BIRON – The Bielefeld Robot Companion

A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis,
G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer
Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
Email: ahaasch@TechFak.Uni-Bielefeld.DE

Abstract

In the recent past, service robots that are able to interact with humans in a natural way have become increasingly popular. A special kind of service robots that are designed for personal use at home are the so-called robot companions. They are expected to communicate with non-expert users in natural and intuitive way. For such natural interactions with humans the robot has to detect communication partners and focus its attention on them. Moreover, the companion has to be able to understand speech and gestures of a user and to carry out dialogs in order to get instructed, i.e., introduced to its environment. We address these problems by presenting the current state of our mobile robot BIRON, the Bielefeld Robot Companion.

Keywords: *human-robot interaction, robot companion*

1 Introduction

The development of cognitive robots serving humans as assistants or companions is currently an active research field. In order to be accepted as a communication partner by non-expert users such robot companions must exhibit a human-like communicative behavior. This raises problems related to the sensors used for observing the environment, the techniques employed for data association, and the cognitive capabilities required for multi-modal interaction with humans.

A robot companion will generally be acting in an unstructured environment, such as an office or a private home, with people roaming around. Since it is not desirable to rely on pervasive sensor technology distributed throughout the environment, the robot companion needs to carry all sensing devices on board.

The "field of view" of these sensors will, however, always be limited and their individual capabilities might not be sufficient for robustly interacting with humans. Thus, it is necessary to combine uni-modal processing results in a multi-modal data-association framework. This method increases both reliability in case of occlusions and robustness against processing errors within a single modality.

Figure 1: A typical interaction with BIRON.

At the cognitive level a robot companion needs to be able to detect humans and to be aware when a person wants to interact with the robot. For an engagement in a dialog the robot needs to focus its attention on the communication partner and maintain mutual attention throughout the dialog by showing appropriate feedback to the human. For the dialog itself the most important modality is spoken language, which can be complemented by other modalities used in natural communication.

In the envisioned scenario human communication partners can not be expected to wear special equipment, such as a close-talking microphone or data-gloves. Therefore, the multi-modal interaction acts produced by the human must be recognized with the limited sensor capabilities onboard the mobile robot platform alone. Given a semantic interpretation of those multi-modal "utterances" and a symbolic description of the observed scene, appropriate verbal or physical actions of the robot companion can be determined by employing a multi-modal interaction model and strategy.

[1]This work has been supported by the European Union within the 'Cognitive Robot Companion' (COGNIRON) project (FP6-IST-002020) and by the German Research Foundation within the Collaborative Research Center 'Situated Artificial Communicators' as well as the Graduate Programs 'Task Oriented Communication' and 'Strategies and Optimization of Behavior'.

In E. Prassler, G. Lawitzky, P. Fiorini, and M. Haegele, editors, Proc. Int. Workshop on Advances in Service Robotics, pages 27-32, Stuttgart, Germany, May 2004. Fraunhofer IRB Verlag.

In this paper we will present the current state in the development of BIRON, the Bielefeld Robot Companion which is a modified PeopleBot from ActivMedia equipped with a pan-tilt camera, a pair of microphones, and a laser range finder (for details see [4]). Our goal is to use BIRON in the so-called *home-tour* scenario. Here, the basic idea is that a human introduces to a newly purchased robot all the objects and places in a private home relevant for later interaction. Figure 1 shows a typical interaction scene where a user gains the robot's attention in order to engage in a dialog.

2 Related Work

The most advanced examples of robots realizing complex multi-modal human-robot interfaces are *SIG* [13] and *ROBITA* [12]. While only ROBITA is a truly mobile system both robots have a humanoid torso with cameras and microphones embedded in the robot's "head". Both use a combination of visual face recognition and sound source localization for the detection of a potential communication partner. SIG's focus of attention is directed towards the person currently speaking that is either approaching the robot or standing close to it. In addition to the detection of talking people, ROBITA is able to determine the addressee of spoken utterances.

There are also several complete service robot systems that integrate capabilities for human-robot interaction. For example, Care-O-bot II [7] is a multi-functional robot assistant for housekeeping and home care, designed to be used by elderly people. It receives input from the user via speech and touch screen. Although the system also produces speech output, it can not carry out natural dialogs with the user. Lino [8] serves as user interface to intelligent homes. It perceives persons by processing visual and auditory information. Since the robot operates in an intelligent environment it makes use of external information sources. The humanoid service robot *HERMES* [3] can be instructed for fetch-and-carry tasks, and it was also adopted as museum tour guide. It integrates visual, tactile, and auditory data to carry out dialogs in a natural and intuitive way, but can only interact with single persons. Jijo-2 [2] is intended to perform tasks in an office environment, such as guiding visitors or delivering messages. It uses data coming from a microphone array and a pan-tilt camera to perceive persons, but a person is only focused after it says "Hello" to the robot.

3 Overall system architecture

Since interaction with the user is the basic functionality of a robot companion, the integration of interaction components into the architecture is a crucial factor. We propose to use a special control component, the so-called *execution supervisor*, which is located centrally in the robot's architecture. We based our robot control system (depicted in Fig. 2) on a three-layer architecture [6] which consists of three components: a reactive feedback control mechanism, a reactive plan execution mechanism, and a mechanism for performing deliberative computations.

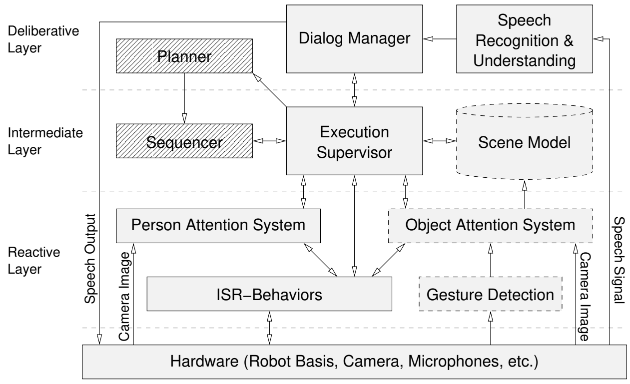

Figure 2: Overview of the BIRON architecture (implemented modules are drawn with solid lines, modules under development with dashed lines).

The execution supervisor, which is the most important component in the architecture, represents the reactive plan execution mechanism. It controls the operations of the modules responsible for deliberative computations rather than vice versa. This is contrary to most hybrid architectures where a deliberator continuously generates plans and the reactive plan execution mechanism just has to assure that a plan is executed until a new plan is received. To continuously control the overall system the execution supervisor performs only computations that take a short time relative to the rate of environmental change perceived by the reactive control mechanism.

While the execution supervisor is located in the intermediate layer of the architecture, the dialog manager is part of the deliberative layer. It is responsible for carrying out dialogs to receive instructions given by a human interaction partner. The dialog manager is capable of managing interaction problems and resolving ambiguities by consulting the user (see section 6). It receives input from the speech understanding system which is also located on the topmost layer (see section 5) and sends valid instructions to the execution supervisor.

The person attention system represents the reactive feedback control mechanism and is therefore located on the reactive layer (see section 4). However, the person attention system does not directly control the robot's hardware. This is done by the ISR software [1]. A parameterization of the attention system leads to the construction of an appropriate network of behaviors inside ISR which then controls the robot's movements.

In addition to the person attention system we are currently developing an object attention system for the reactive layer. The execution supervisor can shift control of the robot from the person attention system to the object attention system in order to focus objects referred to by the user. The object attention will be supported by a gesture detection module which recognizes deictic gestures. Combining spoken instructions and a deictic gesture allows the object attention system to control the robot and the camera in order to acquire visual information of a referenced object. This information will be sent to the scene model in the intermediate layer.

The scene model will store information about objects introduced to the robot for later interactions. This information includes attributes like position, size, and visual information of objects provided by the object attention module. Additional information given by the user is stored in the scene model as well, e.g., a phrase like "This is my coffee cup" indicates owner and use of a learned object.

The deliberative layer can be complemented by a component which integrates planning capabilites. This planner is responsible for generating plans for navigation tasks, but can be extended to provide additional planning capabilities which could be necessary for autonomous actions without the human. As the execution supervisor can only handle single commands, a sequencer on the intermediate layer is responsible for decomposing plans provided by the planner. However, in this paper we will focus on the interaction capabilities of the robot.

4 Person Attention System

A robot companion should enable users to engage in an interaction as easily as possible. For this reason the robot has to continuously keep track of all persons in its vicinity and must be able to recognize when a person starts talking to it. Therefore, both acoustic and visual data provided by the on-board sensors have to be taken into account: at first the robot needs to know which person is speaking, then it has to recognize whether the speaker is addressing the robot, i.e., looking at it. On BIRON the necessary data is acquired from a multi-modal person tracking framework which is based on *multi-modal anchoring* [5].

4.1 Multi-Modal Person Tracking

Multi-modal anchoring allows to simultaneously track multiple persons. The framework efficiently integrates data coming from different types of sensors and copes with different spatio-temporal properties of the individual modalities. Person tracking on BIRON is realized using three types of sensors:

- The laser range finder is used to detect humans' legs. Pairs of legs result in a characteristic pattern in range readings and can be easily detected. From detected legs the distance and direction of the person relative to the robot are extracted [5].
- The camera is used to recognize faces and torsos. Currently, the face detection works for faces in frontal view only [9]. A face provides information about the distance and direction of the person with respect to the robot. In addition, the height of a person can be estimated. Furthermore, the clothing of the upper body part of a person (the color of its torso) can be observed by the camera. If a torso is detected, the direction of the person relative to the robot is known [4].
- The stereo microphones are applied to locate sound sources in front of the robot. By incorporating information from the other cues robust speaker localization is possible [9].

Altogether, the combination of depth, visual, and auditory cues allows the robot to robustly track persons in its vicinity.

In a natural situation, persons are usually moving around. Since also the robot itself is mobile, users can not be expected to be located at a predetermined position. In addition, as the sensing capabilities of the robot are limited, e.g., the camera has only a limited field of view, not all persons in the vicinity of the robot can be observed with all sensors at the same time. To solve these problems an attention mechanism is required.

4.2 Attention Mechanism

The attention mechanism has to fulfill two tasks: On the one hand it has to select the person of interest from the set of observed persons. On the other hand it has to control the alignment of the sensors in order to obtain relevant information from the persons in the robot's vicinity.

The attention mechanism is realized by a finite state machine (see Fig. 3). It consists of several states of attention, which differ in the way the robot behaves, i.e., how the pan-tilt unit of the camera or the robot itself is controlled. The states can be divided into two groups representing *bottom-up attention* while searching for a communication partner and *top-down attention* during interaction.

When bottom-up attention is active, no particular person is selected as the robot's communication partner. The selection of the person of interest as well as transitions between different states of attention solely depend on information provided by the person tracking component. For selecting a person of interest, the observed persons are divided into three categories with increasing degree of relevance. The first category consists of persons that are not speaking. The second category comprises all persons that are speaking, but at the same time are either not looking at the robot or the corresponding decision is not possible,

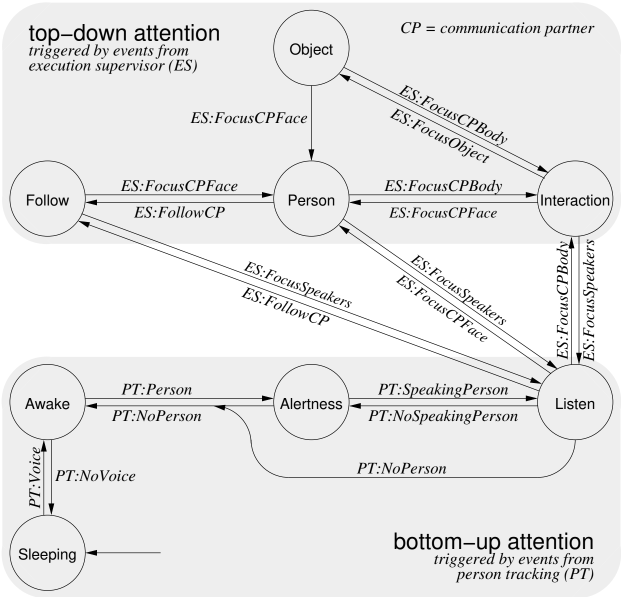

Figure 3: Finite state machine realizing the different behaviors of the person attention mechanism.

since the person is not in the field of view of the camera. Persons assigned to the third category are of most interest to the robot. These persons are speaking and at the same time are looking at the robot. In this case the robot assumes to be addressed and considers the corresponding person to be a potential communication partner. If a person is assigned to this category it is instantly selected and remains selected until the person changes to one of the other categories, e.g., by stopping talking or looking in another direction. If no person has the status of a potential communication partner, the attention mechanism always selects the person that is of most interest, e.g., persons of the second category are selected prior to persons of the first category. If the mechanism has to decide between multiple persons of the same category, it selects the one that for the longest time was not selected. In addition, the mechanism will also switch between persons in order to obtain additional information, e.g., the identities of persons present. For this purpose, a person remains selected only for a limited amount of time, after which it is temporarily blocked for selection, realizing an effect known as inhibition of return.

Top-down attention is activated as soon as the robot starts to interact with a particular person. During interaction the robot's focus of attention remains on this person even if it is not speaking. Here, in contrast to bottom-up attention, transitions between different states of attention are solely triggered by the execution supervisor. The corresponding events sent by the execution supervisor depend on the current state of the dialog.

The behavior of the robot concerning the states of the attention mechanism differs in the way the pan-tilt unit of the camera and the robot itself is controlled. Except for the two states *Sleeping* and *Object* (see Fig. 3) the camera is oriented towards the selected person, primarily towards the user's face, but also towards the torso (*Follow*) in order to robustly track the person while following, or towards the user's hands (*Interaction*) in order to be able to capture deictic gestures. When the attention mechanism is in the state *Listen* or in one of the states of top-down attention, the selected person is likely to speak to the robot. In order to obtain optimal quality of the acoustic signal the robot turns towards the person. Except for the state *Follow* the robot is not moving forward. When the attention mechanism is in the state *Object* the camera is oriented towards a predetermined position in our current implementation. Now we are developing a self-contained object attention mechanism which will replace this state.

5 Speech Recognition and Understanding

Speech is the most important modality for a multi-modal dialog. On BIRON there are two major challenges. First, speech recognition has to be performed on distant speech data recorded by the two on-board microphones. And second, speech understanding has to deal with spontaneous speech phenomena.

The recognition of distant speech with two (or more) microphones can be achieved by reconstructing a single channel representation of the speech originating from a known location on the basis of the different channels recorded by the microphones. This technique is known as beam-forming [10] and calculates a weighted average of the individual channels taking into account the estimated time delay. For recognizing distant speech we calculate this single channel reconstruction by applying beam-forming in the log-spectral domain. This method produces better results on the data recorded via the microphones on BIRON than beam-forming in the time or spectral domain.

The activation of speech recognition is controlled by the attention mechanism. Only if a tracked person is speaking and looking at the robot at the same time, speech recognition and understanding takes place. Since the position of the speaker relative to the robot is known from the person tracking component, the time delay can be estimated and taken into account for the beam-forming process. However, since noise and speech from interfering talkers standing at different positions can only be suppressed to some extent by beam-forming, the recognition quality will never reach the one obtained with a close-talking microphone.

Besides this problem of the speech recognition system the speech understanding component has to deal with spontaneous speech phenomena in dialogs between a user and the robot. For example, large pauses and incomplete utterances can occur in such task oriented and embodied communication. However, missing information in an utterance can often be acquired from the scene. For example the utterance "Look at this" and a pointing gesture to the table concludes to the meaning "Look at the table". Moreover, fast extraction of semantic information is important for achieving adequate response times.

We obtain fast and robust speech processing by combining the speech understanding component with the speech recognition system. For this purpose, we integrate a robust LR(1)-parser into the speech recognizer as proposed in [15]. Besides, we use a semantic-based grammar which is used to extract instructions and corresponding information from the speech input. A semantic interpreter forms the results of the parser into frame-based XML-structures and transfers them to the dialog manager (see section 6). Hints in the utterances about gestures are also incorporated. For our purpose, we consider co-verbal gestures only. An utterance as "This flower at the window" is transformed to the structure in Figure 4. The object attention system is intended to use this information in order to detect a specified object. Thus, this approach supports the object attention system and helps to resolve potential ambiguities.

```
<SPEECH>
  <TIMESTAMP val = "4071866790" />
  <OBJECT type = "plant" >
    <GESTURE val = "probably_pointing" />
    <TITLE name = "flower" />
    <POSITION >
      <RELATION name = "at" />
      <TITLE name = "window" />
    </POSITION >
  </OBJECT >
</SPEECH>
```

Figure 4: Representation of *"This flower at the window"*

6 Dialog Manager

The model of the dialog manager is based on a set of *finite state machines* (FSM), where each FSM represents a specific dialog. The FSMs are extended with the ability of recursive activation of other FSMs and the execution of an action in each state. Actions that can be taken in certain states are specified in the *policy* of the dialog manager. These actions include the generation of speech output and sending events like orders and requests to the execution supervisor. The dialog *strategy* is based on the so-called *slot-filling* method [14]. A slot is an information item for which a value is required. The status of a slot can be empty, filled with an attribute, or in case of a binary entry be *true* or *false*. For every FSM a set of slots is available, which are organized in a so-called dialog *frame*. Every different status combination of the slots in a frame defines a state in the corresponding FSM of the model. The task of the dialog manager is to fill enough slots to meet the current dialog goal, which is defined as a goal state in the corresponding FSM. The slots are filled with information coming from the user and other components of the robot system. This procedure can be viewed as a quantization of a user utterance into required information items.

The dialog management is event-based, where switching between the dialog states is not done by following a transition in the model, but depends on the status composition of all slots in the dialog frame. Several input events like user utterances or information from other components of the robot system change the status of the slots. In an ongoing dialog, the dialog manager compares the slots in the newly updated dialog frame with those in the FSM to find the model's new current state. Thereby, slots are compared only by their status and not by their content. After executing an action, which is determined by a lookup in the dialog policy, the dialog manager waits for new input from the execution supervisor or the speech understanding system.

The slot-filling technique alone is not powerful enough to support the complex interaction scenarios in robot domains [11]. The user intentions are not predictable in such cases. To overcome this limitation, we designed the dialog in a modular way and divided each dialog into a set of sub-dialogs. Each sub-dialog is responsible for a task and is modeled as a separate FSM. This FSM has a goal state which indicates the completion of the current task. The processing of each sub-dialog can be interrupted by another sub-dialog, which makes alternating instruction processing possible. The dialogs are specified using a declarative definition language and encoded in XML in a modular way. This increases the portability of the dialog manager and allows an easier configuration and extension of the defined dialogs.

7 Interaction Capabilities

In the following we describe the interaction capabilities BIRON offers to the user in our current implementation. Initially, the robot observes its environment. If persons are present in the robot's vicinity, it focuses on the most interesting one (cf. section 4). A user can start an interaction by greeting the robot with, e.g., "Hello BIRON". Then, the robot keeps this user in its focus and can not be distracted by other persons talking. Next, the user can ask the robot to follow him to another place in order to introduce it to new objects. While the robot follows a person it tries to maintain a constant distance to the user and informs the person if it moves too fast. When the robot reaches a desired position the user can instruct it to stop. Then, the user can ask the robot to learn new objects. In this case the camera is lowered to also get the hands of the user in the field of view.

When the user points to a position and gives spoken information like "This is my favorite cup", the object attention system is activated in order to center the referred object. If the user says "Good-bye" to the robot or simply leaves, the robot assumes that the current interaction is completed and looks around for new potential communication partners.

8 Summary

In this paper we presented an overview of the robot companion BIRON whose target application is the home-tour scenario. Its natural interaction capabilities are based on a person attention system, a speech recognition and understanding component, and a dialog manager. These components are integrated in a hybrid architecture which is controlled by a central execution supervisor. The architecture's modular design easily allows modifications on the robot companion's skills by replacing and adding new components. Current work focuses on switching to a powerful communication framework [16] and integrating an object attention system for associating gestures with visual features of objects.

References

[1] M. Andersson, A. Orebäck, M. Lindstrom, and H. I. Christensen. ISR: An intelligent service robot. In H. I. Christensen, H. Bunke, and H. Noltmeier, editors, *Sensor Based Intelligent Robots; International Workshop Dagstuhl Castle, Germany, September/October 1998, Selected Papers*, volume 1724 of *Lecture Notes in Computer Science*, pages 287–310. Springer, New York, 1999.

[2] H. Asoh, Y. Motomura, F. Asano, I. Hara, S. Hayamizu, K. Itou, T. Kurita, T. Matsui, N. Vlassis, R. Bunschoten, and B. Kröse. Jijo-2: An office robot that communicates and learns. *IEEE Intelligent Systems*, 16(5):46–55, 2001.

[3] R. Bischoff and V. Graefe. Demonstrating the humanoid robot *HERMES* at an exhibition: A long-term dependability test. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems; Workshop on Robots at Exhibitions*, Lausanne, Switzerland, 2002.

[4] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer. Audiovisual person tracking with a mobile robot. In *Proc. Int. Conf. on Intelligent Autonomous Systems*, pages 898–906. IOS Press, 2004.

[5] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.

[6] E. Gat. On three-layer architectures. In D. Kortenkamp, R. P. Bonasso, and R. Murphy, editors, *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, chapter 8, pages 195–210. MIT Press, Cambridge, MA, 1998.

[7] B. Graf, M. Hans, and R. D. Schraft. Care-O-bot II—Development of a next generation robotic home assistant. *Autonomous Robots*, 16(2):193–205, 2004.

[8] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, and E. Demeester. Lino, the user-interface robot. In *European Symposium on Ambient Intelligence (EUSAI)*, pages 264–274, 2003.

[9] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*, pages 28–35. ACM, 2003.

[10] S. J. Leese. Microphone arrays. In G. M. Davis, editor, *Noise Reduction in Speech Applications*, pages 179–197. CRC Press, Boca Raton, London, New York, Washington D.C., 2002.

[11] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters. The WITAS multi-modal dialogue system I. In *Proc. European Conf. on Speech Communication and Technology*, pages 1559–1562, Aalborg, Denmark, 2001.

[12] Y. Matsusaka, T. Tojo, and T. Kobayashi. Conversation robot participating in group conversation. *IEICE Trans. on Information and System*, E86-D(1):26–36, 2003.

[13] H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Cairns, Australia, 2002. Lecture Notes in Artificial Intelligence, Springer.

[14] B. Souvignier, A.Kellner, B. Rueber, H. Schramm, and F. Seide. The thoughtful elephant: Strategies for spoken dialog systems. In *IEEE Trans. on Speech and Audio Processing*, volume 8, pages 51–62, 2000.

[15] S. Wachsmuth, G. A. Fink, and G. Sagerer. Integration of parsing and incremental speech recognition. In *Proc. European Conf. on Signal Processing*, volume 1, pages 371–375, Rhodes, 1998.

[16] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer. An XML based framework for cognitive vision architectures. In *Proc. Int. Conf. on Pattern Recognition*, Cambridge, UK, 2004. to appear.