An Interactive Interface for Service Robots

Elin A. Topp, Danica Kragic, Patric Jensfelt and Henrik I. Christensen
Centre for Autonomous Systems
Royal Institute of Technology, Stockholm, Sweden
Email: {topp,danik,patric,hic}@nada.kth.se

***Abstract*— In this paper, we present an initial design of an interactive interface for a service robot based on multi sensor fusion. We show how the integration of speech, vision and laser range data can be performed using a high level of abstraction. Guided by a number of scenarios commonly used in a service robot framework, the experimental evaluation will show the benefit of sensory integration which allows the design of a robust and natural interaction system using a set of simple perceptual algorithms.**

I. Introduction

Our aging society will in the near future require a significant increase in health care services and facilities to provide assistance to people in their homes to maintain a reasonable quality of life. One of the potential solutions to this is the use of robotic appliances to provide services such as cleaning, getting dressed, or mobility assistance. In addition to providing assistance to elderly it can further be envisaged that such robotic appliances will be of general utility to humans both at the workplace and in their homes. A number of human-robot interfaces have been built to instruct a robot of what task to perform, ranging from basic screen input to natural language communication [1]–[4]. It is not only necessary to equip a service robot with technical means of communication, but also to make those usable for inexperienced users, which is related to the questions of "How should the communication be performed?" and "How can the robot give feedback about its state?". To answer these questions we have decided to study a set of typical use cases or communication scenarios.

One important issue for giving feedback while communicating with a user is an "attention" mechanism allowing the robot to keep the user in the field of view. The three major problems arising are i) representation (How to connect perception to action?), ii) system's design (What are the necessary control primitives required to control the behaviour of the robot?), and iii) sensory feedback (What types of sensors are needed to achieve a natural way of interaction?). In this paper, we deal with these issues.

Psychological studies presented in [5] have shown that people have different attitudes towards automated systems, often strongly related to system performance and the feedback. A user study reported in [6] pointed out the importance of user-friendly interfaces and the ability of the system to convey to the user how it should be used or what type of interaction is possible. More precisely, it is important to design a system with the ability to show its state to the user. As an example, while the robot is communicating with the user, a camera may be used to keep the user in the field of view corresponding to an "eye-to-eye" relation.

To achieve the above mentioned attention mechanism or focusing ability for an "eye-to-eye" relation, a tracking system is needed. We integrate vision and laser range data for robust person tracking and user detection to establish the communication between a user and the system. We use a state based approach to handle the different phases of communication. Once the user is detected (the communication is established), we integrate speech and gesture recognition for detailed task specification. The modeled states are related to typical communication cases that may occur between a service robot and a user. We will show how integration of different sensory modalities on a high level of abstraction may be used to design an interaction system and describe the advantages of sensory integration.

Related to the design and experimental evaluation both in this paper and in general, we can distinguish between interfaces from a strictly social point of view (evaluation of the interaction system) and, so called, goal oriented interaction. Our approach falls into the latter. The interaction is goal oriented as it is used to specify robot tasks and explain intentions.

The outline of the paper is as follows. We begin with a general description of the system design. In section III we give an overview of related work and use this to motivate some of our design decisions in Section II. Section IV describes the architecture and Section V the implementation. Experimental results are presented in Section VI and a summary is given together with some ideas for future work in Section VII.

II. System Design

The service robot is aimed at operation in a natural domestic setting, performing fetch and carry type tasks. The system is to be used by regular people for operation in an unmodified setting, which implies that it must rely on sensory information for navigation, interaction and instructions. This section presents some of the general design principles used in our approach and proposes a set of basic modalities and sensory data types necessary to design an interactive interface.

A. Use cases

We have based our initial design on the four different interaction principles or use cases common in a service robot framework. These are presented in Figure 1:

- the user wants to provide the robot with information,
- the user requires information from the robot,
- the user gives an action command to the robot and

- the user wants to teach the robot which requires that the robot observes the user's actions.

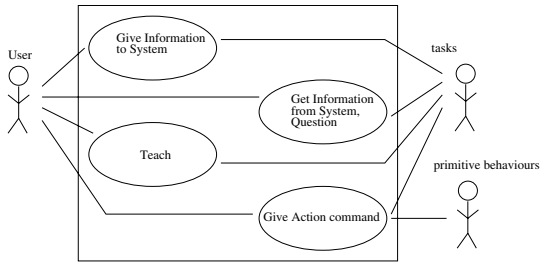


Fig. 1. The four basic use cases for an interactive interface

In all cases, the communication between the user and the robot has to be established before any of the use case scenarios can be initiated, and in all cases the communication has to be ended. Thus, a whole scenario can be divided into three basic phases: i) establish communication, ii) communication phase (involving the use cases), and iii) terminate communication.

This has lead us to use a state based approach with a finite state automaton described as $\{\mathbf{S}, S_0, \mathbf{X}, \delta, \mathbf{S_a}\}$, where the set $\mathbf{S}$ contains the states, S_0 represents the start state of the automaton, $\mathbf{X}$ is the accepted input alphabet and δ defines the transition function. $\mathbf{S_a}$ is the set of accepting states.

In the simplest case, the set of basic states would consist of

- a "wait" state - the system observes the environment for particular events (start and accepting states)
- a "start communication" state - the system actively searches for a user and initiates the communication sequence,
- a "communication" state - the user interacts with the robot, possibly controlling some of its actions and
- a "stop" state - the system goes back to the "wait" state which could, for example, involve moving back to a "home" position.

Depending on the use case scenario, the "communication" state can be modeled as a sub-automaton, with states that represent the respective use cases. To handle unexpected situations or errors an additional error state is introduced that can be reached from other states in cases when the system faces a specific problem.

B. Experimental Platform

The platform used for experiments is a Nomadics Technologies 200 with an on board Pentium 450MHz. On the top of the turret there is a Directed Perception pan-tilt unit with a Sony XC-999 CCD colour camera on it. A SICK PLS 200-114 laser range finder is mounted at a height of 93cm. For low level motor control and coordination the Intelligent Service Robot (ISR, [7]) architecture is used.

III. Motivation

This section gives a short overview of the related work. We will concentrate only on systems that are based on sensory modalities such as vision, laser and speech.

A. Integrated systems

An example of an integrated system, is presented in [1]. The system integrates different modules in a state based control loop where the modalities used for interaction are dialogue and vision based face tracking.

Although dialogue and vision based tracking are run in parallel, there is no specific integration of these modules. In contrary to this, our systems integrates sensory input depending on the current state. The basic design is similar in that it uses a state based approach.

Another system that integrates different modalities is presented in [2]. The authors integrate language (command) processing and gestures for deictic information. Both can be given either "naturally" or by using a hand held PDA.

Our system is based on similar input modalities (language commands and gestures), but considers also the laser data as an additional input. In addition, our design is more general and allows the use of different input states for sensory modalities.

The Nursebot, [3], [4], provides a framework for personal robot assistants for the elderly and is divided into several smaller systems each covering specific applications. Its control architecture is a hierarchical variant of a partially observable Markov decision process (POMDP). It coordinates different functionalities and takes decisions for the interaction with the user. Hierarchy is required to reduce the state space, as stated in [4]. A user study has been conducted where the authors claim that acceptance of the robot was fairly high and problems were mostly caused by a poorly adjusted speech system. This work made clear that it is extremely important to maintain the principle of giving appropriate feedback to the user at all times. Additionally, it states the importance of focusing on the right user when a group of people is present. Compared to this system, we have decided to follow a more general design strategy.

B. Language processing and gesture recognition

Many different approaches to language processing, in this case speech recognition and interpretation, have been presented over the years. For speech recognition, we use the HMM-based system ESMERALDA [8]. In [9], [10] a method that allows to set up a dialogue scheme based on clarifying questions is described. To be able to determine missing or ambiguous information, the user's utterances are represented in typed feature structures (TFS). We use structures inspired by those TFS to assign spoken input to a predefined hierarchy of utterance types.

For gesture recognition we use a face and hand tracker based on skin colour detection, [11]. Here, a combination of Chromaticity coordinates and a di-chromatic reflection model is used to achieve robust skin colour based detection of hands and face. The segmented regions corresponding to the face and hands are tracked directly in the image space using a conventional Kalman filter. The matching between images is performed using a nearest neighbour algorithm, which is adequate when the algorithm is run at 25 Hz.

IV. General architecture

Our general architecture is presented in Figure 2. The control module, labeled with "Coordination and decisions" represents the basic finite state automaton. The incoming sensory data and input from the user are interpreted in the respective modules in the interpretation layer. The control module receives and handles already interpreted information, which also depends on the current state of the system.

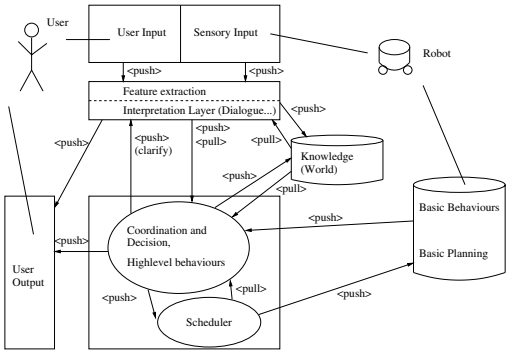


Fig. 2. An architecture for interactive interfaces

A. Modalities

Considering the use cases and typical scenarios a service robot has to handle, we have decided to integrate visual, laser and speech data for interaction as one possible set of sensors and modalities that satisfies our requirements.

To deal with the "attention" problem, we suggest a tracking module based on laser range data. We also consider a camera on a pan-tilt unit (head-neck motion) as an appropriate way to give feedback to the user about what the system currently focuses on. Since it is not possible to derive information about the user's face from laser range data, we use a combination of laser data and image based face detection for more natural (in terms of feedback) and robust tracking and detection of the user.

The most complex use case in the system is the teaching case which involves the ability of the robot to observe the user's actions and understand his/her explanations. Additionally, some control inputs from the user has to be interpreted as a pointing gesture. For this purpose, we use spoken language interpreter and vision based (gesture) tracking.

1) User detection: In its initial state, the system observes the environment until a user is detected. The robot directs its attention to potential users by turning the camera in this direction to try to verify the existence of a user. If the user hypothesis is supported by the image data, the robot starts the interaction by asking this person to verify the hypothesis. If a confirmation is received from speech recognition and interpretation, the hypothesis is marked as the user and system state is switched into the communication state. If a rejection is uttered or no response is perceived during a predefined time period, the hypothesis is marked as no longer of interest and the next hypothesis is chosen. During the verification and the confirmation step the system continuously tracks not only the hypothesized person to confirm as user, but all other person hypotheses respectively.

2) Communication: When a hypothesis is confirmed to be the user, the communication is established and the system accepts various kinds of commands as input. Depending on the received command or utterance, it switches into a certain sub-state of the communication state. Figure 3 shows some of those sub-states. As we are interested in handling a scenario that involves observing the user's actions, we have concentrated on designing the integration of speech and vision based hand tracking modalities for the "teach" sub-state.

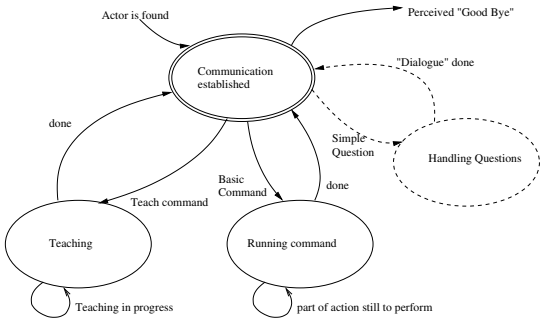


Fig. 3. The communication state with sub-states.

As our approach is state based, it is possible to interpret user's actions – or gestures in general – within the respective context of the scenario. This allows us to make the assumption that an observed movement of one of the user's hands can be interpreted as a gesture. In some cases the system expects a pointing gesture and an speech based explanation, see Figure 4.

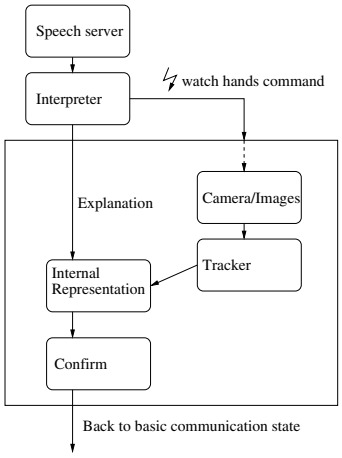


Fig. 4. Integrating gesture and explanation in the respective state.

When the system switches into this particular state, an explanation from the speech interpreter and a pointing gesture from the visual tracking system are expected. If any other spoken input is received, the system informs the user what is the type of explanation or command it expects at this point.

3) Language processing: In our system, spoken input is considered the primary control input which makes it necessary

to provide a representation that facilitates the control of the basic automaton. Consequently, we have chosen to model the control input using a taxonomy of speech acts. Every utterance that is accepted as complete is considered as a speech act. On the second hierarchy level of the system, we propose the following basic speech types: ADDRESS, COMMAND, EXPLANATION, RESPONSE, and QUESTION. These speech acts are represented in structures inspired by the typed feature structures (TFS) presented in [9], which allows us to assign features of arbitrary type to the speech act. Objects and locations for the command type speech act are represented as strings. This is sufficient to demonstrate the general way of integrating different types of information in the structures.

For the interpretation, a word spotting approach is used which is also implemented in form of a state automaton. This is possible because the expected set of utterances for our purpose consists of a small and regular subset of natural English language. Additionally, this approach has the advantage that unexpected input can be ignored. This, in its turn, reduces the number of errors resulting from speech recognition.

V. IMPLEMENTATION

The implemented system is schematically shown in Figure 5. The ISR architecture allows us to use a connection to the planning system for low level control of the robot.

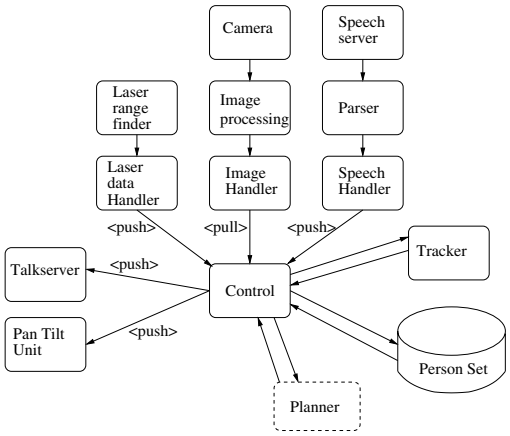


Fig. 5. The implemented system

Different types of connections for data transmission from the interpreting modules of laser, speech and image data are used. A push-type connection is used for laser data, which means that laser data is sent to the system at a fixed rate. Speech input is also received through a push-type connection. In this case it means that as soon as there is some speech data it will be sent to the system. Camera images are only grabbed when required.

Figure 6 shows a schematic overview for detecting the user. Two cues can trigger the system to start searching for the user: a) motion, and b) a spoken command given to the robot. When any of these events occur, a set representing all possible person-like hypotheses is initialized and searched for the actual user. This search is based on the assumption that the user stands rather close to the robot. For each hypotheses from the set, a verification step is performed. Apart from the correct size in the laser scan, the verification relies on the image based face detection.

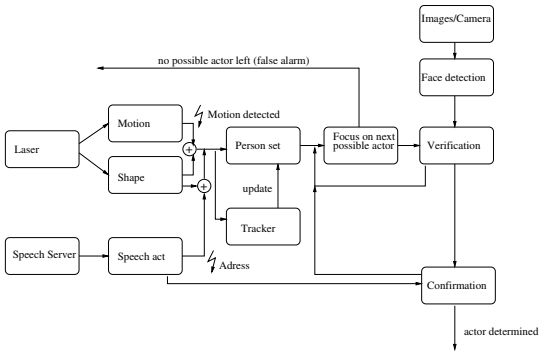


Fig. 6. Schematic overview of detecting the user.

A. Interpreting laser data

To obtain hypotheses about where there are people from laser data, two cues are used: body shape and motion. As our laser range finder is mounted on a height of 93cm, it is too high to be used for detecting "leg-like" structures. Therefore, we use the fact that a body causes a single convex pattern (see [12]) of a certain size and we use this assumption to estimate regions in which a person-like structure exist.

Movement detection can be derived from the subtraction of two consecutive scans under the assumption that the robot is not moving around in this state. This assumption seems natural since, at this stage, we consider a scenario where the user has to take the initiative of approaching and addressing the robot. However, a method to detect moving objects by a moving robot, as for example presented in [13], is one of the modules that we are considering to integrate as a part of our future work. The result of the motion cue is mapped to the observed person-like objects from the shape cue. Moving objects are then considered more likely being a person than static ones.

B. Interpreting visual information

To verify the hypotheses generated by processing the laser data, visual information is used. Face and hand tracking which is based on skin colour detection is used for the verification. The segmented blobs are thresholded based on their approximate size and position in the image. This is possible to do since the distance between the person and the robot is easily estimated from laser data. The interpreting module is responsible for delivering information about the presence of a face or the movement of the user's hands in states that require tracking and gesture recognition.

VI. EXPERIMENTAL EVALUATION

In general, our state based approach represents the three phases of a natural communication between humans quite well. We have performed a number of experiments with different users and the overall results for detecting and verifying the

user are good. The following sections present some of the example scenarios and show the advantages of our integrated system.

A. Cue integration for verification of person hypotheses

Figure 7 shows a panoramic view of the room used for experiments with hypotheses (marked with white crosses) generated by our skin colour detector. Note here that we are not using a panoramic camera - this image is just to show a number of hypothesis generated in general by a colour detector. The blobs are marked with crosses and are note yet pruned depending on their size or position. So, using colour based hypotheses generation without any additional information would give (in this example) 43 hypotheses out of which only one is correct (if only the person's face is searched for). The lower part of the figure shows a corresponding laser scan of the same static scene displayed in polar coordinates and connected to a polyline. In this laser scan, nine hypotheses for convex objects are detected of which four remain after the check for appropriate size. These are marked with arrows pointing up.

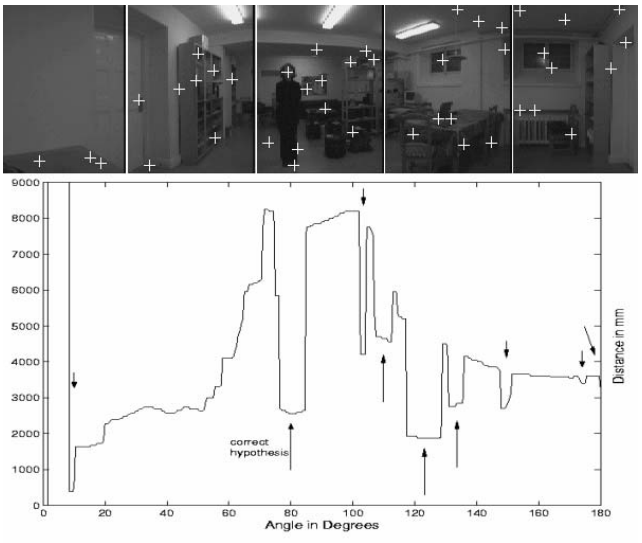


Fig. 7. Generating hypotheses separately from laser and vision data.

Integrating the colour hypotheses with those delivered by the laser data interpreter, the colour hypotheses are checked for appropriate size and position. This allows the verification of person hypotheses by combining respective information. In this particular example, only two of the hypotheses remain.

As this example scene was static, no movement information could be used to help elimination of the remaining false positives. An additional problem here is that the false positive arising from the chair is ranked as the strongest hypothesis due to being closest to the robot. The following example shows how the integration of speech helps to eliminate even this hypothesis. The basic procedure is to verify the hypothesis in a two-step process. This experiment shows the immediate benefit of sensory integration even in case of completely static scenes. In addition, it allows to keep the face of the user in the focus of attention of the camera providing also the necessary feedback to the user about the current state of the interaction system.

The next experiment, presented in Figure 8, shows how even better hypotheses verification can be performed when motion information is available. It can be seen from the figure how the ranking of hypotheses changes when one moving person is present in the scene shown in Figure 7.

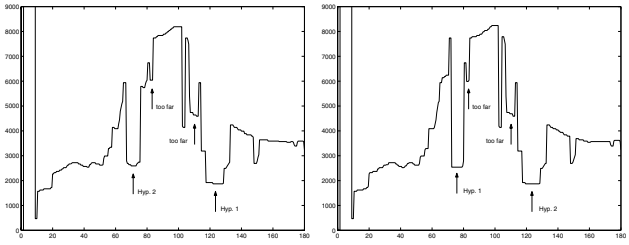


Fig. 8. One person is moving, the other hypotheses represent static objects. When the movement is detected, the ranking for the hypotheses is flipped.

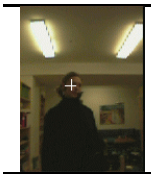
In the following, we show an experiment that presents the behaviour of the system in a scenario. The process is shown as a sequence of images together with the transcript of user utterances and output of the system. First, a new person set is initialized and the user is detected.

```
New person set initialized
SEARCH_ACTOR
```

After detecting the user, the camera is focused on the user's face and she is asked for confirmation.

```
TALK: Can I do something
      for you?
```

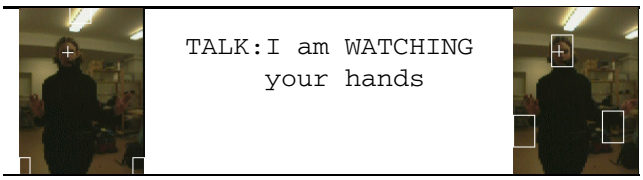
Now, the camera is oriented towards the user and the system asks what should be done since no further information was received. The user explains that she would like to show something, which implies, that a pointing gesture is to be expected.



```
TALK: What do you want me
      to do?
```

The camera moves down to focus on the hands. The visual tracker is initialized by assuming the hands in the lower corners of the image and the face at the middle of the upper bound, as indicated by the boxes in the images. From this initial position it adjusts itself to the actual position of hands and head. Both hands are tracked to determine which hand is

moving. The head is tracked too as to maintain the assumptions required by the tracker (see [11]). One of those assumptions is that the hands are always at a lower position than the head.



When the hand stops, the tracker is also stopped and the final position of the moving hand is used to compute the position relative to the robot (x - and y-coordinates in mm). This demonstrates that it is sufficient to have a gesture recognition running only when required by the communication state. If the visual tracker and interpretation of its results had been running in parallel to the attention part, it would have been obviously very expensive in terms of calculation time. So one of the general results for the integration of speech and gestures is, that both support each other:

- Gestures give the missing deictic information and
- spoken input allows to start a gesture recognition only when necessary.

When the tracker has stopped, the camera is directed to the user's face again and she is asked if something else should be done. In the experiment the answer is "good bye", which makes the system return to the start state.

A comparable experiment with a second user who was introduced to the system for the first time, has shown that a very short explanation was sufficient to use the system.

To summarize, the experiments show that the combination of very simple and therefore computationally inexpensive modalities, helps to achieve an overall robust system that allows to maintain the proposed interaction principles.

B. State based integration of speech and gestures

The approach used for gesture interpretation is a rather simple one. Still, a very important result obtained was: With the help of the context information, which can be derived from the system state, the rate of false positives in terms of pointing gesture recognition can be reduced drastically. The occurrence of a specific gesture is expected only in certain states and the advantage of this approach is obvious:

a) A computationally expensive gesture recognition system can be initiated exactly when required, and
b) the likelihood of recognising a certain type of gesture instead of some arbitrary gesture is therefore higher.

Our next step is to improve the design of the gesture recognition module by using results from extended user studies.

VII. Conclusion

We have presented the initial design of our human robot interaction system. The main contribution of our work is threefold: i) consideration of perception-action loops and their modeling, ii) design based on use cases, and iii) integration of multiple sensory feedback to achieve flexibility and robustness. A number of related systems have been presented and compared to the proposed architecture. We have shown that, by considering an integration framework, even rather simple algorithms can be used to design a robust system that also allows for a natural communication between the user and the robot.

Our future work will concentrate on enhancing the tracking abilities to tracking with a moving robot and further on providing additional algorithms that will allow for more complex action of the robot. Some of them are manipulation of objects where the need for object recognition and pose estimation is an obvious requirement, [14].

Acknowledgment

This paper is based on a master thesis project of Professor R. Dillmann's group "Industrial Applications of Informatics and Microsystems" at the Institute for Computer Design and Fault Tolerance, Fakultät für Informatik, Universität Karlsruhe (TH), Germany. The thesis project was conducted at the Centre for Autonomous Systems, Royal Institute of Technology, Stockholm, Sweden. We would like to thank Professor Dillmann, for making this possible.

References

[1] M. Zobel, J. Denzler, B. Heigl, E. Nöth, D. Paulus, J. Schmidt, and G. Stemmer, "MOBSY: Integration of vision and dialogue in service robots," *Machine Vision and Applications*, 1(14), pp. 26–34, 2003.
[2] D. Perzanowski, W. Adams, A. Schultz, and E. Marsh, "Towards Seamless Integration in a Multi–modal Interface," *Wokrshop on Interactive Robotics and Entertainment*, AAAI Press, 2000.
[3] G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Magaritis, M. Montemerlo, J. Pineau, N. Roy, J. Schulte, and S. Thrun, "Towards Personal Service Robots for the Elderly," in *Workshop on Interactive Robots and Entertainment (WIRE)*, 2000.
[4] M. Montemerlo, J. Pineau, N. Roy, S. Thrun, and V. Verma, "Experiences with a Mobile Robotic Guide for the Elderly," in *National Conference on Artificial Intelligence, AAAI*, 2002.
[5] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, 39(2), pp. 230–253, 1997.
[6] H. Hüttenrauch and K. Severinson-Eklundh, "Fetch-and-carry with CERO: Observations from a long-term user study wit a service robot," in *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 158–163, Sept. 2002.
[7] M. Andersson, A. Orebäck, M. Lindström, and H. Christensen, "ISR: An Intelligent Service Robot," *Lecture Notes in Computer Science* (Christensen, Bunke, and Noltemeier, eds.), vol. 1724, Springer, 1999.
[8] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In Václav Matoušek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Heidelberg, 1999. Springer.
[9] M. Denecke and A. Waibel, "Dialogue Strategies Guiding Users to their Communicative Goals,"*Proceedings of Eurospeech*, 1997.
[10] M. Denecke, "Rapid Prototyping for Spoken Dialogue Systems," in *Proceedings of the COLING'02*, Aug. 2002.
[11] F. Sandberg, "Vision Based Gesture Recognition for Human-Robot Interaction," Master's thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, 1999.
[12] B. Kluge, "Tracking Multiple Moving Objects in Populated, Public Environments," in *Lecture Notes in Computer Science* (Hager, Christensen, Bunke, and Klein, eds.), vol. 2238, pp. 25–38, Springer, 2002.
[13] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers, "Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association," in *Proceedings of the IEEE International Conference on Robotics & Automation (ICRA)*, 2001.
[14] D. Kragic, "Visual servoing for manipulation: Robustness and integration issues," in *PhD Thesis*, (KTH, Sweden), 2001.