User-adaptive Interaction with Social Service Robots

Andrea Scheidig, Steffen Müller, Horst-Michael Gross

This contribution describes the objectives, the application scenarios and first results of the long-term research project Horos **(HOme RObot System). In the focus of this research are perceptual, reasoning and learning techniques to realize user-specific and user-adaptive interaction strategies on mobile service assistants. Expected results of this project are basic methods, algorithms and architectures and their integration and long-term experimentation on social service robots interacting with users in public and domestic environments.**

1 Project Objectives

For robots to exist in everyday human environments, they will have to be able to interact with humans in a naturalistic manner via verbal and non-verbal communication and to adapt their dialog strategy to the specifics of the interaction partner. Therefore, the overall objectives of the Horos project are to study the perceptual, reasoning, learning and motor capabilities required for social service robots acting in human-centered environments.

In the focus of this research is the development of techniques to realize user-specific and user-adaptive interaction strategies taking into account the perceived states and intentions of the current interaction partner. Against this background, there are two main challenges in learning - the offline learning of describing user features, like the user's gender, age or emotions, and, based on this, the learning of an adequate dialog strategy. By adapting the dialog to the current user two different strategies will be distinguished in the project. There are short-term dialogs with changing users, where only few data from a specific user are available, and thus only a restricted dialog adaptation is possible and useful. This is typical for HRI in public environments. Beside this, a long-term interaction with the same user or a small group of users, e.g. in a domestic environment, provides more information about the user: static information, like age or gender, and dynamic information, like his or her current emotions, preferences and intentions. Therefore, for short-term dialogs typically fixed user models initialized by statistical data are utilized while adaptive user-models are better suited for long-term dialogs. However, most of the current applications use only short-term dialogs where the interaction with all users is the same [1].

As soon as a robot becomes a long-term interaction partner, the robot needs to be able to treat each person as a distinct individual and to personalize the dialog. Therefore, in the Horos-project we are focusing on two complementary scenarios to investigate both aspects of HRI, an interactive mobile information system interacting with different people in a public environment, and a personal assistant for a domestic application. In both scenarios, we use our mobile robots Horos as platforms (see Fig. 1).

2 Related Work

Mostly used approaches are rather simple, static and not adaptive like the Stereotypes [2]. Beside these, there are only a few adaptive approaches, like reinforcement learning based systems realizing a simple user-adaptivity based on a person's feedback to the robot's actions. One of the first examples for such a user-adaptive dialog was the learning of user-preferred robot movements based on feedback signals of the interacting person [3]. However, the uncertainty in the communication process between human and robot, especially in naturally spoken dialogs, is an important problem of more natural and complex scenarios [4]. One approach to overcome this uncertainty is to model the cognitive states of the users probabilistically by means of POMDP's. Using this technique, the robot is able to learn an optimal dialog strategy to maximize the dialog reward. An example for that is presented in [5], where a robotic nursing assistant for the elderly plans its verbal dialog strategy by means of this technique. However, approaches that allow a more detailed modeling of the cognitive state and intention of the current user taking into account visual and auditory observations of age, gender, emotions, interest and instructions are not yet known.

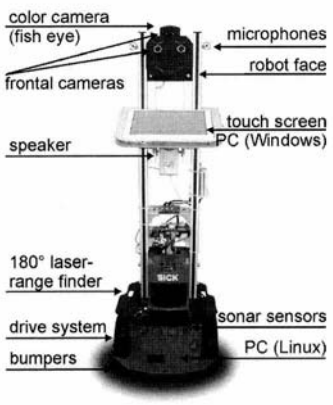

Figure 1: Sensory modalities (visual, auditory, touch, laser, sonar, tactile) and motor components (drive, speech, face) of our mobile interaction-oriented robots Horos I+II.

3 User-adaptive HRI in HOROS

In both scenarios of the Horos project different functional aspects of a user-adaptive HRI have to be investigated. Based on a stable multi-modal person detection and tracking, these include the recognition of specific user features and the estimation of the user's goals (Fig. 2, left). Depending on the concrete scenario, this concerns static features like the user's age or gender (for short-term dialogs with changing users) or additional dynamic features, like the facial expression, gestures or the voice prosody of the current user (for long-term dialogs with the same user). For the association of the perceived user features with specific robot behaviors (Fig. 2, right), we are developing and investigating different adaptive user models described in Section 3.2 (see Fig. 2, middle).

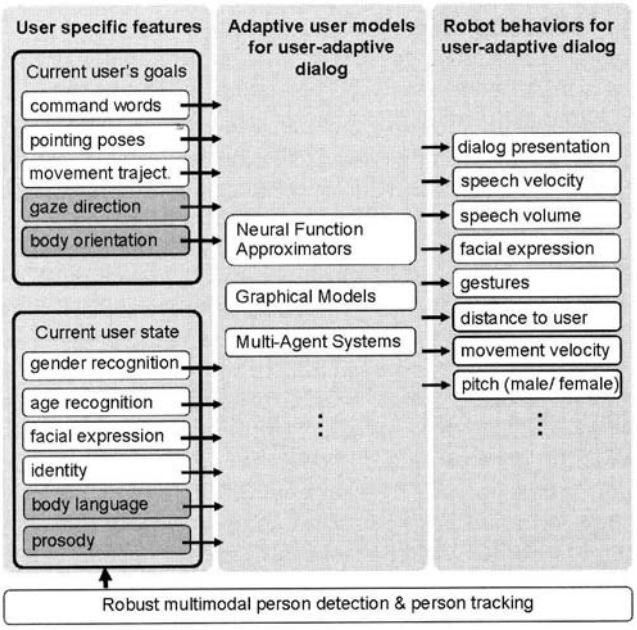

Figure 2: Main aspects to be realized for a user-adaptive HRI in the Horos project: The recognition capabilities already used (white background) or planned (gray background) are shown on the left side. Possible techniques for a user-adaptive dialog are depicted in the middle. The available robot behaviors to be adapted to the current user are given on the right side.

3.1 Recognition Capabilities

3.1.1 Robust Multimodal Multi-Person Tracking

Efficient and robust techniques for people detection and tracking are basic prerequisites when dealing with HRI in real-world scenarios. Therefore, we use a multimodal approach, which can be characterized by the fact, that all utilized sensory cues are concurrently processed and integrated into a robot-centered map using a probabilistic aggregation scheme [6]. The overall computational complexity of our approach scales very well with the number of sensors and modalities. As sensory cues we use the laser-range-finder (to detect typical leg-profiles), the sonar system (to detect any object under a minimum distance), the fisheye-based omni-directional vision system (to detect skin color), and the frontal vision system (to detect faces or body silhouettes) (see Fig. 1). For each of these sensory systems, specific Gaussian distributed object hypotheses are generated. Each Gaussian distribution models the belief to observe a person. These probability distributions are further merged into one map by means of a flexible probabilistic aggregation scheme based on Covariance Intersection (CI). The main advantages of this approach are the simple extensibility by integration of further sensory channels, even with different update frequencies, and the usability in on-line recognition tasks [7]. In our ongoing work, we are extending the system with additional cues to further increase the robustness and reliability for demanding real-world environments (especially in operation areas, where the current sensory cues are insufficient). Currently, we are working on the integration of a voice-based speaker localization and tracking [8] and on the integration of a head-shoulder-based visual tracker.

3.1.2 Estimation of User Intentions

To estimate possible goals of the current interaction partner, we utilize a touch-based tactile dialog, speech commands and non-verbal instructions by means of pointing poses or head poses (see Fig. 2, left).

Movement trajectories: A prerequisite for estimating the user intentions is the capability of a socially interactive robot to estimate the user's interest to interact with it. Based on this estimation, the robot can adapt its dialog strategy to the different behaviors of the respective person. Against this background, we currently use a multi-modal approach to estimate the interest of people to interact with the robot by analyzing the movement paths people typically take in the surroundings of a robot [12]. Relevant known approaches estimate the movement direction of a person only by means of distance sensors [13] or determine the movement goals in a local operation area using several fixed laser sensors [14], techniques which are not suitable for a mobile robot that has to react autonomously in a large-scale and highly populated operation area. In our approach, different movement behaviors, e.g. "slow or fast passing by" or "going ahead to the robot" need to be distinguished to allow an adequate and pro-active reaction of the robot. Therefore, in the project we try to learn a direct mapping from an observed movement trajectory as expression of a current user goal to a specific robot articulation generated by the user-adaptive dialog control (see Section 3.2), e.g. a specific voice response or facial expression or an attracting robot movement. As a preliminary approach, we classify the perceived movement trajectories of tracked people in several categories that could express underlying interest of people [12]. Based on this, an appropriate robot reaction for each classified movement trajectory can be generated.

Pointing poses: Based on this person tracker, we developed a hierarchical neural architecture that is capable of estimating a target point at the floor given a pointing pose, thus enabling a user to command his mobile robot to a specific target position in his local surroundings by means of pointing [11]. The achieved recognition results demonstrate that it is possible to realize a user-independent pointing pose estimation using only monocular images given of the low-cost frontal camera (see Fig. 1), but further efforts are necessary to improve the robustness of this approach for everyday application. Planned improvements concern the application of

a foreground extraction routine based on active shape models [17] or the analysis of the movement of the pointing arm to the final pose, which contains additional information that could be exploited to enhance the precision of the estimation.

Gaze direction and body orientation: To enhance the estimation of the interest of a person to interact with the robot, we are currently integrating further features like the head pose [15, **?**] together with the body orientation.

3.1.3 Estimation of the User State

For the estimation of the user state, we mainly use multimodal cues to estimate age, gender, facial expression, identity, prosody and body language (see Fig. 2 left) in the Horos project.

Age, gender, identity, emotion: For the observation of these features we mainly utilize vision-based cues (see [10]). In this work, alternative approaches for extracting features from face images and several classifiers are compared with respect to their applicability to allow an on-line classification of gender, age, facial expression, and identity of the tracked user. The used models are i) a description of the face images by their projection onto an ICA-based subspace and ii) an Active Appearance Model (AAM) which describes the shape and gray value variations of the face images. Best results for the estimation of gender and facial expressions were obtained by using AAMs and MLP classifiers or ICA-subspace projections in combination with Nearest Neighbor classifiers. In our future work, we will enhance these estimation results by integrating further modalities, like audio-based speaker features analyzing the voice of the user. For example, the prosody can provide very useful information about the current state of the interaction partner (e.g. [18]).

Body language and dynamic gestures: Since both features do also describe the current state of a user and his satisfaction with the interaction process, in the near future new techniques [19, 20, 21] will be investigated with respect to their suitability for classifying dialog-relevant gestures or for finding elementary parts of specific body gestures.

3.2 User-Adaptive Dialog Control

The aspect of a multi-modal user-modeling addresses the transformation of observed user states and intentions into suitable multi-modal robot behaviors, like speech outputs, adequate movements, simple robot gestures or mimics, etc. (see Fig. 2, right). For instance, perceiving an older person which wants to be guided to a specific location, requires a guided tour probably using a lower movement velocity than the same tour for a younger person. Depending on the used scenario, different adaptation methods for the user-model will be required. Thus, scenarios with many different users and short-term dialogs do not require a specific adaptivity for each individual user rather than for typical user groups, like young or old, male or female persons. Unlike, scenarios with long-term dialogs with the same user (e.g. in a domestic application) require a very specific and on-line adaptation to the preferences and intentions of that specific user. A main problem of the short-term dialog is that in the context of a particular real-world application, numerous different dialog strategies have to be tried out by the robot - a process which would be very time-consuming. More difficult is the acquisition of appropriate feedback from the different users about the dialog took place and its success. However, such an immediate feedback is necessary to evaluate the dialog strategy currently executed. Unlike, in long-term dialogs more dialog strategies can be tried out during interaction with the same user, and also alternative feedback signals can be extracted to get an evaluation of the ongoing dialog, like e.g. facial expressions, variations in voice melody, interaction distance between robot and user, or others. In our project, we are currently investigating the following approaches to realize a user-adaptive dialog control:

Neural Function Approximators: This approach is very similar to the RL-based Backgammon playing approach developed by Thesauro [22], where the values of the experienced game states are learned with respect to a particular policy. In our approach, also a neural function approximator is to learn the value of a typical sensorimotor dialog situation for a user-specific successful dialog. The dialog situation can be described by the audio-visual observations of the current user state and extracted user goals and by a set of possible and favorable robot actions. One disadvantage of such a direct mapping is the enormous complexity of the domain, which necessarily limits the approach to a scenario where only a few actions and behavior parameters have to be learned.

Multi-Agent Systems: A first approach to break down the complexity of the problem is the application of Multi-Agent-Systems (MAS). Here, the individual learning instances, the agents, can specialize on different subtasks of the decision problem. So a variety of questions arise. On the one hand, if the structure of the MAS is manually designed, we have to investigate methods for coordination of competing agents or for assignment of rewards. On the other hand, if the MAS structure is not predefined, the field of self-organizing problem decomposition opens for further investigation. One interesting aspect, already mentioned is the fact that one can not be sure about the current cognitive state of the user. This deficit can also be handled using MAS, by means of deploying special agents. These only try to maximize the certainty of the user information and have to compete with the others to apply their actions.

Graphical Models: A more elegant way for integrating the consideration of uncertainty and the aspect of fast learning of strategies is given by probabilistic reasoning. The keyword of Graphical Models subsumes a variety of methods applicable for dialog management, too. On the one side, these approaches are very flexible. For example, multi-modal and diffuse or contradictory inputs can be used for realizing a more intuitive and predictable dialog control. On the other side, enormous effort is still necessary for reasoning in more complex and realistic dialog models. Basic steps for handling the complexity utilizing approximation methods or by factorizing the process have been done by Thrun and colleagues [5]. In our project, we want to take these methods up for integration of extracted multi-modal user features, suitable behavior patterns of the robot, and self-evaluation based on internal drives and external feedback from the user.

4 Summary and Conclusions

In this paper, we presented the general idea and the main objectives of the Horos project, together with first, already published results. For the Horos project and its two applica-

tion scenarios in public and domestic environments, several main challenges were defined and shortly described. These include the estimation of the user's state and the recognition of his intentions and instructions together with the on-line adaptation of the dialog strategy according to these observations. Our future research will therefore be focused on techniques and methods allowing a user-specific on-line adaptation of the dialog strategy by means of an evaluating direct or indirect feedback from the interaction partner. However, a prerequisite for this are robust and fast techniques to recognize the changing user states and their audio-visual expressions.

Acknowledgments: The authors wish to thank H.-J. Böhme, Heike Groß, A. König, Chr. Martin, Chr. Schröter, Sabine Schulz, and T. Wilhelm for conceptual contributions, software implementations, technical support, experimental investigation, and stimulating discussions during the course of this research.

References

[1] Fong, T., Nourbakhsh, I., and Dautenhahn, K.. A survey of socially interactive robots. In: Proc. Robotics and Autonomous Systems, 42, pp. 143 - 166, 2003.

[2] Fong, T., Thorpe, C., Baur, C.. Collaboration, dialogue, and human-robot interaction, In: Proc. Int. Symp. on Robotics Research, 2001.

[3] Dautenhahn, K.. Embodiment and Interaction in Socially Intelligent Life-Like Agents. In: Computation for Metaphors, Analogy, and Agents, LNCS 1562, pp. 102-141. Springer 1999.

[4] Roy, N., Pineau, J., Thrun, S.. Spoken Dialog Management for Robots. In: Proc. of the Association for Comp. Linguistics, 2000.

[5] Pineau, J., Thrun, S.. High-level robot behaviour control with POMDPs. AAAI Workshop on Cognitive Robotics. 2002.

[6] Martin, C., Schaffernicht,E., Scheidig, A. Gross, H.-M.. Sensor Fusion Using a Probabilistic Aggregation Scheme for People Detection and People Tracking. to appear: Robotics and Autonomous Systems, 2006.

[7] Scheidig, A., Mueller,S., Martin,C. and Gross, H.-M.. Generating Movement Trajectories to Estimate the Human's Interest to Interact with a Robot. to appear: In Proc. of RO-MAN 2006.

[8] Brueckmann, R., Scheidig, A., Martin, C., Gross, H.-M.. Integration of a Sound Source Detection into a Probabilistic-based Multimodal Approach for Person Detection and Tracking. In: Proc. AMS 2005, pp.131-137, 2005.

[9] Wilhelm, T., Böhme, H.-J, Gross, H.-M.. Classification of Face Images for Gender, Age, Facial Expression, and Identity. In: Proc. of ICANN 2005, LNCS 3696, pp. 569-574, Springer 2005.

[10] Wilhelm, T.. Nutzerwahrnehmung f"ur eine nat"urliche Interaktion mit mobilen Servicerobotern. In: Künstliche Intelligenz 3, 2006.

[11] Gross, H.-M., Richarz, J. Mueller, S., Scheidig, A. and Martin, C.. Probabilistic Multi-modal People Tracker and Monocular Pointing Pose Estimator for Visual Instruction of Mobile Robot Assistants. to appear: Proc. IEEE World Congr. Comp. Intell., 2006.

[12] Scheidig, A., Mueller, S., Martin, C., Finke, M., and Gross, H.-M.. Recognition of Movement Trajectories as Expression of Users's Interest in Human-Robot Interaction. to appear: Proc. of RO-MAN 2006.

[13] Finke, M., Koay, K., Dautenhahn, K., Nehaniv, C.L., Walters, M.L., and Saunders, J.. Hey, I'm over here - How can a robot attract people's attention? In: Proc. RO-MAN 2005, pp. 7 - 12

[14] Bennewitz,M., Burgard, W., Thrun, S.. Learning motion patterns of persons for mobile service robots. In: Proc. IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 3601-3606, 2002.

[15] Stiefelhagen, R., Yang, J., Waibel, A.. Modeling Focus of Attention for Meeting Indexing. In: IEEE T-NN, 2002.

[16] Brunske, J., Abraham-Mumm, E., Pauli, J., Sommer, G.. Head-pose estimation from facial images with Subspace Neural Networks. In: Int. Neural Netw. and Brain Conf., pp. 528-531, 1998.

[17] Cootes, T., Taylor, C., Cooper, D., and Graham, J.. Active shape models – their training and application. In: Comput. Vis. Image Understand, (61), pp. 18 - 23, 1995.

[18] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.. Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog, In: Proc. ICSLP 2002.

[19] Reng, L., Moeslund,T.B., Granum, E.. Finding Motion Primitives in Human Body Gestures. In: Proc. Int. Gest. Workshop., 2005.

[20] Lv,F., Nevatia, R., and Lee, M.W.. 3D Human Action Recognition Using Spatio-temporal Recognition Motion Templates. In: Proc. HCI Workshop (Part of ICCV 2005).

[21] Park, A.-Y. and Lee, S.-W.. Gesture Spotting in Continous Whole Body Action Sequences Using Discrete Hidden Markov Models. In: Proc. Int. Gesture Workshop, 2005.

[22] Tesauro, G. Practical issues in temporal difference learning. Machine Learning 8, pp. 257 - 277, 1992.

Contact
Ilmenau Technical University
Department of Neuroinformatics and Cognitive Robotics
98684 Ilmenau
Tel.: +49 (0)3677-692858
Fax: +49 (0)3677-691665
http://www.tu-ilmenau.de/neurob

Andrea Scheidig is a Postdoc at TU Ilmenau, Department of Neuroinformatics and Cognitive Robotics. She received her Diploma degree in Computer Science in 1996 and her Doctorate degree in 2003. Her main research interests are human-robot interaction, behavior-based systems, and reinforcement learning.

Steffen Mueller is PhD-student at TU Ilmenau, Department of Neuroinformatics and Cognitive Robotics. He received his Diploma degree in Computer Science in 2005. His main research interests are robotics, learning and Multi-Agent Systems.

Horst-Michael Gross is professor of Neuroinformatics and head of the Department of Neuroinformatics and Cognitive Robotics at TU Ilmenau since 1993. He received his doctoral degree in Computer Science in 1989. Among his research interests are neural computing, robotics, and human-robot interaction.