in: Proc. IEEE Intern. Conf. on Systems, Man and Cybernetics 2000, Nashville, pp. 80-85, IEEE/Omnipress

PERSES - a vision-based interactive mobile shopping assistant

Horst-Michael Gross, Hans-Joachim Boehme
Ilmenau Technical University, Department of Neuroinformatics, 98684 Ilmenau, Germany
Horst-Michael.Gross@informatik.tu-ilmenau.de http://cortex.informatik.tu-ilmenau.de

Abstract

The paper describes the general idea, the application scenario, and selected methodological approaches of our long-term research project PERSES (PERsonal SErvice System). The aim of the project consists in the development of an interactive mobile shopping assistant that allows a continuous and intuitively understandable interaction with a customer in a home improvement store. Typical tasks we have to tackle are to detect and contact potential users in the operation area, to guide them to desired areas or articles within the store or to follow them as a mobile information kiosk while continuously observing their behavior. Due to the specificity of the interaction-oriented scenario and the characteristics of the operation area, we have focused on vision-based methods for both human-robot interaction and robot navigation. Besides some methodological approaches, we present preliminary results of experiments achieved with our mobile robot PERSES in the store with an emphasis on vision-based methods for user localization, map building and self-localization.

1 Introduction

The project PERSES (PERsonal SErvice System) aims to develop an interactive mobile shopping assistant that allows a continuous and intuitively understandable interaction with a human user (customer). Such a shopping assistant must be able to actively observe its operation area, to detect, localize, and contact potential users, to interact with them continuously, and to adequately offer its specific services (Fig. 1). Service tasks we want to tackle are to guide the user to desired areas or articles within the store *(guidance function)* or to follow him as a user-specific mobile information kiosk while continuously observing the user and his behavior *(companion function)*. Intuitively understandable human-robot interaction should at least entail visual and acoustic components. In the context of our application scenario as mobile shopping assistant, we defined the following interaction and navigation tasks, presented as a mix resulting from the interaction sequence and, more general, functional necessities: (a) visual localization of a potential user within a pre-defined operation area, (b) acoustic localization of a potential user clapping his hands or shouting a command to attract attention, (c) fast learning of an initial visual model of the current user and online adaptation of that model due to the varying appearance of the user in the course of the shopping process, (d) robust vision-based user tracking both while standing still and during self-movement of the robot, (e) robust avoidance of static and dynamic obstacles during navigation, (f) continuous self-localization of the robot in the operation area, (g) navigation to desired places, articles, or market areas acting as a guide, (h) recognition of simple spoken commands, and, for the future, (i) recognition of gesticulated user instructions. This spectrum of tasks necessitates adaptive methods at all processing levels using (i) neural networks for visual and acoustic scene analysis and sensorimotor control, (ii) probabilistic methods for map building, robust self-localization, local and global navigation, and mission planning and reasoning, and (iii) concepts from Machine Learning and Control Theory for dynamic coordination of the subsystems responsible for the several interaction and naviga-

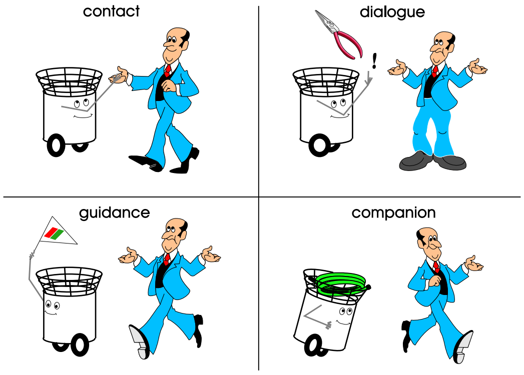

Figure 1: *Necessary skills and typical service tasks of a user-oriented, interactive mobile shopping assistant.*

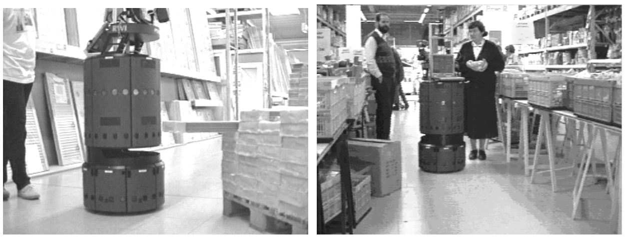

Figure 2: *Our experimental platform* PERSES *operating in a home improvement store, a cluttered and un-engineered environment with numerous critical obstacle configurations.*

tion tasks. To master the specificity of this interaction-oriented scenario and the characteristics of the operation area, a home improvement store, we have focused on vision-based methods for both the interaction and the navigation process. The operation area is characterized by many similar long hallways of equal width and a great number of critical obstacle configurations, for example, objects hanging down from the ceiling or jutting out of shelves, lost shopping carts in the hallways, etc. Many of these obstacles cannot be perceived reliably by distance sensors (Sonar, Laser) which operate in certain planes in 3D space. By contrast, vision-based approaches do not show these limitations, they supply a much greater wealth of information about the structure of the local surroundings. The examples in Fig. 2 are to illustrate some of the challenges of the application scenario. The robot PERSES we use as experimental platform in our experiments in the store is a standard B21 robot additionally equipped with sensor systems for interaction and navigation (Fig. 3).

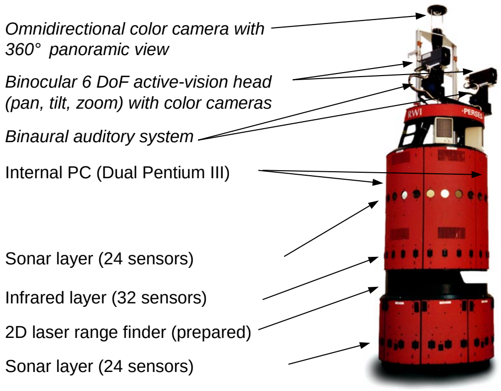
Omnidirectional color camera with 360° panoramic view
Binocular 6 DoF active-vision head (pan, tilt, zoom) with color cameras
Binaural auditory system
Internal PC (Dual Pentium III)
Sonar layer (24 sensors)
Infrared layer (32 sensors)
2D laser range finder (prepared)
Sonar layer (24 sensors)

Figure 3: *Experimental platform PERSES, an extended version of a standard mobile robot B21 by RWI. In addition to the standard equipment of two sonar and one IR-layers, PERSES is equipped with (i) an omnidirectional color camera with a 360° panoramic view used for user localization and tracking, self localization and local navigation, (ii) a binocular 6 DoF active-vision head with 2 frontally aligned color cameras used for user verification and tracking, odometry correction and obstacle avoidance, and (iii) a binaural auditory system for acoustic user localization and tracking.*

2 Overview of the system

Because of the complexity of the "shopping-task" as a whole, we use a behavior-based approach which allows us to decompose the problem into separate behavior modules responsible for several subtasks of the interaction and navigation cycle. As formal framework for behavior coordination, we chose the so-called dynamic approach to robotics [10]. The PERSES-architecture consists of three main subsystems: *User Localization*, *User LogIn* and *Interactive Tour* (Fig. 4).

User Localization: This subsystem is responsible for the robust localization of a potential user in the surroundings. At present, we use a multimodal approach that integrates both visual and acoustic stimuli. The submodule *Visual User Localization* performs a motion-based foreground-background segmentation in the image sequence provided by the omnidirectional camera (Fig. 4 - middle left), and returns the angle to the center of gravity of the largest moving region. While standing still, the motion-based segmentation calculates some candidate regions that indicate if and where potential users could be. Our implemented method is similar to that suggested in the *Pfinder* system [13]. The integration of auditory saliency makes it easy for the user to attract the attention of the robot and to speed up the localization process. For the acoustic localization of a customer clapping his hands or shouting a command, we developed a biologically inspired binaural 360° sound localization system that considers essential functional aspects of the processing in the auditory brainstem and midbrain. This subsystem realizes (i) the detection of the sound direction in the horizontal half-planes by processing of the interaural time-delays (ITD) and (ii) a simple but effective front-behind discrimination on the basis of the differences in the spectral shapes of the left and right sound stream supplied by the microphones mounted on top of PERSES (Fig. 3). Details of this model and localization results are presented in [9].

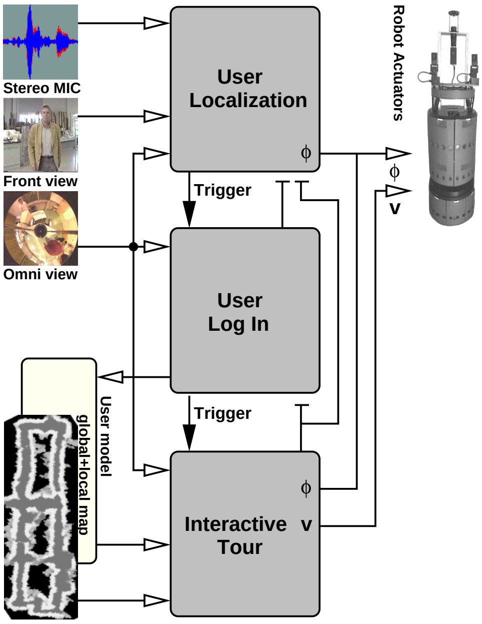

Figure 4: *PERSES system architecture.*

Both submodules *Visual User Localization* and *Acoustic User Localization* make use of the same actuator, namely, they try to turn the robot towards the detected potential user in order to verify the localization hypotheses by means of the frontal cameras (Fig. 4-top left). Due to the turn of the robot, the potential user should be localized in front of the robot allowing the frontal cameras to observe him and to evaluate if he could be willing to interact with the shopping assistant. As a very simple criterion, we assume that a customer may be considered to be a user possibly willing to interact if his face and his upper part of the body are oriented towards the robot. To realize a robust verification of a user localization hypothesis, we use a task-specific multi-cue saliency system that integrates different visual cues: skin color, head-shoulder contour, and facial structure. This way, the system becomes more robust, can handle varying environmental conditions and is less dependent on the presence of any specific feature. Some details of this subsystem are presented in Section 3.1.

User Login: When a potential user has been found and confirmed, and the user has started to interact (by speech and/or touch screen), a visual model of the user is learned, which can be used in the course of the interactive tour to track the current user and to distinguish him from other customers, if he was lost from view. Additionally, this subsystem has to ask the user for the article or area he is looking for. This is also realized by a simple interactive dialog by touch screen. Since development of most parts of this subsystem has only begun, we will not present any experimental results for them.

Interactive Tour: This subsystem is initiated when the *User LogIn* subsystem provides the position of a desired area or article in the store. In this case the internal module *User Guidance* has to plan a route to the desired position. For map building, self-localization, and global navigation, we use very efficient statistical and probabilistic techniques [8, 11, 4, 6, 7]. We currently extend them to the specific visual inputs provided by the on-board cameras (see Section 3.2 and 3.3). In case a user is present, the internal *User Tracking* module is active, too. This module's goal is to realize the companion function by keeping the user within the omnidirectional view. When the user falls behind or moves in another direction, this module takes over control by inhibiting the *User Guidance* module in order to follow the user. Another task of the *User Tracking* module is the online adaptation of the visual user model in order to cope with the varying appearance of the user in the course of the shopping process. Both the *User Guidance* and the *User Tracking* modules compute motor commands for navigation. Before execution, they are passed to an *Obstacle avoidance* module which suppresses those commands impossible according to the current obstacles in front of the robot. The need for supplemental vision-based methods for obstacle avoidance arises from the circumstances mentioned earlier that numerous obstacles cannot be perceived reliably by 2D distance sensors (sonar, laser) because of their specific form, size or height (e.g., boards or pipes jutting out of shelves). In this context, local navigation methods from ecological robotics [5] based on optical flow and inverse perspective mappings of the panoramic image are currently investigated in our lab.

3 Selected methods and results

Of the tasks listed earlier, we consider those of user and robot localization to be of central importance. Therefore, we present methods and results for robust finding of a user in spite of crowded environments containing background disturbances, as well as for vision-based map building, self-localization and position tracking of the robot in the operation area. Detailed emphasis is placed on the latter two aspects.

3.1 Visual user localization/verification

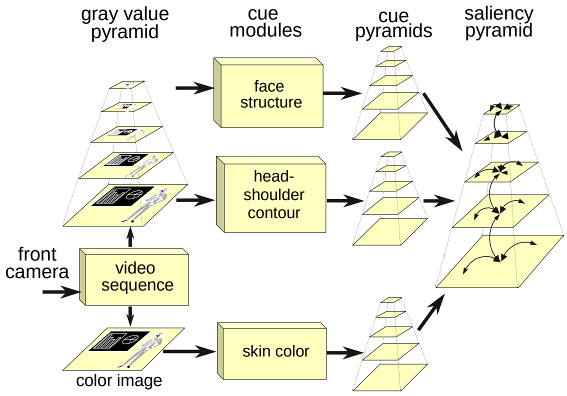

Figure 5: *Multiple-cue approach for user verification.*

To realize the verification of a motion- or acoustic-based user localization hypothesis we use a task-specific saliency system that integrates different visual cues: facial structure, head-shoulder-contour and skin color. This subsystem should highlight all regions that most likely cover the upper part of a person. Figure 5 provides a coarse sketch of our multiple-cue approach. A multiresolution pyramid transforms the images acquired

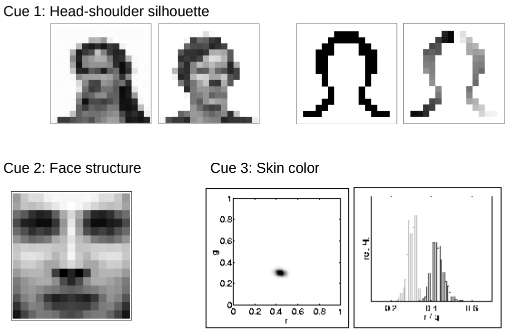

Figure 6: *Overview of the cues used for user localization and verification. (Top-from left to right:) two samples from the data set used for learning the silhouette; binary contour; orientation contour. Local orientation angles are coded by gray values (*0°*: black;* 90°*: medium gray;* 180°*: white).*

by one of the front-cameras into a multiscale representation. Because we want to localize people at different distances from the robot, we use a number of resolution levels. The cue modules sensitive to facial structure and head-shoulder contour operate at all levels of the grayscale pyramid, while the cue module for skin color detection uses the original color image. The output of the cue modules serves as input for the pyramid of saliency maps. To achieve a stable localization result, we utilize dynamic neural fields [1] for selection between alternative hypotheses within the saliency pyramid. To work with the multi-scale representation, we extended the original 2D neural field approach of AMARI to a 3D neural field for selection of the most salient region in depth. Fig. 7 presents typical localization results obtained in the highly structured environment of the home improvement store. More details of our multi-cue approach for user verification can be found in [3].

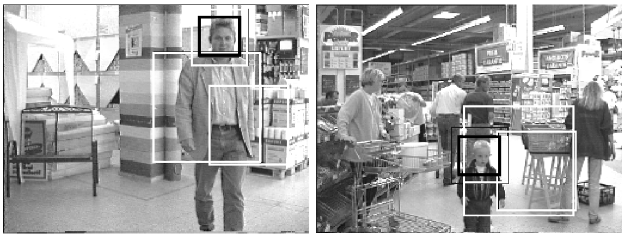

Figure 7: *Results of user localization with the multi-cue approach. Localization hypotheses at different levels of scale space are marked by white frames. The size of the frames corresponds to the respective level of the scale space, small frames correspond to levels of high-resolution and vice versa. Final localization results, are marked as black frames. In the right figure showing a crowded area in the store, the child is selected as final localization result because it is the only subject that fulfills all 3 criteria of our multi-cue approach: face and upper part of the body are oriented frontally towards the robot, skin color can be detected clearly.*

3.2 Visually-controlled map building

To navigate reliably in indoor environments, a robot must know where it is. This includes both the ability of globally localizing the robot from scratch, as well as tracking the robot's position once its location is known. In PERSES, we use two types of maps for self-localization and navigation: (i) grid-based occupancy maps and (ii) a grid of panoramic views of local surroundings (see Section 3.3). The maps are learned from sensor data (sonar, images, odometry) collected when manually joy-sticking the robot through the store or autonomously exploring the operation area. One major problem using odometry data is their increasing error over time, especially concerning the rotation angle. To attenuate this effect, we utilize a specific feature of all home improvement stores. Typically, their floor shows a rectangular structure caused by tiles which are uniquely oriented across the whole store. The idea is quite obvious: a further top-down oriented on-board camera acquires images of the floor in front of the robot (Figure 8). By continuously estimating the dominant orientations within these images, we can calculate the accurate orientation of the robot and, therefore, substitute the rotation angle supplied by odometry by the orientation determined visually. Hence, it is possible to eliminate the orientation error, and subsequently, the position error. If the initial position and orientation of the robot are known, this method allows an accurate, iterative position tracking as required for map building. Figure 9 exemplarily illustrates the efficiency of this specific method for vision-based odometry correction for map building. Of course, the proposed approach does not hold in a more general framework, but is very well suited for our specific environment.

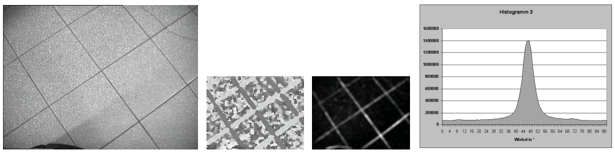

Figure 8: *General idea of our vision-based odometry correction considering a specific feature of the market floor: a) image of the floor in front of the robot, b) local orientation tensors (orientations are coded as gray values), c) confidences of local orientations (low-black, high-white), d) histogram of confidence-weighted local orientations, the dominant orientation (center of gravity) is a significant measure for the accurate orientation of the robot in the interval* $0^\circ - 90^\circ$

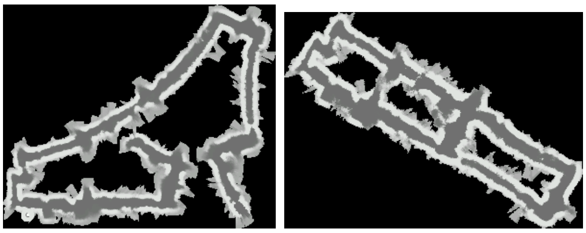

Figure 9: *Sonar-based occupancy maps of a store section (60 by 20 meters; total path length: 250 m). Gray-values code occupancy probabilities: white - occupied (obstacle), gray - free-space, black - unexplored. (Left) without vision-based odometry correction: the closed-loop course cannot be closed, because the error of the odometry-based estimation of the rotation angle finally amounts to* 90°*. The result is an unusable map. (Right) with vison-based odometry correction, now the closed-loop course can be closed exactly.*

3.3 Visual Monte Carlo Localization

The topology of the store area is characterized by many similar, long hallways of equal width. For this reason, self-localization methods based on distance sensors can produce numerous ambiguities preventing a quick self-localization and relocalization in case of a complete loss of positioning. Because the visual input from the omni-

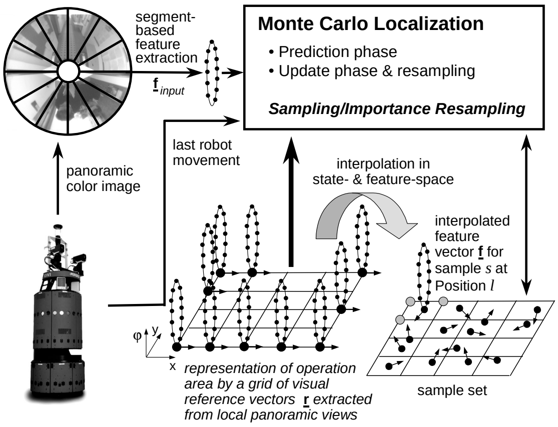
segment-based feature extraction
$\underline{\mathbf{f}}_{input}$
Monte Carlo Localization
• Prediction phase
• Update phase & resampling
Sampling/Importance Resampling
panoramic color image
last robot movement
interpolation in state- & feature-space
interpolated feature vector $\underline{\mathbf{f}}$ for sample s at Position l
φ
y
x
representation of operation area by a grid of visual reference vectors $\underline{\mathbf{r}}$ extracted from local panoramic views
sample set

Figure 10: *General idea of our view-based MCL.*

directional color camera supplies a much greater wealth of information about the structure of the local surroundings, we expect to defuse that problem and to accelerate relocalization significantly. Therefore, we currently develop an approach for vision-based self-localization that combines panoramic views of the omni-camera with the Monte Carlo Localization (MCL) developed by Fox [7].

MCL is a new algorithm for robust and efficient self-localization of mobile robots. It is a version of Markov localization [11, 6], a family of probabilistic approaches for approximating a multi-modal probability distribution coding the robot's belief $Bel(l)$ for being at position $l = (x, y, \varphi)$ in the state space of the robot. x and y are the robot's coordinates in a world-centered Cartesian reference frame, and φ is the robot's orientation. MCL applies sampling techniques to represent the posterior belief $Bel(l)$ for being at position l by a set of N weighted, random samples S. Samples in MCL are of the type $\langle\langle x, y, \varphi\rangle, p\rangle$, where $p \geq 0$ is a numerical weighting factor, analogous to a discrete probability. Because the sample set constitutes a discrete approximation of the continuous probability distribution, the MCL approach is computationally efficient, it places computation just "where needed". Additionally, it is more accurate than Markov localization with a fixed cell size, as state represented in samples is not discretized [7]. This allows a self-localization with sub-grid accuracy. In analogy with the MCL algorithm presented in [7], our view-based MCL (Fig. 10) proceeds in two phases:

Prediction phase (robot motion): In this phase, the sample set computed in the previous iteration (or during initialization) is moved according to the last motion of the robot. This way, MCL generates N new samples that approximate the predictive probability density of the robot's position after the motor command. In our approach, we use a discrete representation of the operation area by a coarse grid of visual reference vectors $\vec{r}(x, y, 0^o)$ extracted from the respective panoramic view at this position in the reference orientation $\varphi = 0$ (Fig. 10). Because of the discrete grid representation, our approach requires interpolations both in state and feature space to determine the unknown feature vectors $\vec{f}(l'))$ of the moved samples in the new positions l' within the grid. First, we interpolate linearly between the three reference vectors $\vec{r}(x, y, 0^o)$ closest to the sample position l'. After this, this new vector is rotated according to the orientation φ of the sample. This is possible, since the visual features are extracted from annulus segments of the omnidirectional image (see Fig. 10-top left). A rotation of this image corresponding to a turn of the robot simply results in a circular shift of the components of the feature vector. Since the feature vector only has a discrete number of components, for continuous rotation angles, we additionally use a linear interpolation between the features of neighbored segments. This way, we obtain a set of N interpolation-based feature vectors $\vec{f}(x, y, \varphi)$ describing the moved samples at the new positions l'.

Update phase: In this phase, the panoramic view at the new robot's position has to be taken into account in order to re-weight the sample set. For this, the weighting factor p_i of each sample s_i describing the probability that the robot is located at the position of the sample s_i is computed. We determine the similarity E_i between the current input feature vector $\vec{f}_{input}$ extracted from the panoramic view at the current robot's position and the interpolation-based feature vector $\vec{f}_i(x, y, \varphi)$ of each sample s_i simply by computing the angle between both normalized vectors. Now $p_i = 1 - \alpha E_i$ can be determined, where α is a normalization constant that enforces $\sum_{n=1}^{N} p_n = 1$. The new sample set S for the next iteration is obtained by resampling from this weighted set. The resampling selects with higher probability samples that have a high likelihood (weighting factor) associated with them. Samples with low weighting factors p_i are removed and randomly placed in the state-neighborhood of samples with high weighting factors.

Subsequently, we present first promising experimental results of this approach, also considering the specificity of the environment in the market. Fig. 11 and 12 illustrate empirical results of experiments recently executed in a section of the store. Despite the uniformity of the two hallways and the coarse grid-space of 90 cm, the view-based MCL yields accurate localization results already after a few movements of the robot. In the normal case (no occlusions), this approach allows a correct localization with sub-grid accuracy. At the end of the sequence shown in Fig. 12, four people were standing around the robot in a very low distance and occluded large regions of the panoramic image (40-50%). Despite this occlusion, the MCL still generates good localization results. In this critical situation, the difference between estimated and correct position is not larger than the grid space. The empirical experiments confirm the robustness of this vision-based localization and tracking method: the influence of lighting, changes within the operation

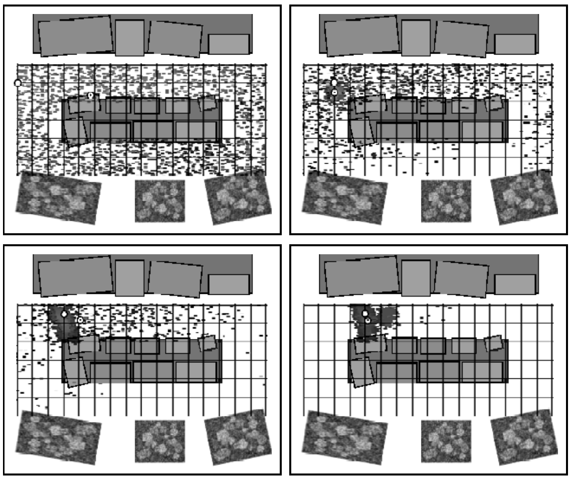

Figure 11: *View-based self-localization and tracking experiment realized in a section of the store ($6 \times 15m^2$, grid space 90 cm). Series of 2D sample sets using panoramic views as sensory input for MCL. Sequence depicts the temporal condensation dynamics of the samples - as result of local robot movements and the sampling/importance re-sampling cycle. In the beginning, the robot is globally uncertain, the samples are spread uniformly throughout the free space. Already after five movements, MCL has disambiguated the robot's position - the majority of samples is now centered tightly around the correct position.*

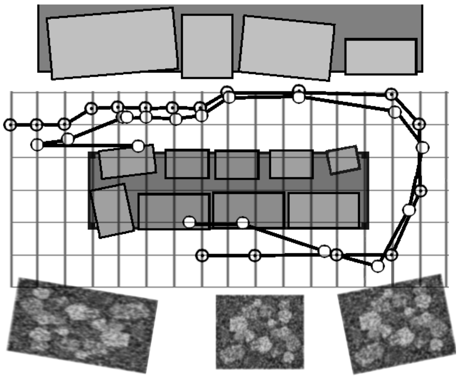

Figure 12: *Complete sequence of a self-localization and tracking experiment illustrated in Fig. 11. The correct positions of the robot are marked by dotted white circles ⊙ and the estimated positions (centers of gravity of the samples with highest weighting factors) by white circles.*

area and local occlusions caused by customers or other objects is of low significance. The view-based MCL approach presented here turned out to be a robust online, any-time algorithm which can generate an answer at any time, but the quality of solution increases over time.

4 Conclusions and Outlook

The PERSES project contains a collection of new and known approaches, addressing challenges arising from the characteristics of the scenario and the environment, and from the need to continuously interact with customers. In the paper, special emphasis has been placed on vision-based methods for user localization, map building and self-localization to accommodate the challenges that arise from the scenario. The overall system presented here should be understood as work in progress, undergoing continuous changes. Therefore, so far, we can only present preliminary results demonstrating the function of principle of selected subsystems of the overall architecture. Besides the implementation of vision-based methods allowing a robust self-localization and navigation in larger areas of the whole store ($80 \times 100m^2$), the continuous vision-based interaction between robot and user still remains a challenge to realize a user-friendly guidance and companion function as mobile shopping assistant.

Acknowledgement: The authors wish to thank the PERSES and MILVA mobile robot groups, particularly the PhD-students A. Corradini, J. Key, C. Schauer, T. Wilhelm, and the students A. Liebetrau, A. Koenig, and T. Hempel, for implementations, experiments, and discussions.

References

[1] S. Amari. Dynamics of Pattern Formation in Inhib. Type Neural Fields, *Biol. Cyb.*, 27 (1977) 77-87

[2] T. Bergener et. al. Complex behavior by means of dynamical systems for an anthropomorphic robot. *Neural Networks*, 12 (1999) 1087-1099

[3] H.-J. Boehme, U.-D. Braumann, A. Corradini, H.-M. Gross. Person Localization and Posture Recognition for Human-Robot Interaction. In *Proc. of 3rd Int. Gesture Workshop (GW'99)*, 105-116, Springer 1999

[4] W. Burgard, A.B. Cremers, D. Fox, D. Haehnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an Interactice Museum Tour-Guide Robot. *Artificial Intelligence*, 114 (1999) No. 1-2

[5] A.P. Duchon, W.H. Warren, and L.P. Kaelbling. Ecological Robotics. *Adaptive Behavior*, 6 (1998) 473-507

[6] D. Fox, W. Burgard, and S. Thrun. Markov Localization for Mobile Robots in Dynamic Environments. *Journal of Artificial Intelligence Research*, 11 (1999) 391-427

[7] D. Fox, W. Burgard, F. Dellaert, and S. Thrun. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In: *Proc. AAAI-99*

[8] H.P. Moravec. Sensor fusion in certainty grids for mobile robots. *AI Magazine*, Summer 1988, 61-74

[9] C. Schauer, Th. Zahn, P. Paschke, and H.-M. Gross. Binaural sound localization in an Artificial Neural Network. In: *Proc. IEEE-ICASSP'2000*, II 865-68, IEEE Press

[10] G. Schoener. Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16 (1995) 213-245

[11] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artif. Intell.*, 99 (1998) 21-71

[12] S. Thrun et al. MINERVA: A Second-Generation Museum Tour-Guide Robot. In: *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'99)*, 1999

[13] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time Tracking of the Human Body. *IEEE Trans. on PAMI*, 19 (1997) 780-785