simultaneous Localization And Mapping (SLAM) has been coined for this problem [2]. The mutual dependency between pose and map estimates requires to model the state of the robot in a high-dimensional space, consisting of a combination of the pose and map state. An effective means of handling the high-dimensionality in the SLAM problem has been introduced in the form of the Rao-Blackwellized Particle Filter (RBPF) [3]. In this approach, the state space is partitioned into the pose and map state. A particle filter approximates the pose belief distribution of the robot, while each particle contains a map which represents the model of the environment, assuming the pose estimation of that specific particle (and its history, which is the estimation of the entire robot path) to be correct.

In recent years, the RBPF approach to SLAM has been very successfully used with landmark-based maps [4] as well as with gridmaps [5]. However, only very few approaches are known that use topological maps (graphs) for map representation in RBPF [6]. Taking the challenging operational environment in a shopping center and the project-specific focus on low-cost and vision-based sensor systems into account, we developed two complementary RBPF approaches using either metric gridmap models or topological graph models labeled with appearance views and metric data for environment modeling. While in most RBPF implementations the evaluation of state hypotheses (particles) is based on the compliance of the current observation with the global map through a sensor model, both of our approaches use two different types of maps for evaluation: a *global map*, which represents the already known environment model learned so far and a *local map* representing the current and the latest observations (see Fig. 3 and Fig. 5). This way, the likelihood of a given global map to be correct can be evaluated in a simple way by comparing the local and the global map directly. This has several advantages: any sensor can easily be integrated in the SLAM framework, as long as a gridmap or a topological map can be built from the sensor observations. Furthermore, because the local map can incorporate subsequent observations, it is particularly well suited for sensors with low spatial resolution or high noise (e.g. sonar), where a sequence of measurements (preferably from slightly different positions) will yield significantly richer information than a single measurement. In the following, first we will introduce our gridmap match SLAM approach and show its successful application to SLAM with low resolution/high uncertainty sonar range sensors. After that, we'll present our graph-match and vision based SLAM approach.

A. Gridmap match SLAM with RBPF

In contrast to the often used landmark representations, gridmaps [7] do not make assumptions about any specific features to be observable in the environment. They can represent arbitrary environment structures with nearly unlimited detail. However, known gridmap solutions mostly use laser scanners and exploit their high resolution and accuracy in order to reduce state uncertainty and therefore computational cost [8]. In contrast to most of the approaches that have been proposed so far, we aim at developing a SLAM algorithm that is not adapted to any specific sensor characteristics. Instead, a widely applicable model and algorithm should be used to represent the robot's observations and to build the global environment map.

For this purpose, additionally to its global gridmap, each particle contains a local gridmap (Fig. 3), which is built from

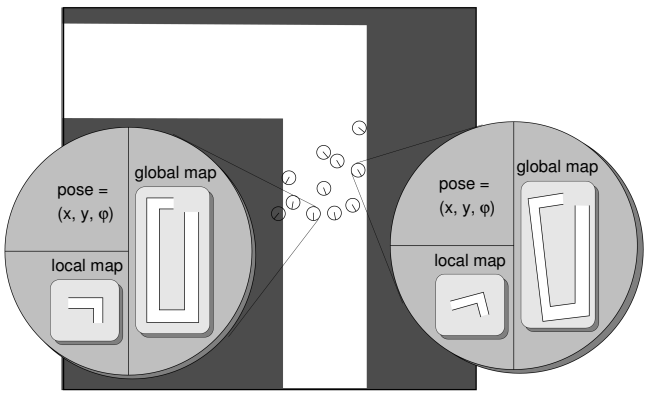

Fig. 3. Data representation overview: The particles model the distribution of the robot pose belief. Each particle contains a full map of the environment, which is a combination of the particle trajectory and the sensor observations. Furthermore, each particle contains a local map, which only contains the most recent measurements and depends on the particle's current pose belief. The situation shown is shortly before a loop closing. Apparently, the left particle is a better approximation of the true pose than the right one.

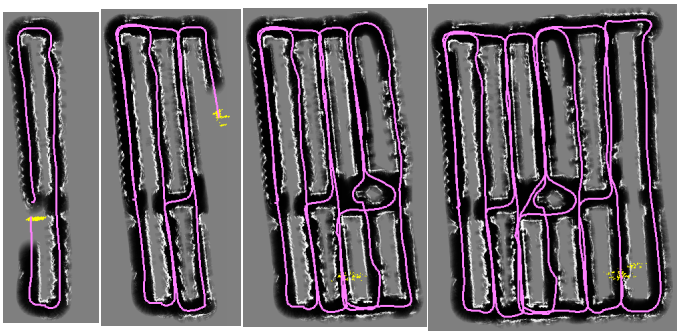

Fig. 4. Temporal evolution of the SLAM-process using 500 particles for the RBPF. In this experiment, only the robot odometry and the sonar sensors were used. The environment, the home improvement store, consists of a large number of more or less long loops of hallways with 50, 100, or 160 meters loop length. The entire path length in this experiment is about 2000 meters. In the figures from left to right, after 10, 20, 40, and 60 minutes of joysticking the robot through the store, for the best matching particles the corrected paths (magenta lines) and the corresponding maps from our Map Match SLAM approach are shown: all position errors are corrected and the obstacles (shelves, etc.) and free space in the map are clearly defined, despite the limitations of the sonar sensor and its local field of view.

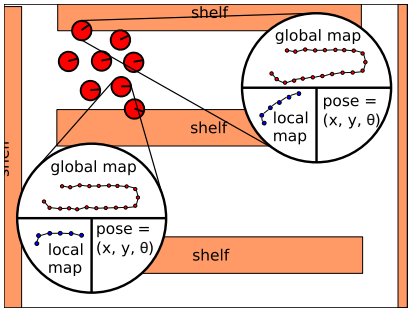

Fig. 5. The graph representation of the particles in our approach: Each particle models its current pose estimation based on its path history, a full global graph map, and a local graph. Due to the path history, all particles are slightly different and their maps differ as well. Some maps are more likely to be correct and consistent than others.

sensor observations in the same way as the global map, but consists of the most recent observations only [9], [10]. By delaying the updates for the global map until the robot has moved on for several meters, we can ensure that local and global map consist of disjoint sets of observations. Because of the delayed updates, from the start of mapping, as long as the robot moves forward, the global map at the current position is unknown, and there is no overlap between local and global map. Only in case of loop closing, i.e. when the robot returns to a known area (more precisely, when a particle believes to return to a known area), the local and global map overlap at the estimated position and can be compared. For a correct particle position estimate, local and global map of the particle should be compliant with each other (assuming a static environment), while for wrong position estimates, there will be discrepancies between the maps. Therefore, the matching of local and global map is an appropriate measure for evaluating the correctness of the particle's position estimate. The calculation of the match value between the local and global map of each particle is simple: for each occupied cell in the local map, the occupancy value of the corresponding cell in the global map is tested. The match value becomes positive if local and global map are very similar, and negative if many objects exist in the local map where there is free space in the global map. Details of this comparison are given in [10]. A major problem with using gridmaps in RBPF is the memory cost. Therefore, we introduced a simple and fast, but very efficient shared representation of gridmaps, which reduces the memory cost overhead caused by inherent redundancy between the particles (see [9]). This makes the maximum memory cost dependent on the loop size instead of the overall map size.

To verify our approach, we built maps of the "TOOM Baumarkt" home improvement store in Erfurt (Germany) which has been the public test site for all our experiments since 2002. This environment is very well suited for our online SLAM approach as it essentially consists of more or less large loops of hallways (50 to 160 meters loop length) (see Fig. 4). In this experiment, only the robot odometry and the low-quality sonar range sensors were used. The robot was moved about 2,000 meters through the operation area by joysticking while the online built global map of the best matching particle was used by the operator as indicator for necessary loop closings. A map update was not done with every observation, but in intervals: whenever the robot has moved on for 0.2 meters, first the new particle position is sampled from the odometry motion model, then the particle's map is updated with the observation at that believed position. An importance weight calculation and resampling of the particles was done in intervals of 1 meter.

After completing the on-line mapping of the whole store, the global map of the best matching particle was stored to be used later on in the routine operation with customers as global map for Monte Carlo Localization [11], [13], [2], and path planning to the article locations. Beforehand, however, this global map had to be labeled with the locations of all articles (about 60,000) available in that store. Fortunately, this could be done semi-automatically, because all article locations are stored in the merchandize management system of the store with an exact reference to the shelf number and the position within the shelf. It is an essential advantage of our approach and the used metric representation that the built gobal gridmap can easily be made fitting with the metric CAD map of the merchandize management system by a simple manual map transformation. Results of long-term field trials using an on-line built gridmap subsequently labeled with article locations are presented in Section V.

B. Appearance-based visual SLAM with RBPF

Another objective of our research was to clarify whether vison-based SLAM approaches are also suited for this type of indoor environments, and if so, how they can be made capable of working online and real-time. In our research on vision-based SLAM, we prefer the appearance-based approach for the following reasons: in a highly dynamic, populated and maze-like environment, a robust recognition of earlier selected natural landmarks cannot be guaranteed. Furthermore, the need for a robust and invariant detection of visual landmarks often results in highly computational costs and, therefore, map building is often performed off-line by these approaches. Based on own earlier experiences in view-based Monte Carlo Localization [12], [13], we developed an appearance-based visual

SLAM approach that is also using the RBPF concept [14]. Similar to the gridmap approach described before, each particle incrementally constructs its own environment model, in this case a graph-map, which is labeled with visual observations and the estimated poses of the places where the snapshots were taken from.

As observations we utilize holistic appearance features extracted from panoramic snapshots obtained from the 360° camera located at the top of our experimental robot platforms (see Fig. 2). A number of possible appearance-based features has been studied in our lab with respect to their capability to visually distinguish neighbored positions in the environment, the computational costs, and the preservation of similarities under changing illumination conditions and occlusions. The respective features experimentally investigated include local RGB mean values [12], HSV-histograms [13], FFT-coefficients [15], and SIFT features [16], [17] as appearance-based image description. Considering the findings and constraints of these investigations (see [14]), in our final implementation we decided in favor of the HSV histogram features (lower memory usage and computational costs than SIFT), especially because of the real-time requirements of our SLAM approach and the high robustness of these features.

Basic idea of this approach: Our appearance-based SLAM approach also utilizes the standard RBPF approach to solve the SLAM problem, where each particle contains a pose estimate $\mathbf{x}_i$ (position x, y and heading direction φ) as well as a map estimate (see Fig. 5). In contrast to the gridmap-based approach described before, the environment model used here is a graph, where each node i representing a place in the environment is labeled with the features $\mathbf{z}_i$ extracted from the panoramic view captured at that place and the estimated pose $\mathbf{x}_i$. As in our gridmap-based approach, here we again use the two different types of maps, a *global map* representing the already known environment model, and a *local map* representing the current and the latest observations, and the local path between them. This way each particle (see Fig. 5) estimates and stores a local map, a global map, and the currently estimated pose. Serving as a short-term time-window of observations, the local map is used to compute the likelihood of each global map to be correct (Fig. 6). This has a substantial advantage: the local map provides both geometric and visual information about the lastly observed places. This way, correct comparisons can be made taking both spatial relations and visual observations into consideration. We achieved the best results with a local trajectory length of about 2 meters and an average distance between the nodes of 0.2 meters.

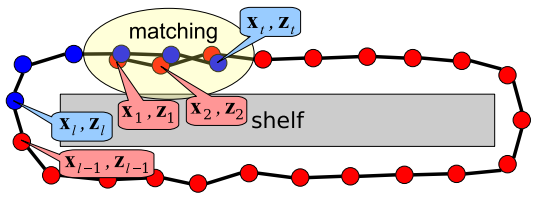

Fig. 6. Graph-based environment representation of our appearance-based approach: The red nodes show the global map of a single particle with respect to the path estimated by this particle. The blue nodes code the local map, whose data represent a short-term time-window of observations (the current and the n latest observations) used for map matching to determine the likelihood of the respective global map. The idea of our appearance-based RBPF is, that only particles with correctly estimated trajectories are able to build correct maps, so that the matching between local and global map provides a higher matching value than wrongly estimated trajectories.

To determine the importance weight of each particle, their local and the global graph maps have to be compared. Thereto, corresponding pairs of nodes in both maps are selected by a simple nearest neighbor search in the position space. The relation between each pair of corresponding nodes i and j of the local and global map provides two pieces of information, a geometric one (spatial distance d_{ij}) and a visual one (visual similarity S_{ij}). Both aspects of each relation ij are used to determine a matching weight w_i for the respective node i of the local map. In the context of appearance-based observations, the visual similarity between observations is not only depending on the difference in position but also on the environment itself. If the robot, for example, moves in a spacious environment with much free-space, the similarity between observations from slightly different positions will be very high. In a narrow environment with many obstacles, however, observations at positions with low spatial distance are already drastically influenced, which leads to low visual similarities. To that purpose, we developed an adaptive sensor model that estimates the dependency between surrounding-specific visual similarities $\hat{S}_{ij}$ of the observations $\mathbf{z}_i$ and $\mathbf{z}_j$ and their spatial distance $\hat{d}_{ij}$. Different approaches to approximate this sensor model have been investigated, e.g. Gaussian Process Regression (GPR) as a non-parametrical description, or a parametrical polynomial description. Since the model has to be computed after each motion step, and applied to each node of the local map of each particle during the graph-matching process, we decided for the parametrical polynomial description. Mathematical details of the developed adaptive sensor model and the graph comparison to determine the importance weight of each particle, i.e. the likelihood of the respective global map, are given in [14].

The computational and the memory costs of this algorithm only depend linearly on the number of particles and quadratically on the number of nodes in the local maps required for graph matching. At a single-core CPU with 1.8GHz, the computation of the adaptive sensor model is done in approximately 60 ms (independent from map and particle size), whereby 20 nodes of the local map were used for the estimation of the similarity model. The rest of the computational cost is spent for map updates, weights determination, and the resampling process. For a small number of 250 particles as used in the following experiment, one complete iteration cycle requires 0.08 s, which allows real-time and on-board mapping.

Experimental results: For the experimental investigation, we used the aforementioned home improvement store in Erfurt again. All data for the analysis were recorded during routine operation of the store, i.e. people walked through the operation area, shelves were rearranged and with that their appearance, and other dynamic changes (e.g. illumination, occlusions)

Fig. 7. The red path shows the robot's movement trajectory estimated by our appearance-based visual SLAM approach. The map shows a high accuracy with only small alignment errors. Only for visualization of the localization accuracy, an occupancy map was built in parallel using raw laser data arranged along the pose trajectory estimated by means of our appearance-based SLAM approach and superimposed to the estimated movement trajectory.

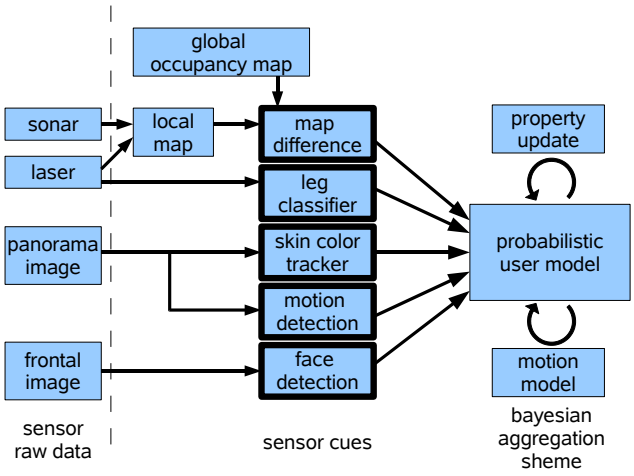

Fig. 8. Architecture of the multimodal tracker, left: the sources of information, middle: preprocessing systems, delivering Gaussian hypotheses, right: probabilistic model and update mechanism.

happened. In this experiment, the robot was moved several times through the home store along repeated loops around the goods shelves. The resulting graph (Fig. 7) covers an area of 20 x 30 meters and was generated by means of 250 particles in the RBPF. Thus, 250 global graph maps had to be built, whereas Fig. 7 only shows the most likely final trajectory and an occupancy map superimposed for visualization only. This occupancy map built by laser scannings and the visually estimated path gives an impression of the high quality and accuracy of our appearance-based SLAM approach. It only shows marginal alignment errors, all hallways and goods shelves are arranged very precise along straight lines. To evaluate the visually estimated path shown as red trajectory in Fig. 7, in addition a ground truth path and map built by means of a laser-based RBPF SLAM algorithm were calculated. A first result was, that the trajectories estimated by both SLAM approaches are very similar. This is also expressed by a mean localization error of 0.27 meters (with a variance $\sigma \approx 0.13$ meters) compared to the laser-based SLAM results. The maximum position error in this experiment was 0.78 meters. These experimental results demonstrate, that our vision-based approach is able to create a consistent trajectory and, for this reason, a consistent graph representation too. Furthermore, in contrast to the gridmap approach introduced before, topological maps require significantly less memory because of the efficient observation storage, where the feature sets of each snapshot are shared and only repeatedly linked.

IV. Robust User Detection and Tracking

The aim of a shopping companion is to assist customers during their purchase. Therefore, at first, people, who seem to need assistance have to be found, while the system is patrolling through the operational area. Indications for the interest of customers to interact with the robot are given, when a person is standing still and facing the robot for a while. During the guided tour, the robot has to continuously observe the user to detect if the person stops or keeps on following the robot. For this purpose, usually visual cues for face detection and tracking are used, but for our application, tracking people only in image space is not sufficient. The goal is to track them in a two dimensional reference frame, allowing to estimate their distance to the robot, and to infer about their behavior and movement trajectories. Therefore, in [18] 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 in mobile robotics, e.g. in [19], [20], [21], [22]. 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 as primary sensor 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. Due to these findings, for the application of the shopping assistant, we concentrated on the improvement of the already existing sensor fusion method [18], which additionally has the advantage that the perceivable area is not limited by the range of one sensor. In [23], the algorithm of [18] had already been improved by representing the position of people in an Euclidean world space, allowing to generate movement trajectories and to estimate their velocity. For the application presented here, however, the model had to be extended again by components needed for making decisions on the robot's behavior.

A. Multi-modal probabilistic person tracker

As mentioned above, different sensory cues are used for estimating the positions of people nearby the robot. The main sources of information are the images and the occupancy map of the local environment, integrating all the range information from sonar and laser (see Fig. 8). The centered column of Fig. 8 shows the preprocessing modules. Each

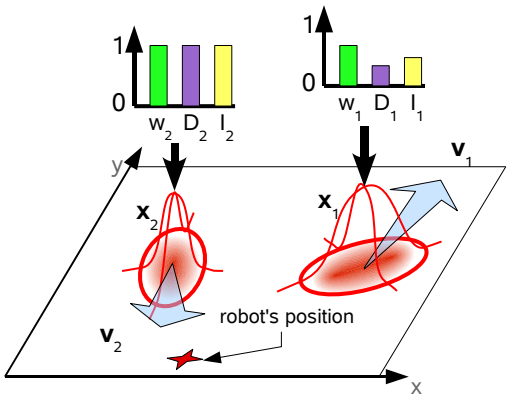

Fig. 9. Two hypotheses of the probabilistic user model are shown. Each one has a position $\boldsymbol{x}$ and a velocity $\boldsymbol{v}$ which are modelled by Gaussians in 2D world coordinates. Further, for each hypothesis the probability of being a human w, the need for interaction I, and the probability of being the current dialog partner D are modeled by discrete distributions over $\{0, 1\}$.

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. 8 is using these observations to estimate the users' positions and further properties, like the estimated need for interaction, or the probability of being the current dialog partner (see Fig. 9). In extension to the algorithm of [18] and [23], 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.

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 of the sonars in a local occupancy gridmap (similar to that one used for SLAM), representing the current local surroundings of the robot. Based on this occupancy map, 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. Objects in that virtual scan are then classified as person or background. For that purpose, a second scan $R_g(n)$ is extracted for the estimated pose from the global occupancy map built during SLAM (see Section III). By comparing these two scans, Gaussian observation hypotheses of *non-static, moveable objects* can be generated [24]. Because in a dynamic environment, the probability for these detections to be a person is relatively weak, the hypotheses resulting from this cue only get a low weight w_s. By means of a set of simple rules, *leg pair hypotheses* can be generated from the laser scan as described in [24]. For each of these pairs, a Gaussian observation hypothesis is generated at the average position of the legs. Due to the cluttered environment, the reliability w_s of observing a person this way is also chosen as low.

To find and track regions of *skin color* in the panoramic image, a set of particle filters is used. The concept introduced in [25] allows to track multi-modal distributions of skin color in the image resulting in a set of disjoint hypotheses for people's skin regions. To generate observation hypotheses in 2D world coordinates, the average direction $\bar{\varphi}$ and a typical distance to the robot $r_{fix} = 1.5$ m 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 particles' variance in φ. The average weight $\bar{\pi}$ of all particles of the respective filter then defines the weight w_s for these hypotheses.

A reliable feature of humans 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. This allows the utilization of a simple difference image based approach, which gives a reliable hint for motion in the panoramic image (for more details see [24]).

A very promising hint for people in the surroundings of the robot is a *face* in the image. For finding faces, the well known Viola and Jones detector [26] is applied. The likelihood for an image patch to be a face is estimated, which is yielding a further reliability w_s. For generation of Gaussian observation hypotheses, besides the given direction, a distance is needed again. The typical size of faces (about 15 cm) is utilized to triangulate the distance with a deviation of about 40 cm.

All these sensory systems are chosen to complement each other. We have range 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 low spatial accuracy, but with a higher selectivity. Thus, the main task of skin color is to generate hypotheses at distinct directions to the robot. Later, they can be refined in their distances 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 user model up to date, which is introduced subsequently.

B. Probabilistic user model

The probabilistic user model for tracking potential interaction partners in the surroundings of the robot is illustrated in Fig. 9. There is a varying number of hypotheses $H_k = (\boldsymbol{x}_k, \boldsymbol{v}_k, I_k, D_k, w_k)$, 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 $\boldsymbol{C}_k$ and $\boldsymbol{\nu}_k$ and $\boldsymbol{V}_k$ respectively. 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_k and the probability D_k of being the user who is currently in dialog with the robot are estimated internally. I_k is used to decide whether to offer service to a detected person while

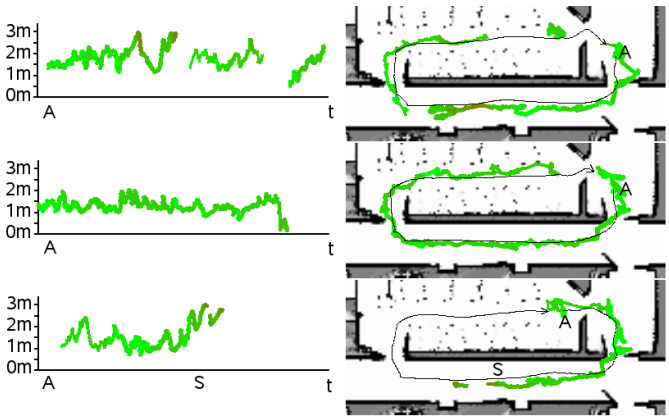

Fig. 10. Exemplary tracking results of a person instructed to follow the robot. (Right) trajectories of robot (black) and user (green) following in different distances, top: $2m$, middle: $1.5m$, bottom: stopping at point S. (Left) estimated distance of the robot to the user over time.

searching for new users. The position of the hypothesis with maximum D_k and a 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. Algorithmic details of the update process for the user model are described in [24].

C. Experimental results in user tracking

In order to evaluate the ability to track a customer during a guided tour, the robot gave several tours, while the customer had to follow the robot in a particular distance. In Fig. 10 three of these runs are shown. The plots visualize the continuity of the tracking result for different average distances. During the tours with a length of about 50 m, the robot lost contact only two times in five trials, where the people where asked to keep a distance of about 1 to 1.5 m, which is a normal displacement for people doing a shopping tour. In a second experiment, the number of false positive hypotheses was evaluated. Here, the robot should interrupt its tour, if a customer stopped following. Therefore, at the marked point S in Fig. 10 the customer stopped, while the robot continued its guiding tour. In all of these trials, the robot detected the disappearance of the user correctly after at most 5 m and interrupted its guidance tour.

V. Acting as Shopping Robot - Field Trials

Our long-term field trials started at the end of March 2008 and still ongoing (at the end of July 2008) aim at studying whether and how a group of interactive mobile shopping robots can operate completely autonomous without any assistance by roboticists in an everyday environment, an ordinary shopping center, and how they are accepted by non-briefed customers. To this purpose, a home improvement store with a size of about 7,000 square meters situated in Bavaria (Germany) was equipped with 3 robotic shopping assistants allowing a working in parallel. With the objective of finding people who might need assistance, the robots move around in the store and patrol between particular points of interest. If one of the robots detects a potentially interested customer during its tour (see Section IV), it stops and offers its service by using a short voice output like ”Hello! May I help you to find an article?” If the customer is interested in assistance, he/she starts the communication with the robot by pressing the ”Start” button on the touch screen. Then, the menu for selecting the several modi for goods search are displayed and also verbally explained. If the customer has found the requested article in the database, a map of the store is shown, in which the current position, the position of the article in the store, and the suggested path planned by the robot are drawn in. If the customer presses the button ”Go”, the robot moves along its planned path to the requested article. At the arrival point, the robot drives a bit aside, so that the customer has enough space for choosing the preferred article, and offers additional services like a video conference to a salesperson, a price scanning, or a new search. If the customer wants to bring the assisted shopping process to an end, she can finally press a ”Good bye” button. To get an impression whether the users are contented with the robots or not, we inserted two dichotomous questions on the last screen page: First, the robot asks if the customer was satisfied, and then, if she would use the shopping robot again in the future.

During the first four months of this still ongoing long-term field test, there were about 210 kilometers traveled by each robot on average. The run-time was limited by the battery capacity and the decision to use not more than two robots at the same time in the store. With a full battery charge, the robots are running between 6 and 8 hours. The mean daily run time was 4.1 hours a day. All in all, 3,764 customers have started the dialogue with the robots. On average, a customer wanted to reach 1.53 arrival points during an interaction. The ”Go” button was pressed 5,781 times; if we adjust this figure by subtracting the events when the customer has manually stopped the guidance process, we get 2,799 real guidance tours. Based on the last named figure, in 83,7% of all cases the robots have reached their arrival points. The price scanning function was used 382 times. The results from the survey at the end of the interaction are very promising; about a fifth of the users has taken an interest in giving a feedback. 93,4% of the interviewed persons are contented with the shopping robot, and 92% would use the robot again.

Finally, we made the experience that the *patrol mode* is indeed a good choice to advise the customers to the shopping robot, but there are times, where it does not make sense. Particularly, in the case of very crowded stores during the rush hours, it is better that the robot remains at a waiting position defined in advance, otherwise it happens, that it obstructs a passage without giving a service. Furthermore, due the fact that a shopping robot is an unexpected innovation in a store, the customers should be advised of the services of the robot. This increases the numbers of interactions and, therewith, the benefit for the customers, who want to find articles faster with the guiding help of a robot. Our experiments were successful in two main dimensions. First, they demonstrated the robustness of the various probabilistic techniques in a really challenging real-world task. Second, they provided some evidence towards

the feasibility of using mobile robots as assistants to people without any background knowledge in robotics.

VI. Related Work and Discussion

A comprehensive overview of the employment of mobile robots as tour guides in expositions or museums is given in [27]. Among them are such well known robots like Rhino, Minerva, and Sage, or the exposition guide RoboX. Usually, these robots guide visitors to a set of predefined exhibits following a planned path while offering exhibition-related information. They navigate in densely populated, structured, but completely known operation areas as they are typical for expositions - in the most cases limited to a few hundred square meters. Unlike this, the shopping robot scenario and the respective operational environments in supermarkets or home stores make more challenging demands. Amongst others, there are the very large operation area up to 10,000 square meters, the uniformly structured, maze-like environment consisting of numerous similar, long hallways, the enormous total length of all hallways (up to several kilometers), and the continuous changing of the environment (e.g. due to rack filling or rearrangements). Besides the work presented here, there are only a few approaches known that tried to meet at least some of these challenges. RoboCart [28], or the approaches of [29] and [30] belong to this. However, these systems don't navigate autonomously, some require engineered environments (e.g. equipped with RFID-tags for localization), others are remote-controlled from outside [30]. They are proof-of-concept prototypes or even only design concepts, show very limited functionality only, often require assistance by roboticists, and are not yet ready for everyday use. None of these systems has continuously been involved in shopping tasks over longer periods of time, or was subject of long-term field trials during routine operation of the store.

VII. Summary and Conclusions

This paper describes recent progress in developing a mobile shopping robot for interactive user guidance in shopping centers. Building on former work in robot navigation and human-machine interaction, new approaches specifically aimed at robust navigation and human-robot interaction in complex and populated public environments, and results of still running long-term field trials are presented. With the successful development of the world's first shopping robot, that supports people in looking for goods in a completely autonomous fashion, one important step towards assistive robotics for everyday use has been completed. Our shopping companion intuitively offers its services, keeps continuously contact to the current customer, and guides its user to the desired good's location. In four-month field trials, the robot has shown its suitability for a challenging real-world application, as well as the necessary user acceptance. The companion robot offers unequaled opportunities and demonstrates that such service robots can be usefully utilized in a variety of applications in daily life. Areas of application, that could benefit from such interactive, mobile robot assistants, are all those, where people have need for individual assistance along their way, because of missing knowledge of place and contact persons, respectively.

References

[1] H.-M. Gross and H.-J. Boehme, "Perses - a Vision-based Interactive Mobile Shopping Assistant", *in Proc. IEEE-SMC'00*, 80–85

[2] S. Thrun, W. Burgard, D. Fox, *Probabilistic Robotics*, MIT Press, 2005.

[3] K. P. Murphy, "Bayesian map learning in dynamic environments," in *Proc. NIPS'99*, 1015–1021.

[4] M. Montemerlo et al. "FastSLAM: A factored solution to the simultaneous localization and mapping problem", in *Proc. AAAI'02.*, 593–598

[5] D. Haehnel, W. Burgard, D. Fox, and S. Thrun, "An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements," in *Proc. IROS'03*, 206–211.

[6] A. Ranganathan, F. Dellaert, "A Rao-Blackwellized particle filter for topological mapping", in *Proc. ICRA'06*, 810 - 817

[7] H. Moravec, "Sensor fusion in certainty grids for mobile robots", *AI Magazine*, 9:61–77, 1988.

[8] A. I. Eliazar and R. Parr, "DP-SLAM: Fast, robust simultaneous localization and mapping without predetermined landmarks", in *Proc. IJCAI'03*, 1135 – 1141.

[9] C. Schroeter, H.-J. Boehme, H.-M. Gross. "Memory-Efficient Gridmaps in Rao-Blackwellized Particle Filters for SLAM using Sonar Range Sensors", in *Proc. ECMR07*, 138-143

[10] C. Schroeter, H.-M. Gross. "A sensor-independent approach to RBPF SLAM - Map Match SLAM applied to Visual Mapping", in *Proc. IROS'08*

[11] D. Fox, W. Burgard, F. Dellaert and S. Thrun, "MCL: Efficient Position Estimation for Mobile Robots", in *Proc. AAAI'99*, 5398 – 5403.

[12] H.-M. Gross, A. Koenig, H.-J. Boehme, and C. Schroeter, "Vision-based Monte Carlo self-localization for a mobile service robot acting as shopping assistant in a home store", in *Proc. IROS'02*, 265–262.

[13] H.-M. Gross, A. Koenig, Chr. Schroeter and H.-J. Boehme. Omnivision-based Probabilistic Self-localization for a Mobile Shopping Assistant Continued. In *Proc. IROS'03)*, 1505–1511.

[14] A. Koenig, J. Kessler, H.-M. Gross, "A Graph Matching Technique for an Appearance-based, visual SLAM-Approach using Rao-Blackwellized Particle Filters", in *Proc. IROS'08*

[15] E. Menegatti, M. Zoccarato, E. Pagello, and H. Ishiguro. "Hierarchical Image-based Localisation for Mobile Robots with Monte-Carlo Localisation." in: *Proc. ECMR'03*, 13-20.

[16] D. G. Lowe. "Object recognition from local scale-invariant features." In *Proc. ICCV'99*, 1150–1157.

[17] H. Andreasson, T. Duckett, and A. Lilienthal. "Mini-SLAM: Minimalistic visual slam in large-scale environments based on a new interpretation of image similarity." In *Proc. ICRA'07*, 4096–4101.

[18] C. Martin, E. Schaffernicht, A. Scheidig, and H.-M. Gross. Sensor fusion using a probabilistic aggregation scheme for people detection and people tracking. *Robotics and Autonomous Systems*, vol. 54:721-728, 2006.

[19] D. Schulz, W. Burgard, D. Fox, and A. Cremers. "Tracking multiple moving objects with a mobile robot", In *Proc. CVPR'01*, 371–377.

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

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

[22] J. Fritsch et al. "Audiovisual person tracking with a mobile robot", In *Proc. ICRA'04*, 898–906.

[23] A. Scheidig, S. Müller, C. Martin, and H.-M. Gross. "Generating person's movement trajectories on a mobile robot", In *Proc. ROMAN'06*, 747–752

[24] St. Mueller, E. Schaffernicht, A. Scheidig, H.-J. Boehme, H.-M. Gross. "Are you still following me?" in: *Proc. ECMR'07*, 211-216

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

[26] P. Viola and M. Jones. "Fast and robust classication using asymmetric adaboost and a detector cascade", In *Proc. NIPS'01*, 1311–1318.

[27] B. Jensen et al. "Robots meet Humans - Interaction in public Spaces", *IEEE Trans. Industr. Electronics* 52: 1530-46, 2005

[28] V. Kulyukin et al. "RoboCart: Toward Robot-Assisted Navigation of Grocery Stores by the Visually Impaired." in: *Proc. IROS'05*, 2845-50

[29] M. Shieh, Y, Chan, Z. Lin, J. Li. "Design and Implementation of a Vision-Based Shopping Assistant Robot" *in Proc. SMC'06*, 4493-98

[30] T. Tomizawa, A. Ohya, S. Yuta. "Remote Shopping Robot System", in: *Proc. IROS'06*, 4953 - 4958