In **Interspeech'2005 — 9th European Conf. on Speech Communication and Technology**, Lisboa, 2005.

Modality Integration and Dialog Management for a Robotic Assistant

Ioannis Toptsis, Axel Haasch, Sonja Hüwel, Jannik Fritsch, Gernot A. Fink

Faculty of Technology, Bielefeld University
33594 Bielefeld, Germany
{itoptsis,ahaasch,shuewel,jannik,gernot}@techfak.uni-bielefeld.de

Abstract

The communication with robotic assistants or companions is a challenging new domain for the use of dialog systems. In contrast to classical spoken language interfaces users interact with mobile robots mostly in a multi-modal way. In this paper we will present the integration of several modalities in the dialog system of BIRON – the Bielefeld Robot Companion. Besides speech as the main modality the system integrates deictic gestures and visual scene information in order to resolve object references in a task oriented dialog. We will present example interactions with BIRON and first qualitative results from the "home-tour" scenario defined within the COGNIRON project.

1. Introduction

The development of mobile robots serving humans as assistants or companions is a challenging research field. In order to be accepted as a communication partner by naive users such robot companions must support a natural interaction with them. Their acceptance in a private household environment depends on the exhibition of a basic social behavior.

Experiments in household domains have shown that humans are using multiple modalities in the communication with robots, which is a natural concept for human communications. The naturalness of human-machine interaction can be increased with support of spoken language as the main modality and additional natural modalities like, e.g., gestures and mimics. In these domains the dialog manager has to deal with incomplete information from the speech side. The missing information can be obtained from the other modalities and helps to complete or disambiguate the information provided via spoken language.

In contrast to telephony-based services – currently the preferred application area of dialog systems – the mobile robot domain is much more challenging. The dialog management complexity is not only increased by the use of additional modalities but also by the dynamic environment. Both the robot and the user do not have a fixed spatial position but move around in an unpredictable way. Furthermore, the position of the objects in the scene is not static. Consequently, the robot's behavior depends not only on the communication with the user but also on the complex interaction of the mobile platform and its environment. Furthermore, the necessity to consider additional modalities beside speech increases the complexity of the dialog system and especially of the dialog manager.

In this paper we will present the integration of multi-modal information in the dialog system of BIRON – the Bielefeld Robot Companion [1]. The system is connected by several perceptive channels to the environment, namely visual and acoustic sensors to acquire visual object information from the scene, deictic gestures and speech data. The sensory input has to be interpreted in an integrated way. In our work we propose an integration of these modalities on several levels. In the lower level object recognition is fused with a recognition module for deictic gestures to a so-called *object attention system*, in order to focus the robot's attention on certain objects. On the higher level the dialog manager interacts with the object attention module via semantic interpretation structures. The object attention module tries to match the sematic representation with current objects in the environment.

In the following section we will give a short overview about related work on dialog systems used for the interaction with mobile robots. Then we will describe the architecture of the proposed system and the modality fusion in the object attention module. In section 4 the focus will be on how the dialog manager deals with information from the different modalities in order to control the dialog and carry out the task. Then we will outline the capabilities of the current dialog model and present first results of interactions with our mobile robot BIRON.

2. Related Work

Most of the dialog systems today and particularly those in commercial applications only handle speech input. This enables rather natural communication, because spoken language is the most important modality in human-human interaction. However, mobile robot systems need advanced capabilities of dialog management. Only a limited number of dialog systems supports multi-modal interaction with mobile robots. Such a system is on *Hygeiorobot*, a mobile robotic assistant for hospital use [2]. It is able to carry out several tasks like delivering medicine or messages and answering questions about patients. Its dialog system supports only speech-based interaction and is designed for relatively short dialogs only.

A mobile robot which can handle multi-modal input is CARL [3]. It is a service robot with navigation abilities which can greet guests in a reception or serve food to them. Besides spoken language it also accepts as input pointing gestures on a touch screen and supports through this simulation of deictic gestures a multi-modal interaction. Its dialog system is built modular with three stages, speech and linguistic processing and dialog control. The user input is interpreted through high-level reasoning and the interaction with the user is modeled as a message exchange. The output is also multi-modal, because an animated face is used with labial movements synchronized to spoken output. The stationary robot *Leonardo* [4] is capable of detecting the interaction of a human with colored buttons arranged around it. Leonardo recognizes deictic gestures combined with

*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 (DFG) within the Graduate Program 'Task Oriented Communication'.

speech. It is possible to label buttons by giving verbal information and to teach the robot how to use these buttons. The verbal abilities of Leonardo are limited to the input side. It can carry out tasks like pushing a button without spoken feedback.

3. System Architecture

The dialog system developed for BIRON is designed in a modular way with a separation of the dialog management from speech processing which enhances the portability and extensibility of the system. Additional modules process the different input modalities. For the communication between the modules we use a specially developed XML communication framework.

3.1. Speech Understanding

The speech understanding component has to deal with spontaneous speech phenomena. 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 by scene information, e.g., by resolving gestures or by object detection.

To obtain fast and robust speech processing, we combine 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 [5]. Furthermore, we use a grammar based on semantically defined constituencies which represent the relevant data for the robot interaction such as information about instructions. A semantic interpreter forms the results of the parser into frame based XML-structures and transfers them to the dialog manager. Hints in the utterances about gestures are also incorporated. We assume that people in this context only use co-verbal gestures. The object attention system uses this information in order to detect a specified object. Thus, this approach supports the object attention system and helps to resolve potential ambiguities.

3.2. Object Attention System

Opposed to many other approaches using a visual attention system in static scenes, we use a mobile robot having a camera with a limited field of view. Therefore, we need some kind of sensor control to reorient the camera to a relevant part of the current scene. For instance, in the context of a home environment objects to which a user refers to are considered important. However, private homes contain a tremendous variety of different objects. Determining the object the user refers to requires to recognize known objects and, more importantly, to interactively learn unknown objects. In order to accomplish this task, the object attention system (OAS) processes input from the camera, the dialog manager, and the gesture recognition module. Due to limited space we present a more detailed description of the OAS in [6]. The coordination of the input modalities and the control of the hardware components (e.g., the pan-tilt camera) is realized by a finite state machine (FSM). This FSM controls the processing of the OAS (see Fig. 1).

In order to represent the objects in the environment of the robot, a knowledge base called *scene model* is used. This scene model is an active memory [7] and can be considered as a mixture of short-term spatial memory and long-term object memory. It stores not only the current position of an object, but also the object's properties that are extracted from the scene (e.g., its visual appearance) and given verbally by the user (e.g., the attribute "blue").

The processing of the OAS is started when the dialog manager sends a request to identify an object in the scene that has been referenced by the user. As a simple example, a lexical cue like "this" or "that" triggers the speech understanding component to determine that a gesture is expected. As a consequence, the dialog manager provides the OAS with corresponding information and the gesture recognition module [6] is activated by the FSM. After finding a pointing gesture, the camera is reoriented based on the hand coordinates and the pointing direction that are provided by the gesture recognition module. This information is also used to restrict the region of interest (ROI) in the camera image that needs to be searched for the object referenced by the user. If the dialog manager sends the description of the object (e.g., type, color, owner), an inquiry to the scene model is initiated. The object is considered as known if the scene model already contains an object with this description, otherwise the object is unknown. The two different processing strategies for learning and localizing known and unknown objects in the robot's environment are described in the following two sections using the example "This is Gernot's blue cup".

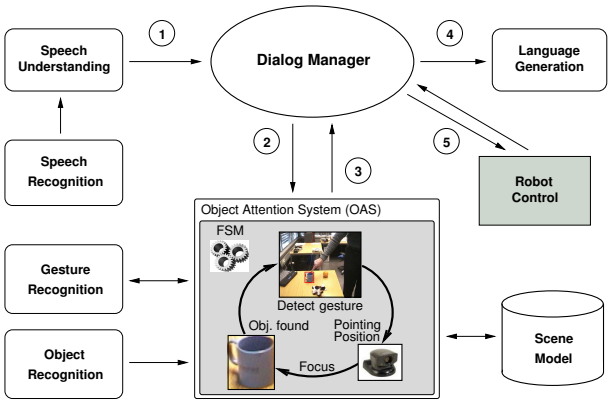

Figure 1: Overview of the BIRON dialog system architecture

3.3. Recognizing known objects

The search for a known object type involves an object detection process. For this task we use an object recognizer based on the visual appearance of objects. However, as we focus on the interactive object learning and not on the issue of object recognition, the Normalized Cross-Correlation (NCC) [8] is sufficient here.

In the case when the scene model contains a known object with properties matching the verbally referenced object (e.g., color=blue, type=cup, owner=Gernot), the OAS fetches appropriate image patterns for recognizing the cup from the scene model and the object detection is initiated. If the object is found within the ROI, a success message is sent to the dialog manager and the obtained position of the object is stored in the scene model. If two or more objects of the same type are detected in the camera image, a message is send to the dialog manager to initiate a clarification sub-dialog in order to decide which of the objects is referenced by the user.

3.4. Finding unknown objects

For detecting and learning unknown objects we assume that an accompanying pointing gesture is present and a query to the scene model is not successful, i.e., there is no object with matching properties stored in the scene model. In this case, the OAS uses several different predefined color filters that are similar to

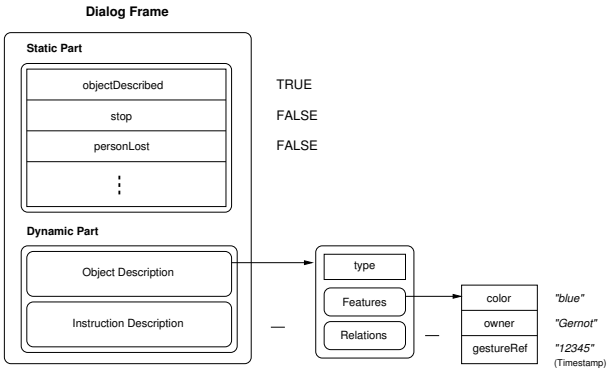

Figure 2: Dialog frame including information of the utterance "This is Gernot's blue cup"

the attention maps described in [9]. Based on the additional verbal information given by the user (e.g., "blue", see Fig.2) an appropriate color filter is selected. Subsequently, a bounding box in front of the hand position is set around the image area that contains the color which is supported by the corresponding attention map. This bounding box is used to extract a view of the blue cup from the camera image.

4. Dialog Manager

The model of the dialog manager is based on a set of *Finite State Machines* (FSM), where each FSM represents a specific sub-dialog [10]. The overall dialog is divided into sub-dialogs, where each sub-dialog corresponds to a task. Thus we overcome the limitations of state-based approaches in robot domains. The dialog strategy is based on the so-called *slot-filling* method [11]. A slot is an information item and 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 in the corresponding FSM. Each state of the dialog model is determined by the status of its slots. Incoming information from every channel, not only from speech processing components, can change the status of the slots.

The slots are grouped in a structure called *dialog frame*. In an ongoing dialog, the dialog manager compares the newly updated slot status configuration with those in the FSM in order to find out the current dialog state. In relatively simple applications, like information inquiry systems, all the necessary information to achieve the dialog goal can be stored in the dialog frame. But in the service robot domain the user can describe the objects in several ways, e.g. through *features* like color and size or *relations* to other objects in the scene. There could be several types of objects each of them with different features which results in a variable number of required slots. The number and the type of relations to other objects are also not predictable. In order to handle such tasks we divided the dialog frame in a static and a dynamic part, as shown in Fig. 2.

The *static* part has the function to acquire information which controls the dialog flow. In the case of information inquiry application it is sufficient to model the dialog. All the incoming information from users utterances will be stored there and used for further interaction. In service robot applications it is necessary to use also the *dynamic* part of the dialog frame. The object description, including features and relations to other objects, can be stored in this dynamic tree-structure. As the dialog flow control based on the static part, it is required to define slots to represent the existence of information in the dynamic part. For example, a slot objDescribed is declared in the static part and switched from status *false* to *true* if an spoken object description is received.

In addition to the modality fusion of recognized gestures and objects in the object attention system, an integration of the modalities is carried out at a higher level. The dialog manager receives semantic interpretations of user utterances from the speech understanding module (1), as shown in Fig. 1. During the dialog, the dialog manager collects information to build a dynamic object description including object features and relations. If the user references an object by a pointing gesture and in combination with a phrase like "*this* is a blue cup", the dialog manager inserts a specific slot into the object description (*gestureRef* in Fig. 2). This description will be sent as an object request to the object attention system (2). After the internal processing the object attention system will respond with an object list, which includes ideally one object with its ID, or more objects in case of ambiguity (3). The object representations are enriched with additional properties from the scene. Then the dialog manager tries to confirm or clarify the object interacting with the user (4). When the appropriate object is confirmed, the dialog manager sends its ID and further information about its usage to the robot control (5). If the user issues an instruction relating to the object, the instruction and its properties are sent in combination with the ID to the robot control for execution.

Summarizing, the whole process of higher level modality integration characterized by three steps: receiving instructions and object descriptions (including indication for expected gestures) from speech modality, sending an object request based on this description to the object attention system, receiving an object description including an ID and additional information from scene, confirming/clarifying by the user through the dialog, transmitting the object reference where necessary combined with an instruction description to the robot control.

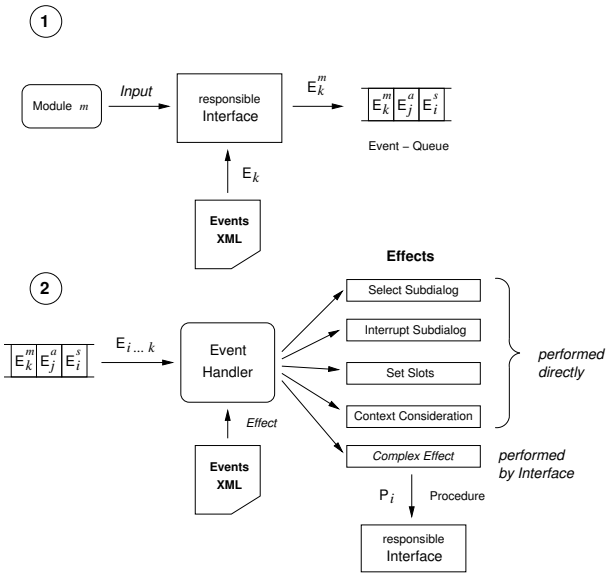

Figure 3: Mapping of input to events and their effects

The dialog manager is not only responsive to speech input, but also to other modalities and messages from the robot control system. The mapping of all kinds of input data to *events* enables such a behavior. A set of domain dependent events and their *ef-*

fects is defined. The input from every channel is mapped into appropriate events by the responsible interface which are put in an event queue (1), as shown in Fig. 3. The queued events are processed by an event handler (2). The majority of the events has effects which are defined in a declarative way and can be processed by the event handler directly. Few of them have to be processed by a procedure in the responsible interface. Our goal is to minimize such hard-coded effects and allow a flexible design of multi-modal input processing. The execution of the actions defined in the effect declarations changes the dialog frame configuration, which triggers further dialog steps. The actions in each dialog step are not limited to spoken output, but can be also commands to robot control or requests to the OAS. Thus the multi-modal support is also extended to the output side.

5. Domain and Results

Within the COGNIRON project we are implementing the *home tour* experiment. In this scenario a robot is delivered at home and the user familiarizes it with the new environment by showing it several rooms and objects. During the home tour the robot should build internal representations of the environment and objects, in order to carry out some tasks with them at a later date. In the following we present a typical dialog of this scenario. (U: User, R: Robot, DM: dialog manager, OA: object attention)

U: "Hello BIRON!"

R: "Hello, how can I help you?"

U: "Please follow me, BIRON."

R: "OK, I follow you."

U: "I'll show you some objects."

R: "OK, I'm ready."

U: "This is a pencil." + *deictic gesture*

DM→OA: object(type(pencil), gestureIndication)

OA→DM: objList(object(ID, type(pencil), color(green)))

R: "OK, I've seen the pencil."

U: "This is Gernot's cup." + *deictic gesture*

DM→OA: object(type(cup),owner(Gernot),gestureIndication

OA→DM: objList(object(ID, type(cup), color(red)),
object(ID, type(cup), color(blue)))

R: "Do you mean the red or the blue cup?"

U: "The blue one."

R: "OK, I've seen Gernot's cup."

U: "Thank you BIRON. That's it."

R: "Goodbye!"

At the *IST Event 2004* exhibition we proved the robustness of the dialog system in a real environment. During three days BIRON has been successfully interacting with many persons (11 dialogs recorded), which were non-experts in this application area. The flexibility of the dialog manager was also proved by some extentions to the dialog model which were necessary during the preparations for this event. Furthermore, we conduct an experiment with 12 naive users interacting with BIRON. The task was to guide the robot through a room with several objects and to show them by the use of deictic gestures combined with spoken naming of the objects. The results of the experiment shows that the users found the interaction with the robot natural (83%) and easy (92%).

6. Conclusion

In this paper we presented the integration of several modalities in the dialog system of BIRON. Speech is the main modality used for the interaction but in robot domains and especially in our home tour scenario, speech includes only partial information. Therefore, we consider further modalities like deictic gestures and visual scene information to resolve object references. The integration of the modalities is divided in two parts, a lower level integration of gesture and object recognition results in the object attention module and a higher level integration with the transmission of semantic object representations from speech side to the object attention system. The response to the dialog manager, a semantic object reference, is used in the further dialog or is associated with a given instruction.

7. References

[1] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON – The Bielefeld Robot Companion," in *Proc. Int. Workshop on Advances in Service Robotics*, Stuttgart, Germany, 2004, pp. 27–32.

[2] D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proc. European Workshop on Service and Humanoid Robots*, Santorini, Greece, 2001.

[3] L. S. Lopes, A. Teixeira, M. Rodrigues, D. Gomes, C. Teixeira, L. Ferreira, P. Soares, J. Girão, and N. Sénica, "Towards a personal robot with language interface," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[4] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg, "Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots," *Artificial Life*, 2004.

[5] S. Wachsmuth, G. A. Fink, and G. Sagerer, "Integration of parsing and incremental speech recognition," in *Proc. European Signal Processing Conference*, vol. 1, Rhodes, Greece, 1998, pp. 371–375.

[6] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer, "A multi-modal object attention system for a mobile robot," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, 2005.

[7] S. Wrede, M. Hanheide, C. Bauckhage, and G. Sagerer, "An Active Memory as a Model for Information Fusion," in *Proc. Int. Conf. on Information Fusion*, vol. 1, Stockholm, Sweden, 2004, pp. 198–205.

[8] J. P. Lewis, "Fast template matching," in *Proc. Conf. on Vision Interface*, Quebec, Canada, 1995, pp. 120–123.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] I. Toptsis, S. Li, B. Wrede, and G. A. Fink, "A multi-modal dialog system for a mobile robot," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 1, Jeju, Korea, 2004, pp. 273–276.

[11] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant – Strategies for spoken dialog systems," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 1, pp. 51–62, 2000.