Künstliche Intelligenz (KI), vol. 4 (2000), pp. 34-36

The PERSES Project – a Vision-based Interactive Mobile Shopping Assistant

Horst-Michael Gross, Hans-Joachim Böhme, Jürgen Key, Torsten Wilhelm

This paper describes the general idea, the application scenario, the system architecture and selected methodological approaches of our long-term research project PERSES started in the middle of 1999. This project deals with the mainly vision-based interaction of a human user with a mobile service-robot operating as an interactive shopping assistant in a home improvement store. Due to the specificity of the scenario, the characteristics of the operation area as highly unstructured, dynamic and crowded environment, and our methodological interest in visual information processing, we have focused on vision-based methods for both human-robot interaction and robot navigation.

1 Project Idea and Scenario

The aim of the project PERSES (PERsonal SErvice System) consists in the development of an interactive mobile shopping assistant that allows a continuous and intuitively understandable interaction with a human user. 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 services in the context of the present situation in the interaction cycle. Typical service tasks we want to tackle are to guide the user to desired market areas or articles or to follow him as a mobile information kiosk while continuously observing him and his behavior. In the chosen scenario, a mainly vision-based dialogue seems to be the most natural way for a robust and continuous interaction between customer and robot. But, most of the known approaches for visually guided human-machine interaction require certain constraints concerning the environmental conditions. However, during interaction with a mobile service robot operating in an unconstrained indoor area, one cannot assume predefined circumstances. Therefore, such a system must be able to handle highly varying environmental conditions which can neither be estimated nor influenced in advance.

In the context of our scenario, we have to cope with the following interaction and navigation tasks: 1) visual localization of a potential user within a pre-defined operation area, 2) acoustic localization of a potential user clapping his hands or shouting a command, 3) fast learning of an initial visual model of the current user and online adaptation of that model handling the stability-plasticity dilemma due to the varying appearance of the user in the course of the shopping process, 4) robust vision-based user tracking, 5) robust avoidance of static and dynamic obstacles during navigation, 6) navigation to desired places, articles, or market areas acting as a guide, 7) continuous self-localization of the robot in the operation area, 8) recognition of simple spoken commands, and, for the future, 9) recognition of gesticulated user instructions and body language. This spectrum of tasks necessitates adaptive methods at all processing levels using neural networks for visual and acoustic scene analysis, probabilistic methods for map building, self-localization and navigation (Moravec 1988, Thrun 1998, Burgard 1999, Fox 1999a,b), and concepts from Machine Learning and Control Theory for dynamic coordination of the subsystems responsible for the several interaction and navigation tasks (Schoener 1995, Bergener 1999). Due to the specificity of this interaction-oriented scenario and the characteristics of the operation area as highly unstructured and dynamic environment with atypical obstacle configurations (e.g., boards and pipes jutting out of shelves), we have focused on vision-based methods for both the interaction and the navigation process.

2 System Architecture

Because of the enormous complexity of the "shopping-task" as a whole, we use an approach which allows us to decompose the problem into separate behavior modules or subsystems 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 (Schoener 1995). The PERSES-architecture consists of three main subsystems or "meta-behaviors": User Localization, User LogIn and Interactive Tour (see Figure 1). The subsystem User Localization is responsible for extracting the position of a potential user in the surroundings. It supplies the resulting coordinates to the User LogIn subsystem, which learns a vision-based user model used for user tracking throughout the interactive shopping tour. Now it is the turn of the user to ask the robot for an article he is looking for. Because this part is not of direct research interest to us, we prefer a simple solution for the dialog, e.g., a touch screen based user interface. As a result of the dialog, the meta-behavior Interactive Tour is activated. Now the robot shall guide the customer through the market, using an omnidirectional camera to track the user, stopping the tour in case the user stops or begins an individual shopping tour.

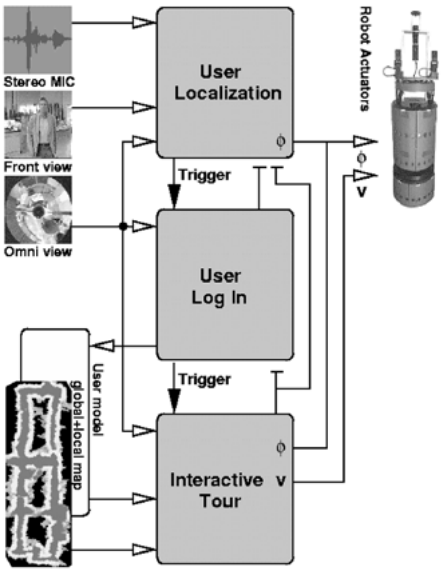


Figure 1: System architecture of our Personal Service System PERSES.

Künstliche Intelligenz (KI), vol. 4 (2000), pp. 34-36

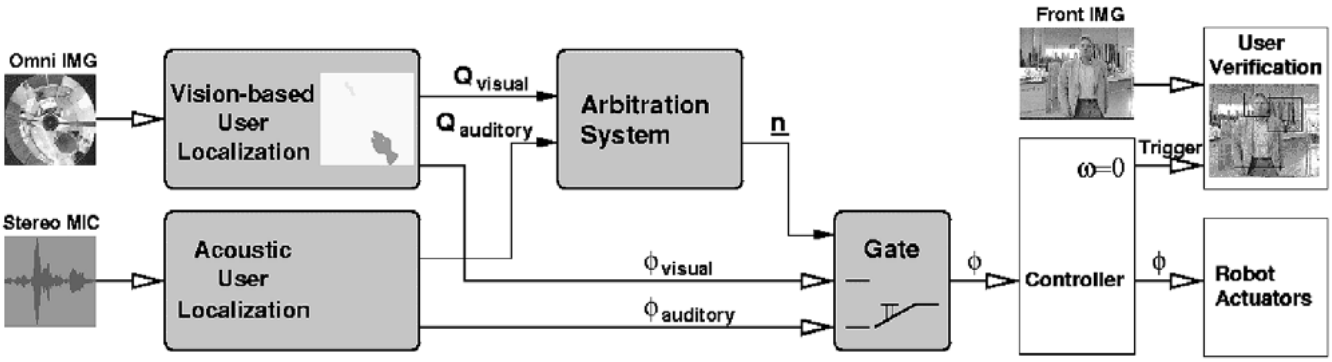


Figure 2: Subsystem for user localization.

User Localization: One of the major problems of our scenario consists in the robust localization of a potential user in the market. At present, we use a multimodal approach that integrates both visual and acoustic stimuli (see Figure 2). The module Vision-based User Localization performs a motion-based foreground-background segmentation in the image sequence provided by the omnidirectional camera, and returns the angle to the center of gravity of the largest moving region. 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. 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 (Schauer 1999). According to this model, the module Acoustic User Localization listens to a stereo sound stream supplied by the microphones mounted on top of our robot. It detects pitch onsets in the signals and calculates the angle to the sound source from the phase shift between the binaural signals. 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 significantly. The modules both for Vision-based user localization and for Acoustic user localization make use of the same actuator, namely, they try to turn the robot towards the detected potential user. If both modules generate alternative hypotheses, or if the Acoustic User Localization detects a new sound source while the visual module is turning the robot to a potential user, the contradictory motor commands have to be coordinated properly. This competitive access to the actuators is controlled by the Arbitration System.

The verification of a localization hypothesis is realized by the User Verification module. Its execution is triggered, when the robot was turned by one of the localization modules 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. As a very simple approach, 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 saliency system that integrates different visual cues: skin color, head-shoulder contour, and facial structure. This way, the system becomes much more robust, can handle highly varying environmental conditions and is less dependent on the presence of any specific feature. Details of our multi-cue approach are presented in (Boehme 1999, Corradini 1999).

User Login: When a potential user has been localized visually or acoustically, and the user has confirmed his willingness for interaction, a vision-based model has to be learned, which can be used in the course of the interactive tour to track the current user, to keep the distance between user and robot constant, to recognize the right customer, if he was lost from view, etc. Although this part has not been realized yet, there exist a number of conceptional ideas that will be investigated in the near future.

Interactive Tour: This module is activated when the User LogIn subsystem provides the position of a desired area or article in the market, or, in case of a logout of the user, activates the homing position. In both cases 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 (Moravec 1988, Thrun,1998, Fox 1999a,b). We currently extend them to the specific visual input provided by the on-board cameras. PERSES uses two types of maps to self-localize and navigate: occupancy maps and a grid of omnidirectional views of the local surroundings in the market. Both maps are learned from sensor data (US-scans, camera images, and odometry readings) that are collected when manually joy-sticking the robot through its environment or autonomously exploring a local area in the market. In case a user is present, the internal User Tracking module is active, too. This module's goal is to keep the user within the omnidirectional view and, moreover, keep a constant distance to him. When the user falls behind or moves in another direction, this module takes over execution by inhibiting the User Guidance module and follows 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 the next movement. Before execution, they are passed to the Obstacle Avoidance module, which suppresses motor commands that are impossible according to the actual obstacle configuration of customers and market-specific obstacles (shelves, pallets, shopping carts, special offers, etc.). The need for vision-based methods for obstacle avoidance arises from the circumstances in our scenario that numerous obstacles cannot be perceived reliably by ultrasonic or laser sensors because of their specific form, size or height (e.g., shopping carts, boards or pipes jutting out of shelves). To handle this problem, besides visual sonar (Horswill 1994), local navigation methods based on optical flow and inverse perspective mappings are currently investigated in our lab.

At present, the map building in PERSES is based on ultrasonic distance measures and odometry readings. 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 our market: the floor shows a rectangular structure caused by tiles which are uniquely oriented across the whole market area. The idea is quite obvious: another 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, 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. Of course, the proposed approach does not hold in a more general framework, but is very well suited for our special environment. Figure 3 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.

Most of the known approaches use laser or ultrasonic scan-based occupancy grids for navigation. Already in the context of the MINERVA tour-guide project (Thrun 1999), Dellaert (1999) presented a vision-based approach for self-localization and navigation using a camera pointed at the ceiling. Such ceiling mosaics are more difficult to generate than occupancy maps, because the height of the ceiling is unknown or might differ from point to point, which makes it difficult to translate coordinates in the image plane into real-world coordinates. In our experimental field, we would have to handle more than six different heights of ceiling in the range between 2.50 to 7 meters. Therefore, we currently devise an alternative approach for a vision-based map building and self-localization that combines the panoramic view of the omnidirectional color camera with the Monte Carlo localization method presented by Fox (1999b). First promising experimental results of this approach can be found in (Koenig 2000). They demonstrate the accuracy and robustness of this vision-based localization method despite crowded surroundings and significant modifications within the operation area.

Figure 3: Results of the occupancy map building; Without (left) and with (right) vison-based correction of odometry.

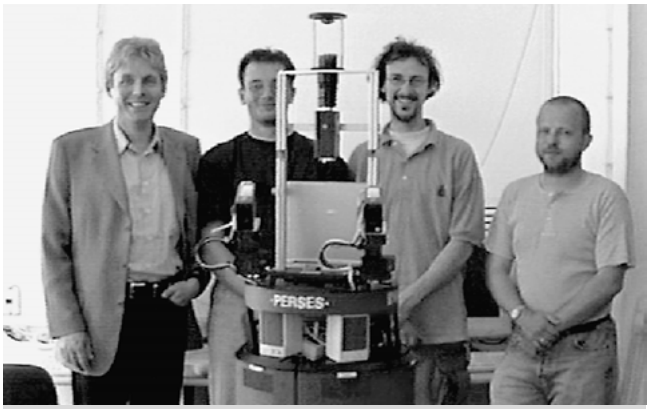
3 Conclusions and Outlook

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. The 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 issues include the visual human-robot interaction, especially the recognition of gestures (Corradini 1999) and body language, and the flexible integration of all subsystems in our control architecture. Besides the implementation of robust vision-based localization and navigation methods, the continuous interaction between robot and user still remains a challenge.

References

T. Bergener et al. (1999). Complex behavior by means of dynamical systems for an anthropomorphic robot. Neural Networks, 12, 1087-1099.
H.-J. Boehme et al. (1999). Person Localization and Posture Recognition for Human-Robot Interaction. In: Proc. of 3rd Int. Gesture Workshop, 105-16.
W. Burgard et al. (1999). Experiences with an Interactice Museum Tour-Guide Robot. Artificial Intelligence, 114 (1-2).
A. Corradini et al. (1999). A Hybrid Stochastic-Connectionist Architecture for Gesture Recognition. In: Proc. of IEEE Int. Conf. on Information, Intelligence, and Systems, 336-341.
F. Dellaert et al. (1999). Using the Condensation Algorithm for Robust, Vision-based Mobile Robot Localization. In: Proc. IEEE Int. Conf. on Comp. Vision and Pattern Recognition.
D. Fox et al. (1999a). Markov Localization for Mobile Robots in Dynamic Environments. Journ. of Artificial Intelligence Research, 11, 391-427.
D. Fox et al. (1999b). Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In: Proc. 16th Nat. Conf. on Artificial Intelligence (AAAI).
I. Horswill (1994). Visual Collision Avoidance by Segmentation. In: Proc. of IROS'94, 902-909.
A. Koenig et al. (2000). Visuell-basierte Monte Carlo Lokalisation für Roboter mit omnidirektionalen Kameras. In: Proc. SOAVE2000, 32-38, VDI.
H.P. Moravec (1988). Sensor fusion in certainty grids for mobile robots. AI Magazine, Summer, 61-74.
C. Schauer, P. Paschke (1999). A spike-based model of binaural sound localization. International Journal of Neural Systems, 9, 447-452.
G. Schoener et al. (1995). Dynamics of behavior: theory and applications for autonomous robots. Robotics and Autonomous Systems, 16, 213-245.
S. Thrun (1998). Learning metric-topological maps for indoor mobile robot navigation. Artificial Intelligence, 99, 21-71.
S. Thrun et al. (1999). MINERVA: A Second-Generation Museum Tour-Guide Robot. In: Proc. IEEE Int. Conf. on Robotics and Automation (ICRA).

Contact
Technische Universität Ilmenau, Fachgebiet Neuroinformatik, 98684 Ilmenau, homi@informatik.tu-ilmenau.de



Authors from left to right:

Horst-Michael Gross is Professor of Neuroinformatics at the TU Ilmenau. He received his Diploma degree in EE and his Doctorate degree in Neuroinformatics in 1985 and 1989, resp. Research interests: neural computing, autonomous robots, reinforcement learning, vision.

Jürgen Key received his Diploma degree in CS from the TU Ilmenau in 1998. Currently he works as research collaborator in the PERSES project at this university. Research interests: image processing, robotics, neural networks.

Torsten Wilhelm obtained the Diploma degree in CS from the TU Ilmenau in 1999. Since 1999, he has been working as academic collaborator at the Dept. of Neuroinformatics, TU Ilmenau. Research interests: multi-agent systems, learning.

Hans-Joachim Böhme received his Diploma degree in EE and his Doctorate degree in Neuroinformatics from the TU Ilmenau in 1989 and 1991, respectively. Currently he is an academic assistant at TU Ilmenau. Research interests: intelligent human-machine interaction, gesture recognition, robotics.