The Robotic Bar – An Integrated Demonstration of a Robotic Assistant

G. v. Wichert[1], C. Klimowicz[1], W. Neubauer[1], Th. Wösch[1], G. Lawitzky[1], R. Caspari[1], H.-J. Heger[1], P. Witschel[1], U. Handmann[2], and M. Rinne[2]

[1] Siemens AG, Corporate Technology, Information and Communications, 81730 Munich, Germany, georg.wichert@siemens.com
[2] Viisage Technology AG, Universitätstrasse 160, D-44801 Bochum, Germany

Summary. Coming out of the labs, the first robots are currently appearing on the consumer market. Initially they target rather simple application scenarios ranging from entertainment to home convenience. However, one can expect, that they will capture more complex areas soon. These robots will have a higher and higher level and a broad range of functional competence, and will collaborate and interactively communicate with their human users. All this requires considerable cognitive abilities on the robot's side and appropriate man-machine interaction technologies. Apart from further development of individual functions and technologies it is crucial to build and evaluate fully integrated systems. This paper describes our approach to construct a robotic assistance system. We present experience with an integrated technology demonstration and the exposure of the integrated system to the public.

1 Introduction

Coming out of the labs, the first robots are currently appearing on the consumer market. Initially they target rather simple application scenarios ranging from entertainment (e.g. Sony's Aibo) to home convenience (e.g. vacuum cleaning). However, one can expect, that they will capture more complex areas like elderly or health care as well as commercial services not too far from now.

In all these applications the robots have to be able to do useful things, which certainly requires considerable cognitive abilities (perception, situation awareness, reactivity, and plausible behaviour) and adequate robotic (manipulation) skills. Basic ingredients are: Object recognition, collision-free, purposeful arm and platform motions, grasping and manipulation, and tactile interaction.

Co-existence of robots and humans in shared workspaces is one of the major characteristics of service robot applications. Furthermore, robots and humans will not only co-exist, but collaborate and interactively communicate. So technologies like people detection, face recognition, speech recognition and understanding and dialog control will play a key role. Communication will not only be a means to command the robot, but also to exchange information, the human might as well ask the

robot for certain things as the robot might ask the humans for information it cannot acquire on its own. The latter is extremely important, since today's and certainly also tomorrow's robots will still be of very limited intellectual competence.

Customer acceptance of service robots in everyday applications will depend very much upon the whole setup's plausibility. The robot should behave "naturally", so that the human interaction partner can predict the robot's actions. It should become active on its own, if suggested by its assigned tasks. Performance and task execution time play an important role, and the need for human interventions should be limited as much as possible.

Ultimately the prospective users will judge the integrated system and not the individual algorithm or subsystem. Thus, at the current state of the technology, where satisfactory solutions are available for a number of subtasks, system integration and evaluation of the integrated system become increasingly important.

This paper describes our approach to construct a robotic assistance system – architectural principles, basic functions and higher-level skills. We present experience with integrated technology demonstrations and the exposure of the integrated system to the public first at the 2002 Hannover Messe, the world's major industrial trade show, and later e g. at a Man-Machine-Interaction conference held in Berlin 2003.

2 The MobMan Robot – System Architecture and Basic Skills

The robot *MobMan* (see fig. 1) used throughout this paper consists of a mobile base, an 8-DoF anthropomorphic arm and a 2-DoF head. It uses a 2D laser scanner for navigation, a (double) stereo vision system (mounted on the robot's pan/tilt-head) for object recognition, and several gripper mounted tactile and force sensors to support grasping operations plus a camera and a structured light depth sensor also mounted at the end-effector.

Fig. 1. The MobMan robot at the Robotic Bar during the 2002 Hannover Messe (left) and in Berlin (right).

In addition, the robot arm and parts of the mobile base are covered with an artificial skin that enables the robot to sense force and location of environmental contacts, which either result from collisions with the environment (e.g. furniture) or from the user touching the robot either by purpose (to push it away, to teach a motion,...) or unintentionally. Finally the system employs speech recognition and voice output to communicate with the user.

2.1 System Architecture

Complex robot systems acting in realistic environments require control systems, which are at the same time powerful, robust and versatile in the sense that they are able to control the large variety of different tasks, which will be typical for service robots, that operate in every-day environments. The high environmental complexity of service robotics scenarios prohibits the construction and maintenance of complete and detailed models. In recent years reactive control schemes embedded in multi-layer architectures have proven to be appropriate for such systems [5]. Reactivity is important in particular for robots working close to people for safety reasons and in order to react properly on often unpredictable human behavior. The reactive capabilities implemented on the lower level of such multi-layer control systems are commonly referred to as skills.

We assume, that most complex real world tasks can be decomposed into a sequence of elementary subtasks. This sequence can either be planned or it can be determined on-line by comparatively simple reactive behaviours. Since planning needs a complete environment model, which is hard to maintain in a dynamic and complex environment, it does not make sense to use motion and task planning on a very *detailed* level. However, in many cases it is possible, to supply the system with coarse models of task and environment, which can be used to select and parameterize appropriate predefined behaviours to be executed at the Sequencing Layer of the control system (see section 2.1). These behaviours provide a breakdown of the complex overall task into a (not necessarily linear) sequence of elementary subtasks. In turn these subtasks are executed by the Sensorimotor Skill Control (see section 2.1). At this system level necessary robustness requirements and the lack of precise a-priori knowledge demand closed loop control using concurrent sensor readings[10]. Fig. 2 gives an overview of our control system architecture.

Sequencing layer

At the time being behaviours are selected and parameterized directly via a natural speech interface. Another possibility we use are idle tasks which lets the robot get proactive. The selected behaviour is loaded from the database and scheduled onto the sequencer. Behaviours can call other behaviours, sensorimotor skills or perception modules. All of them give feedback on erroneous or successful execution. Using the object oriented scripting language *Ruby* for implementation, all types of conditional constructs like, if, for, while, case, etc. can be used to design the control flow. Additionally there are constructs to implement parallel processes or finite state machines.

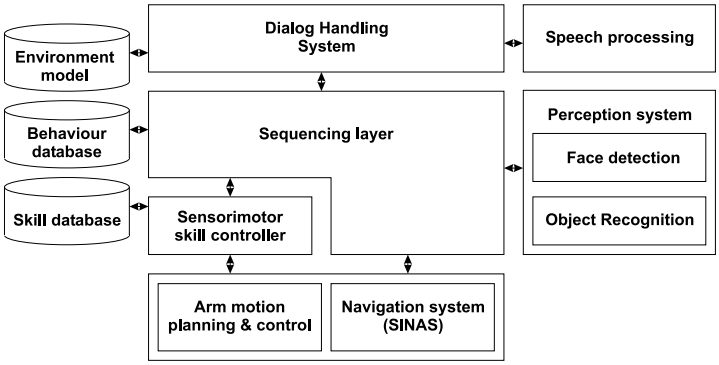

Fig. 2. The robot control system.

To design robust and mighty behaviours, it is necessary to generate and handle *all possible* feedback values.

Sensorimotor Skill Control

The elementary tasks called by the sequencing layer are referred to as 'sensorimotor skills'. They should deal with all details that are not covered by the behaviours of the control layer above. A skill requires knowledge on *what information to extract* from the sensor data and *how to extract* it, i.e. which sensor features are relevant to the task at hand. In addition it needs to know the *desired configuration* for the relevant features and *what to do* if the current feature values do not match the desired configuration (feature based servoing). Of course, the configuration of sensors, controllers, and actuators will vary during the task execution. The following basic components have been identified as building blocks for the definition of arbitrary robotic skills:

- **Feature:** Features represent the sensor data. The value of a Feature can be determined using real as well as virtual sensors.
- **Controller and Action:** Controllers evaluate Features and determine system Actions to control the features towards their "desired configurations". The Controller may use arbitrary algorithms to calculate the appropriate Actions. This includes, but is not limited to classical control engineering methods. In particular controllers can also be used to maintain the informational status of the system, triggering "cognitive" processes if needed.
- **Task Phase:** The concept of a Task Phase is used to maintain the changing arrangement of Controllers. During each Task Phase a predetermined set of Controllers is used to steer the system towards an intermediate target state. The Task Phase is completed, when the intermediate task state is reached, i.e. all controlled features have reached their desired configurations. Additionally the Task Phases can be left due to varying error conditions defined by the skill designer.

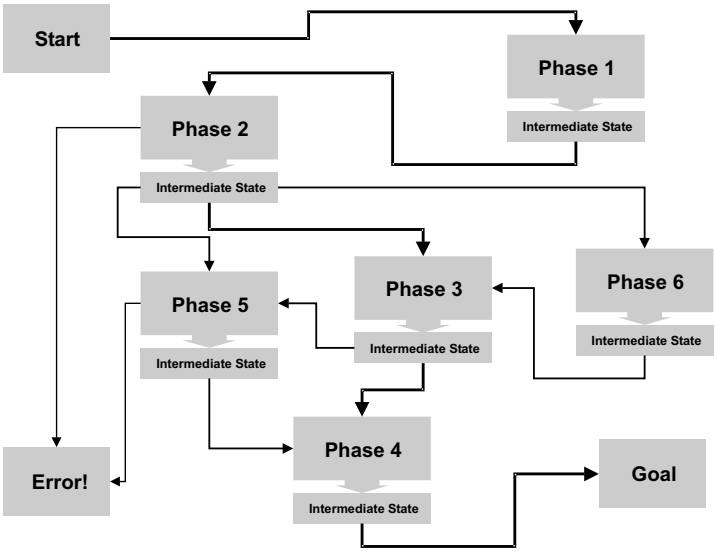

Fig. 3. The skill model as a network of Task Phases. The nominal path through phases 1,2,3 and 4 is emphasized using the thicker arrows.

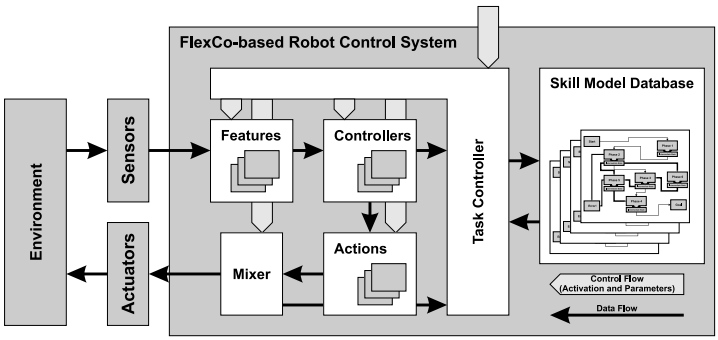

Fig. 4. The structure of a the low-level skill control system.

Task Phases can be freely arranged in a network that resembles the complete skill (see skill model in fig. 3). During each Task Phase several controllers may be active.

These building blocks are the basis for defining control components that are used to implement the actual skills. The scheduling and execution of the skill is performed by a software component, the **Task Controller**. It allows the implementation of branches and loops in the skill model. The overall structure of the low-level skill control system is depicted in fig. 4.

2.2 Basic Robot Functions

Using the control system of section 2.1 we have implemented a set of basic robot functions which are detailed in the follow sections.

Navigation

The robot is able to navigate within indoor environments. We use the SINAS navigation system [6], commercially available from Siemens.

Object Recognition and Localization

The robot's main sensor for perception of the environmental state is a stereo vision system. The perception process generates lists of classified objects with their respective position and dimensions. This information is used both for grasping and manipulation.

The raw camera images are passed to the stereoscopic scene analysis with range-data output. The classification step follows an hierarchical approach and consists of the steps: 3D-segmentation, (geometric) reconstruction and pose estimation (see fig. 5). The same image data along with the hypotheses of the geometric scene interpretation is passed to a second appearance based classifier which evaluates the color-information and performs a probabilistic fusion of object appearance and geometric hypothesis.

Scene reconstruction using a stereo vision is a popular research topic. A recent overview of stereo correspondence algorithms includes [7]. For depth image computation we use SRI's Small Vision System, which is a commercial solution for stereo analysis [3, 4]. The software provides dense range data using two video-cameras. All passive stereo systems provide the depth information only for image regions with sufficient structure. Therefore, the resulting 3D-reconstruction exists only for some parts of the image. Thus, only parts of the relevant objects can be properly reconstructed.

For the segmentation of range data various approaches have been proposed [2]. In our implementation the segmentation of the range data is based on a split and merge method. The split step separates the data into connected components. An additional splitting is performed at range data discontinuities. The merge-step is a model-based [1] approach. Currently simple geometric features for all segments resulting from the splitting steps are computed. These features are compared to previously acquired geometric descriptions for the known objects. As this step is based only on geometric information, this part of the classification procedure is robust to changing lighting conditions.

The hypotheses resulting from the analysis of the range image are passed to a second probabilistic appearance-based classification step using color features. It has the advantage, that the generation of the visual models can be reduced to the estimation of feature histograms from a set of training images. Our solution generates color-features and performs a probabilistic classification similar to [8]. The region

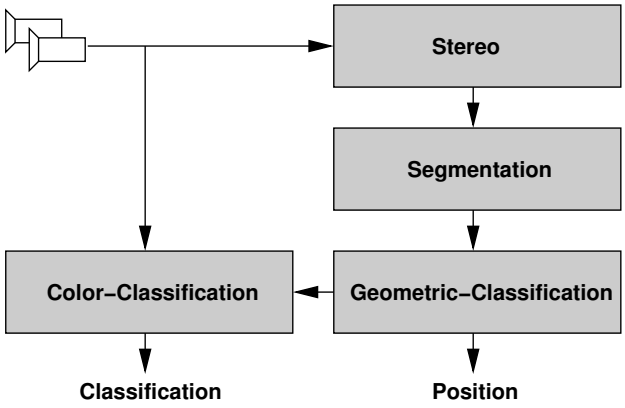

Fig. 5. The structure of the vision system used for object recognition and localization.

of interest for the feature computation is limited to the image segments which are previously created by the geometric-segmentation of the range image.

For the 3D localization of the classified objects again range information for the corresponding image segments is evaluated. An example of the classification and pose estimation result can be seen in fig. 6.

Manipulation Skills

The robot's manipulation skills are implemented in the framework of section 2.1. Up to now we have implemented several elementary skills (e.g. for door opening [9]), in the context of the bar scenario described below motion planning and grasping skills are of interest. Arm motions are executed using a hybrid motion planning and execution system providing trajectory generation and collision avoidance functionality as well as arm-base coordination and safety reflexes based on an artificial skin. Details on this can be found in [12] and in [11], here we focus on the actual grasping procedure.

The precondition for our grasping skills is that the object to be grasped was recognized and localized by means of the vision system and the gripper was moved to an "optimal" approach position with the object being in the field of view of the hand camera. From there visual servoing based on a structured light approach is used to align the gripper so that the object is in between the gripper jaws. Finally the gripper is closed using force feedback from it's tactile sensors.

Several different grasping strategies were implemented to grasp objects from the side, from the top, with rotational alignment (e.g. grasping cuboid objects or lying cylindrical objects lying on the table) and without rotational alignment (vertical cylinders to be grasped from the side).

The implemented grasping strategies enable the robot to successfully grasp all convex objects that geometrically fit into the gripper.

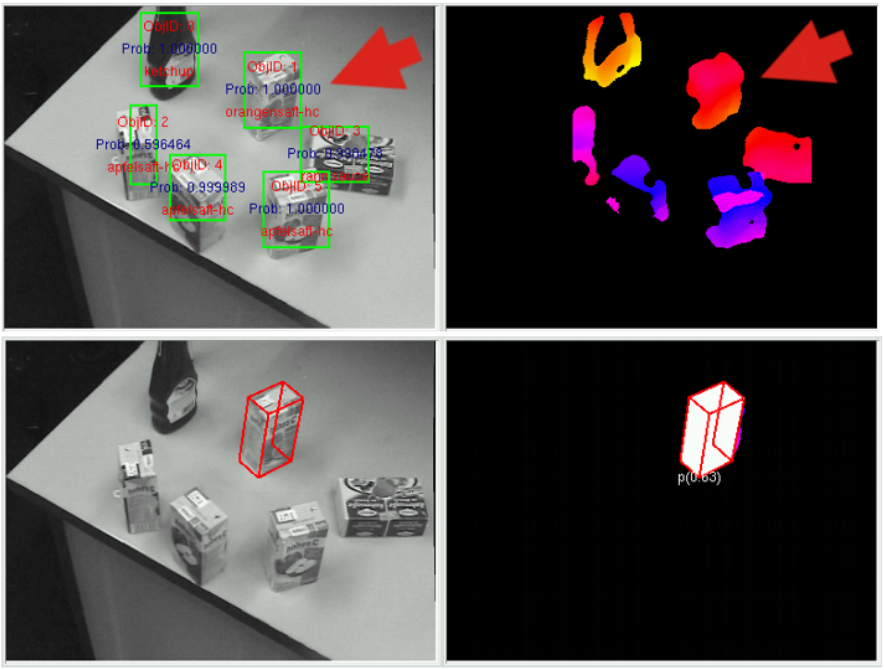

Fig. 6. The top left image was rectified for the stereoscopic 3D-reconstruction and the classification results are plotted into the image. The top right image shows the segmented depth data. The images in the lower row display the results of the pose and shape reconstruction.

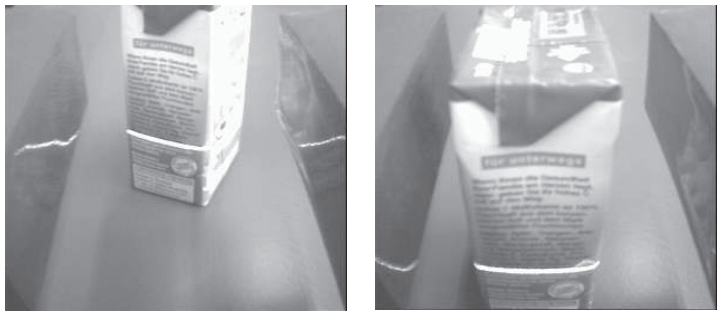

Fig. 7. The view from the gripper camera at the beginning and endpoint of the grasping motion.

Face detection

The hybrid face detection module was developed by Viisage Technology AG. Originally it is part of FaceFINDER, a real-time face recognition system[3]. Inside the face

[3] For more information on FaceFINDER see: http://www.viisage.com

detection module several biologically motivated cues (feature and model based) are used to detect faces present in the current camera image in real-time. In Fig. 8 the structure of the face detection module is given.

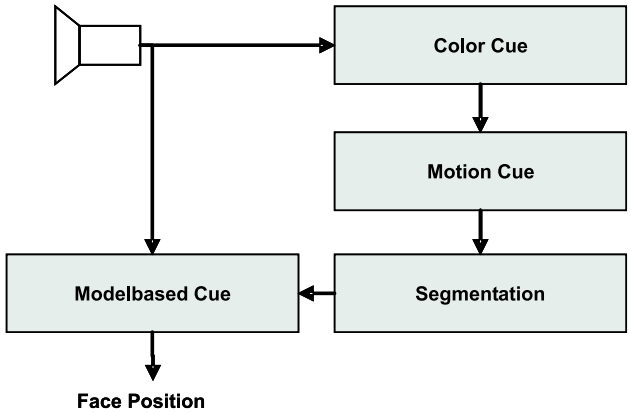

Fig. 8. Structure of the face detection module used for face finding in the camera images.

The camera images are analyzed by using color and motion cues to separate relevant segments from the background. The resulting segments are classified by a neural net, trained with representative face models. The classification results are used to extract the corresponding face position inside the given image. Fig. 9 depicts a typical result along with the individual cues used by the face detection module.

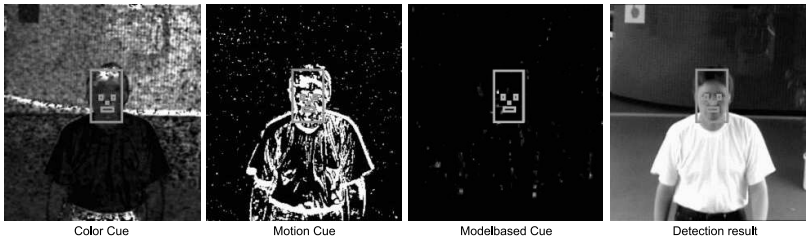

Fig. 9. Face detection cues and detection result.

The face detection results are then fed into the robot's control system to trigger appropriate actions (see section 2.3).

Speech I/O

Speech is the major communication channel used by humans and should also be used by service robots targeted at consumer applications. In the MobMan system the

Siemens Corporate Technology integrated environment ViCA performs speech I/O as well as dialog control.[4]

Within ViCA the recognizer component performs robust speaker independent recognition of continuous speech or keyword spotting. The recognizer is especially focussed on recognition in noisy environments.

The mixed initiative dialog control component is based on declarative action descriptions rather than on procedural finite state transition networks. Actions are described at a conceptual level as n-place predicates. The dialog interpreter maps user utterances e.g. ”bring me some orange juice” to action descriptions e.g. ”bring (receiver, object)” thereby asking the user if some information is missing.

Finally the dialog result is put into the robot’s control system and appropriate robot actions are triggered.

2.3 Higher-Level Behaviours

The basic robot functions described in the previous sections are the base for building the higher level behaviours needed for applications of the system (e.g. the integrated demonstration of section 3). These higher-level behaviours are controlled using the sequencing layer described in section 2.1. Implemented higher-level behaviours include (in the order of increasing complexity):

- Recognize, localize and grasp objects
- Look for a person
- Place objects on bar or table
- Hand over objects to humans
- Open door
- Navigate to a specified place
- Clean up table/bar
- Go to next room and get a specified object

The complexity of the environment the system can deal with depends on the performance of the underlying functional modules, primarily on the power of the perception system and the robot’s fine manipulation capability. Currently our system is able to robustly recognize in the order of five to ten known real-world objects in a moderately complicated real (uncontrolled) environment. The manipulation capability of the system is limited by its simple parallel jaw gripper which allows the robot to grasp objects, but prohibits any kind of fine manipulation. An additional limitation for many real world tasks (e.g. open a bottle), arises from the lack of a second arm.

3 Real-life Test and Experience

The demonstration scenario foresees that the robot should assist a human barman by fetching objects (i.e. soft drink cans or tetrapacks) from a repository (at the Han-

[4] For more information about ViCA see: https://partnerdialog.siemens.com/hipathready/show.php?mode=product&CatID=634&NewsID=7683&id=161&lang=de

nover fair just a table, in Berlin extended to a cupboard with several sliding doors) and handing them over to the barman or the customer. Alternately the objects should be placed onto the bar. Furthermore, the system should detect the presence of possible guests and ask them for their wishes. During idle time, i.e. if it is not busy serving other requests, the bar should be cleared from any objects left there by guests. All these tasks should be fulfilled in close cooperation and dialog with the human barman. During all motions the artificial skin should enable the user to move the robot's arm by simply pushing it.

The goal of the demonstration was to have the robot interacting with people in a realistic environment for about one week. The robot performed as expected despite of the very demanding (with respect to lighting conditions, noise etc.) trade fair environment. In particular object and voice recognition, as well as face detection proved to be very robust. The system successfully managed a lot of unforeseen visitor activity. With regard to user acceptance we can say, that reactivity and plausibility of the robots behaviour are extremely important. These criteria are immediately followed by sufficient motion speed; people expect robots to move with a speed comparable to human beings.

The anthropomorphic interaction channels provided by the system – tactile interaction and speech – as well as special behaviours purely targeted at interaction (e.g. that the robot looks into the users face when talking to him) are extremely important for the overall system, simply because they significantly lower the barrier between human and robot.

4 Conclusions and Future Work

We have described our robot MobMan, its architecture and functions and some experience in exposing the system to the real world. From this experience we conclude, that it is possible to build robots that robustly perform service tasks outside the labs. The system can of course not cope with the full complexity of unconstrained environments (e.g. in private households). However, with moderate restrictions on the environmental complexity today's technology allows for robust, and dependable systems of some use. Of course, there is still a strong need for substantial improvements, primarily in the areas of perception and manipulation. One of our next steps will be to systematically extend the perception capabilities of the system with respect to scene complexity (including dynamic scenes) and number of robustly distinguishable objects.

For the user acceptance of such systems appropriate communication and interaction skills are crucial. The MobMan robot possesses all necessary communication channels, but we feel that the overall behaviour should also contain some amount of entertaining, affective features.

References

1. H. Bunke and X. Jiang. *Dreidimensionales Computersehen. Gewinnung und Analyse von Tiefenbildern*. Springer, 1997.
2. Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J. Flynn, Horst Bunke, Dmitry B. Goldgof, Kevin K. Bowyer, David W. Eggert, Andrew W. Fitzgibbon, and Robert B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):673–689, 1996.
3. K. Konolige. Small vision system: Hardware and implementation, 1997.
4. K. Konolige and D. Beymer. Sri small vision system. user manual – software version 1.4, 1999.
5. David Kortenkamp, R. Peter Bonasso, and Robin Murphy, editors. *Artificial Intelligence and Mobile Robots: Case Studies of Sucessful Robot Systems*. The MIT Press, 1998.
6. Gisbert Lawitzky. A navigation system for cleaning robots. *Autonomous Robots*, 9:255–260, 2000.
7. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, April-June 2002.
8. Henry Schneiderman. *A Statistical Approach to 3D Object Detection Applied to Faces and Cars*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2000.
9. Georg von Wichert, Jochen Bauer, and Björn Magnussen. Mobile Manipulation in Alltagsumgebungen. In Gisbert Lawitzky, Wolfgang Grimm, Erwin Prassler, and Peter Weierich, editors, *Intelligente Serviceumgebungen*, pages 135–150. Shaker, Aachen, 1999.
10. Georg von Wichert, Thomas Wösch, Steffen Gutmann, and Gisbert Lawitzky. MobMan – Ein mobiler Manipulator für Alltagsumgebungen. In R. Dillmann, H. Wörn, and M. von Ehr, editors, *Autonome Mobile Systeme 2000*, pages 55–62. Springer, 2000.
11. Th. Wösch and W. Neubauer. Grasp & place tasks for domestic robot assistants. In *2nd International Workshop on Advances in Service Robots (ASER'04)*, Stuttgart, Germany, May 2004.
12. Th. Wösch, W. Neubauer, G. v. Wichert, and Z. Kemény. Robot Motion Control for Assistance Tasks. In *11th IEEE Intern. Workshop on Robot and Human Communication (ROMAN'02)*, pages 524–529, Berlin, Germany, September 2002.