Proc. SOAVE2000 - SelbstOrganisation von Adaptivem VErhalten, Ilmenau, Fortschritt-Berichte
VDI, Reihe 10: Informatik/Kommunikationstechnik, Vol. 643, pp. 57-66, VDI-Verlag 2000

PERSES - a PERsonal SERvice System*

H.-J. Boehme, H.-M. Gross, J. Key, T. Wilhelm

Ilmenau Technical University, Department of Neuroinformatics

98684 Ilmenau (Thuringia), Germany

http://cortex.informatik.tu-ilmenau.de

Abstract

This paper describes our long-term research project PERSES (PERsonal SErvice System), which deals with the mainly vision-based interaction of human users with a mobile service-robot. As application scenario we chose a home improvement store, where the robot is to operate as a mobile information kiosk or shopping assistent. Against this background, several functional necessities have to be fulfilled by the robot, beginning with the ability to navigate savely in a crowded environment, and ending with an intuitively understandable and natural interaction with its users. In many mobile robot applications, the robots perceive their surroundings mainly by means of distance measuring devices (laser, sonar). Due to the characteristics of our operation area as highly unstructured, dynamic and crowded environment, and our methodological interest in visual information processing for both human-robot interaction and navigation, we have focused on *vision-based methods* to augment the perceptual space of the robot. In this paper, we concentrate on selected, methodically interesting aspects of our approach: sonar-based map building in a large scale environment, vision-based local navigation, obstacle avoidance, and self-localization, as well as user localization and verification. For the already implemented behavioral submodules, we present preliminary results of experiments achieved with our robot platform PERSES in a home improvement store in Erfurt. The results are promising and illustrate the functionality, but also the strengths and weaknesses of the already realized subsystems for navigation and human-robot interaction.

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 and model 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:

*Supported by the Thuringian Ministry of Science, Research, and Art (Grant B 611-98041, PERSES-Project)

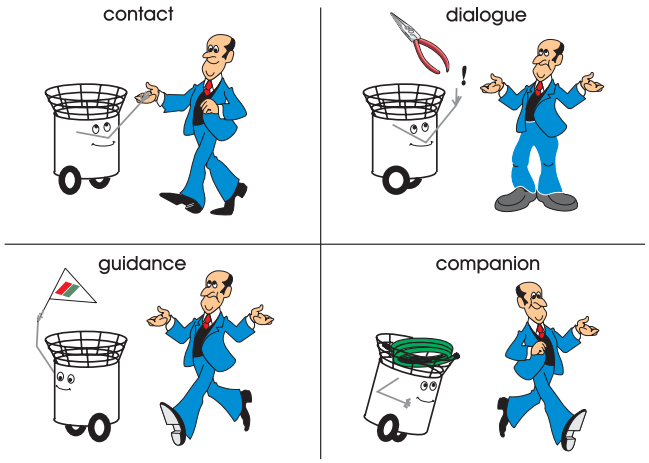


Figure 1: *Necessary skills and typical service tasks of a user-oriented, interactive mobile shopping assistant.*

(a) *building and maintaining large-scale maps as well as continuous self-localization of the robot in the operation area,* (b) *robust avoidance of static and dynamic obstacles during navigation,* (c) *navigation to desired places, articles, or market areas acting as a guide,* (d) *visual localization of a potential user within a pre-defined operation area,* (e) acoustic localization of a potential user clapping his hands or shouting a command to attract attention, (f) 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, (g) robust vision-based user tracking both while standing still and during self-movement of the robot, (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

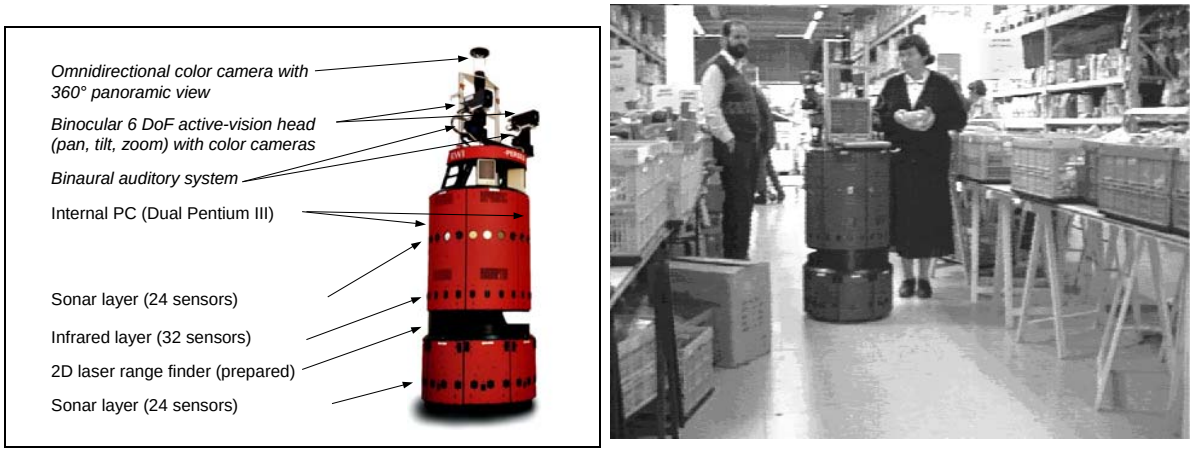


Figure 2: *Left: Experimental platform* PERSES. *In addition to the standard equipment of two sonar and one IR-layers,* PERSES *is equipped with (i) an omnidirectional color camera with a* 360^o *panoramic view used for user localization and tracking, self localization and local navigation, (ii) a binocular 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. Right:* PERSES *during a test run in a home improvement store, a cluttered and un-engineered environment with numerous critical obstacle configurations.*

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 navigation 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, left 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. In contrast, vision-based approaches do not show these limitations, but supply a much greater wealth of information about the structure of the local surroundings (see section 3.1). The robot PERSES we use as experimental platform in our experiments in the store is a standard B21 robot by RWI, additionally equipped with sensor systems for interaction and navigation (Fig. 2).

2 System Architecture

2.1 Overview

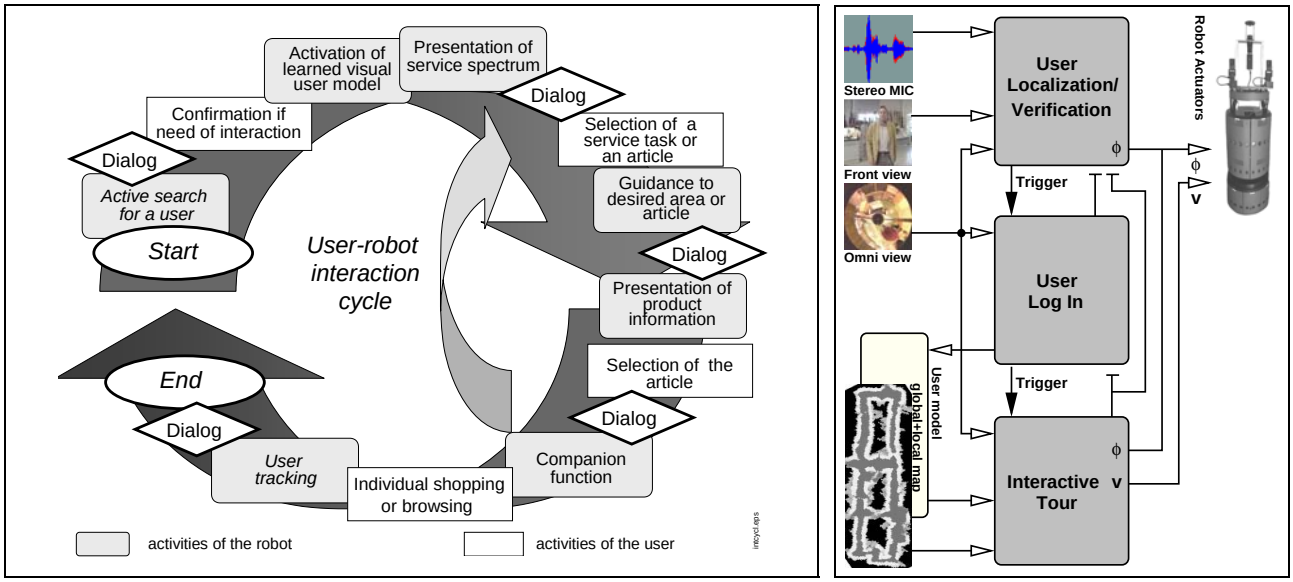


Figure 3: *Left: Schematic sketch of a typical interaction cycle between a customer and the mobile shopping assistent which determines the PERSES system architecture (right).*

Fig. 3 (left) illustrates the typical phases of an interaction process between a customer and the mobile service system. Because of the complexity of the „shopping-task“ as a whole, for the development of the system architecture, we use an approach which allows us to decompose the problem into separate behavior modules responsible for several subtasks of the interaction and navigation cycle (Fig. 3, right). As formal framework for behavior coordination, we chose the so-called dynamic approach to robotics [13]. The PERSES-architecture consists of three main subsystems: *User Localization/Verification*, *User LogIn* and *Interactive Tour* (Fig. 3).

2.2 User Localization and Verification

This subsystem is responsible for the robust localization and verification of a potential user in the surroundings. For localization, 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. While standing still, the motion-based segmentation calculates some candidate regions that indicate if and where potential users could be. 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 (see Section 3.4).

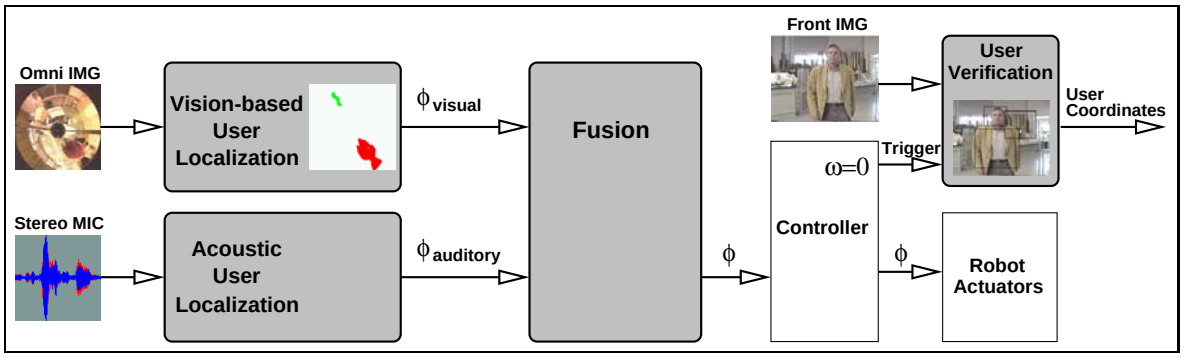


Figure 4: *Subsystem for localization of a potential user in the operation area.*

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. 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. 3-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. This evaluation is done by the *User Verification* module, which uses a multiple cue approach to confirm people possibly willing to interact with the system. 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 (see Section 3.4).

2.3 User Login

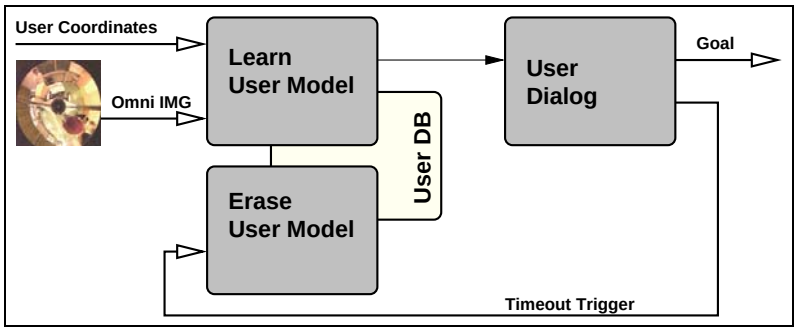


Figure 5: *Modules of the User LogIn subsystem.*

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.

2.4 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

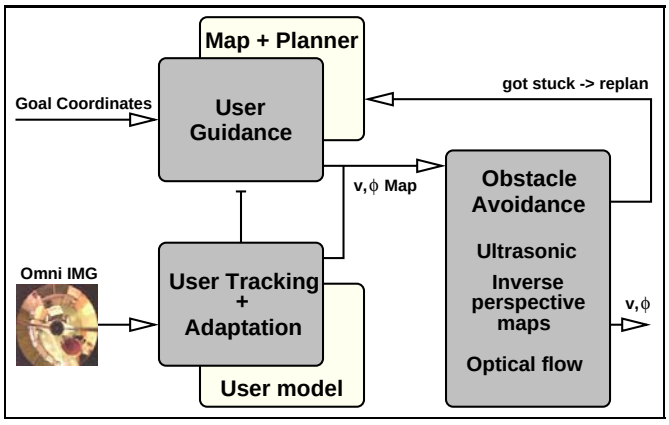
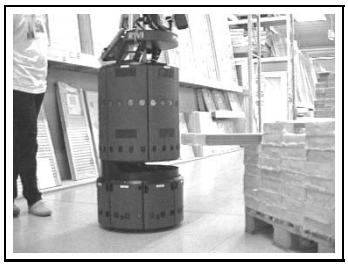


Figure 6: *Modules of the Interactive Tour subsystem.*

to the desired position. For map building, self-localization, and global navigation, we use very efficient statistical and probabilistic techniques known from literature [10, 14, 3, 6, 7]. We currently extend them to the specific visual inputs provided by the on-board cameras. 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 (see Section 3.1).

3 Description of selected subsystems

3.1 Local maps and obstacle avoidance



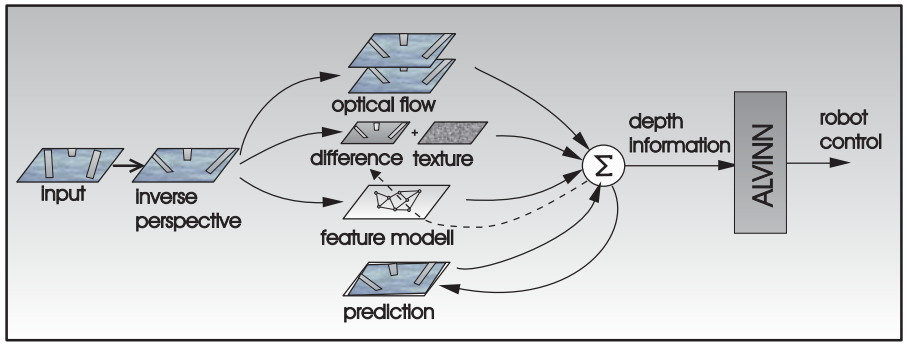


Figure 7: *Left: Typical obstacle configuration in our application scenario. Right: Multiple-cue architecture for visual obstacle avoidance: after an invers-perspective mapping, the perception layer calculates optical flow and difference images, and additionally uses an adaptive color feature model to segment unstructured image regions; the prediction layer produces an expectation attitude for future steps. The action layer uses an* ALVINN*-architecture to provide the motor output. The whole system was originally implemented for visual guidance of the robot* MILVA *(see [9] for details of this approach) and is currently transfered to the* PERSES*-platform.*

As stated above, navigation and obstacle avoidance purely based on sonar or laser sensors is

insufficient in our szenario. To emphasis this fact, fig. 7 (left) contains a typical obstacle jutting out. Unfortunately, the robot cannot perceive such an obstacle with its distance measuring sensors. Therefore, the sonar based navigation has to be augmented by vision-based methods. Fig. 7 (right) sketches an multiple-cue approach that was developed in our department ([9]). The images are captured by a frontally aligned camera tilted to the floor. Provided egomotion of the robot, this approach allows to segment the scene into free space and obstacles, even in the absence of texture. Fig. 8 shows one exemplary segmentation result.

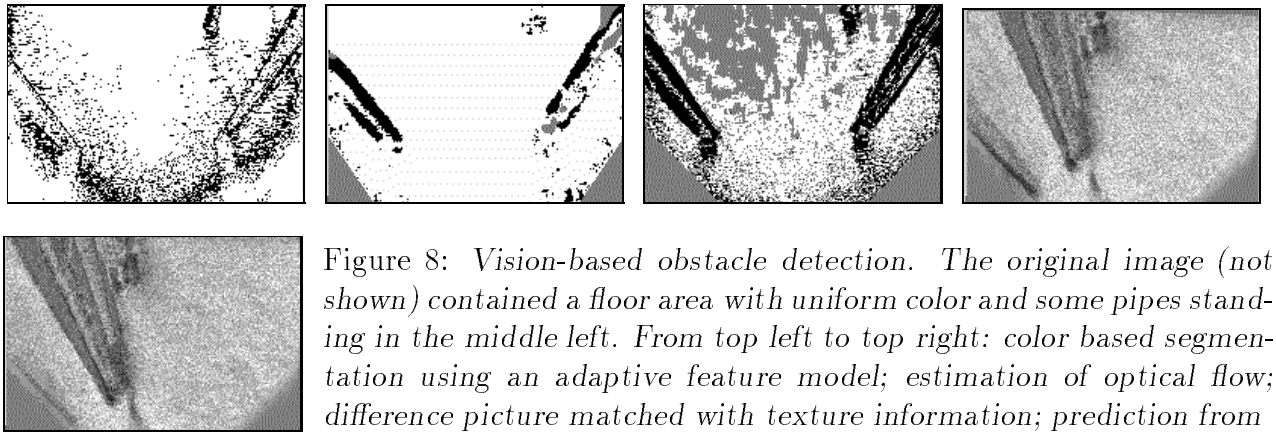


Figure 8: *Vision-based obstacle detection. The original image (not shown) contained a floor area with uniform color and some pipes standing in the middle left. From top left to top right: color based segmentation using an adaptive feature model; estimation of optical flow; difference picture matched with texture information; prediction from* $t-1$*; Bottom: final result of feature-extraction (light: free space; dark: obstacle)*

Currently, we try to combine the local map supplied by the sonar sensors with the vision-based obstacle detection in order to make the local navigation more robust and to improve the local maps used for global map building.

3.2 Building and maintaining a global map

PERSES uses an occupancy map to orient itself (see figure 10): This map is learned from sensor data (sonar scans and odometry readings) that are collected when manually joy-sticking the robot through its environment or autonomously exploring a local area in the market.
Up to now, the map building is based on ultrasound distance measures and odometry readings. One major problem using odometry data is their increasing error over time, especially concerning the rotation angle. This problem is well known and leads to the fact that a global map generated along a closed-loop course cannot be really closed without additional efforts (see fig. 10, left). To attenuate this effect, we utilize a specific feature of our market floor: the floor (ground) of our operational area shows a rectangular structure caused by tiles which are uniquely oriented across the whole market area (see Figure 9, left). The idea is quite obvious and illustrated in Figure 9: a camera acquires images of the surrounding floor, and by continuously estimating the dominant orientations within that image, we can calculate the accurate orientation of the robot and, therefore, we can substitute the orientation measure supplied by odometry with the correct orientation value. Hence, it is possible to eliminate the orientation error, and subsequently, the position error. Under the assumption that the initial position and orientation of the robot are known, the described method yields a very accurate position tracking. Consequently, the risk of loosing the actual position can be minimized. Of course, the proposed approach does not hold in a more general framework, but is very well suited for our special environment. Figure 10 illustrates the efficiency of this scenario-specific method for vision-based odometry correction. It shows the resulting occupancy maps without (left) and with (right) odometry correction. Here, a section of the market of about 60 by 20 meters (path length about 250 meters) was explored. The proposed method for visually-based odometry cor-

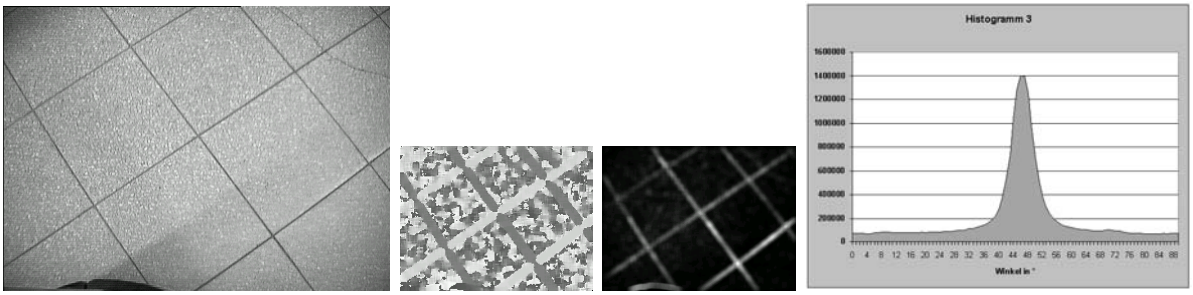


Figure 9: *General idea of our vision-based odometry correction considering a specific feature of the market floor: a) image of the floor in front of the robot, b) local orientation tensors (orientations are coded as gray values), c) confidences of local orientations (low-black, high-white), d) histogram of confidence-weighted local orientations, the dominant orientation (center of gravity) is a significant measure for the accurate orientation of the robot*

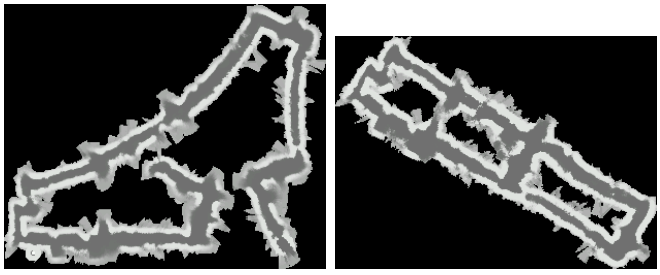


Figure 10: *Results of the occupancy map building; (Left) without vison-based correction of odometry: the closed-loop course cannot be closed, because the error of the odometry-based estimation of the rotation angle finally amounts to 90°; (Right) with vison-based correction of odometry: the closed-loop course can be closed exactly.*

rection allows to avoid the computationally expensive EM-algorithm for localization estimation and map building [4], which is extensively exploited for mobile robot applications ([15]).

3.3 Vision-based self localization

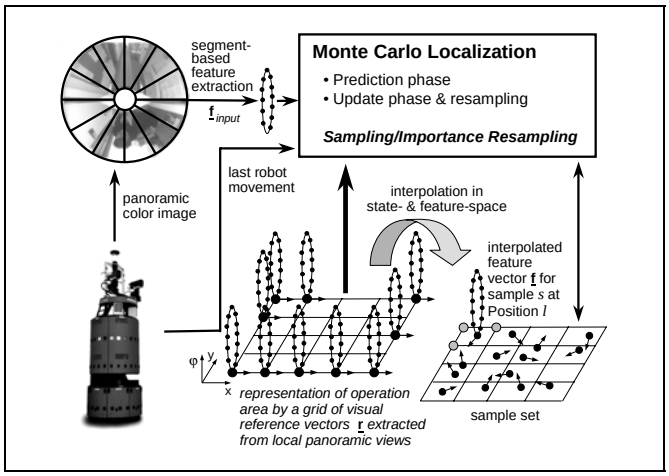


Figure 11: *General idea of our view-based Monte Carlo Localization. The approach is based on a grid-based representation of the operation area by a set of panoramic views of local surroundings.*

*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 re-*localization in case of a complete loss of positioning. Because the visual input from the omni-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 (fig. 11) that combines panoramic views of the omni-camera with the Monte Carlo Localization (MCL) developed by Fox [7].

Fig. 12 illustrates the results of ongoing 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

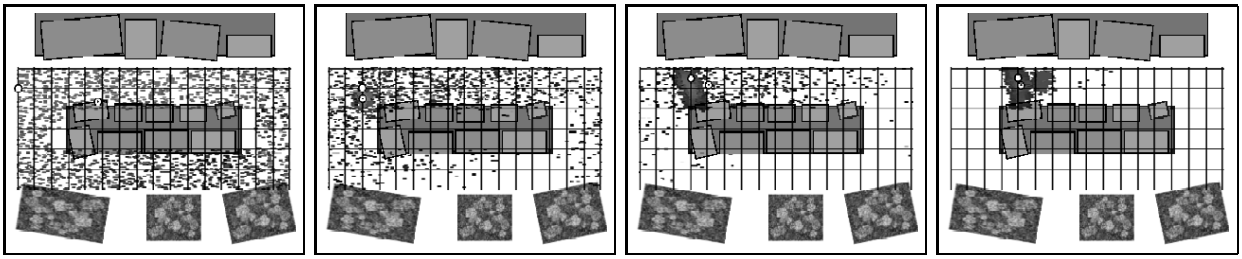


Figure 12: *View-based self-localization and tracking experiment realized in a section of the store ($6 \times 15m^2$, grid space 90 cm). Series of 2D sample sets using panoramic views as sensory input for MCL. Sequence depicts the temporal condensation dynamics of the samples - as result of local robot 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.*

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. The experiments confirm the robustness of this vision-based localization and tracking method: the influence of lighting, changes within the operation area and local occlusions caused by customers or other objects seem to be of low significance. A more detailed description of the proposed method can be found in [8].

3.4 Localization and verification of a potential customer

One of the major problems of our scenario consists in the robust localization of a potential user in the operation area. At present, we use a multimodal approach that integrates both visual and acoustic stimuli into the localization process, and both cues are combined into one behavioral module. The *Vision-based User Localization* performs a motion-based foreground-background segmentation in the input images provided by the omnidirectional camera, and returns the angle to the center of gravity of the largest moving region. In the waiting position or while standing still, the motion-based segmentation gives us some candidate regions that indicate if and where persons could be in the surroundings of the robot (Figure 13). Our implemented

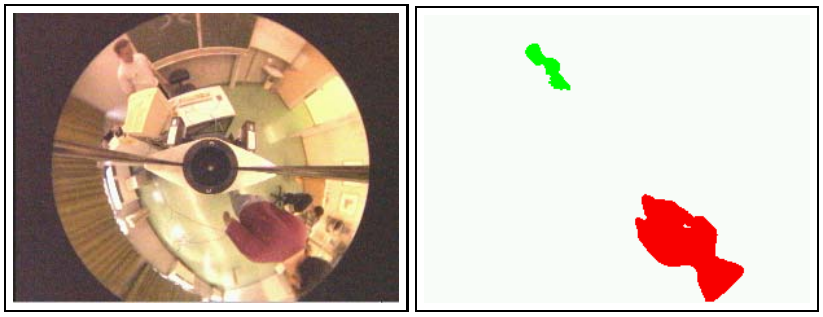


Figure 13: *Motion-based segmentation of potential users in the image sequence of the omnidirectional camera. (Left) original image. (Right) segmented image, the two regions correspond with two persons at different distances to the robot.*

method is similar to that suggested in the *Pfinder* system [16], but differs in the following aspects: (i) the statistical models for foreground and background pixels were simplified to boxes, and (ii) the foreground and background models are continuously adjusted. The model simplification led to a lower computational complexity resulting in a performance speed-up, surprisingly without almost no lost in sensitivity. By the adaptation of the models, we take into account that the robot cruises its surroundings which makes it impossible to use only one stationary background model. After the alignment of all image pixels to the foreground and

background model, respectively, a simple grouping mechanism is applied to get closed moving regions and to suppress noise and very little regions. The segmented candidate regions are labeled according to their distance to the robot.
For the acoustic localization of a potential user clapping his hands or shouting a command, we developed a biologically inspired model of binaural sound localization using interaural time differences and spikes as temporal coding principle [11]. This subsystem realizes (i) the detection of the sound direction in the horizontal half-planes by processing of the interaural time-delays 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. 2). It detects pitch onsets in the signals and calculates the angle to the sound source from the phase shift between the binaural signals. Details of this model and localization results are presented in [12]. The integration of auditory saliency makes it easy for the user to attract the attention of the robot to accelerate the localization process significantly.

Figure 14: *Typical verification results for different situations in the home improvement store. 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. The left image shows a back light scene taken in the entrence area, where only the contour detection can provide a confident contribution for verification. 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.*

The verification of the localization hypothesis is realized by the *User Verification* (see Figure 14, and [2] for a more detailed description). Its execution is triggered, when the robot was turned by the localization module and the controller reached its final position. Due to the body movement of the robot, the potential customer 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. To realize a robust verification of a customer hypothesis, we use a task-specific multi-cue saliency system that integrates different visual cues. A multiresolution pyramid transforms the images acquired by one of the frontally aligned cameras into a multiscale representation. Because we want to localize persons even at different distances from the robot, we use five fine-to-coarse resolutions in our scale space. Two cue modules sensitive to *facial structure* and *structure of a head-shoulder contour*, respectively, operate at all levels of the grayscale pyramid. The cue module for *skin color* detection uses the original color image. After superposition of the corresponding feature maps, a 3D-Winner-Take-All process [1] within the saliency pyramid selects that region most likely covering the upper part of a person. This way, the system becomes more robust, can handle varying environmental conditions and is less dependent on the presence of any specific feature. Hence, we can handle varying environmental circumstances much easier, which, for instance, can make the skin color detection difficult or almost impossible. Fig. 14 shows typical verification results obtained in the home improvement store.

4 Conclusion 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. The experimental results are promising and illustrate the functionality, but also the still existing weaknesses of the already realized subsystems for human-robot interaction and navigation. Our overall system must be understood as work in progress, undergoing continuous changes. Therefore, so far, we can only present preliminary results demonstrating the operation of selected subsystems. Future research will concentrate on robust person tracking algorithms needed to keep continuously contact with the actual user, and on the design and implementation of a dialog component integrating speech recognition and verbal articulation of the robot itself. Furthermore, research issues will be dedicated to the recognition of gestures (body language) for human-robot interaction, and the flexible integration of all subsystems in our control architecture. Besides the implementation of robust vision-based methods for localization, tracking, and navigation, the continuous interaction between robot and user still remains a challenge.

References

[1] Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] Boehme, H.-J., Braumann, U.-D., Corradini, A., and Gross, H.-M. Person Localization & Posture Recognition for Human-Robot Interaction. In *GW'99 - The 3rd Gesture Workshop, Gif-sur-Yvette, France*, pages 105–116. Springer, Lecture Notes in Artificial Intelligence 1739, 1999.

[3] Burgard, W., Cremers, A., Fox, D., Lakemeyer, G., Hähnel, D., Schulz, D., Steiner, W., and Thrun, S. The interactive museum tour-guide robot. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[4] Dempster, A., Laird, A., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[5] Duchon, A.P., Warren, W.H., and Kaelbling, L.P. Ecological Robotics. *Adaptive Behavior*, pages 473–507, 1998.

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

[7] Fox, D., Burgard, W., and Thrun, S. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In *Proceedings 16th National Conference on Artificial Intelligence (AAAI-99)*, 1999.

[8] König, A., Key, J., and Gross, H.-M. Visuell basierte Monte-Carlo Lokalisation für mobile Roboter mit omnidirektionalen Kameras. In *Proceedings SOAVE'2000*. VDI Verlag, 2000.

[9] Krabbes, M., Weber, S., Boehme, H.-J., and Gross, H.-M. Monokulare visuelle Hindernisdetektion auf Basis merkmalsbasierter Bildsegmentierung. In *AMS'98 - 14. Fachgespräch Autonome Mobile Systeme*, Informatik aktuell, pages 85–92. Springer-Verlag, 1998.

[10] Moravec, H.P. Sensor fusion in certainty grids for mobile robots. *AI Magazine*, pages 61–74, 1988.

[11] Paschke, P. and Schauer, C. A spike-based model of binaural sound localization. *International Journal of Neural Systems*, 9(5):447–452, 1999.

[12] Schauer, C., Zahn, T., Paschke, P., and Gross, H.-M. Binaural sound localization in an Artificial Neural Network. In *Proceedings IEEE-ICASSP'2000*, volume II, pages 865–868. IEEE Press, 2000.

[13] Schöner, G., Dose, M., and Engels, C. Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16:213–245, 1995.

[14] Thrun, S. Learning Maps for Indoor Mobile Robot Navigation. *Artificial Intelligence*, 99(1):21–71, 1999.

[15] Thrun, S., Burgard, W., and Fox, D. A Probabilistic Approach to Concurrent Mapping and Localization for Mobile Robots. *Machine Learning and Autonomous Robots (joint issue)*, 31(5):1–25, 1998.

[16] Wren, C., Azarbayejani, A.and Darrell, T., and Pentland, A. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997. M.I.T. Media Lab Techreport TR 353.