Interaction Awareness for Joint Environment Exploration

Thorsten Spexard\textsuperscript{1}, Shuyin Li\textsuperscript{2}, Britta Wrede\textsuperscript{3}, Marc Hanheide\textsuperscript{4}, Elin A. Topp\textsuperscript{5}, Helge Hüttenrauch\textsuperscript{6}

\textsuperscript{1,2,3,4}Faculty of Technology, Applied Computer Science, Bielefeld University, 33594 Bielefeld, Germany
email:\{tspexard, shuyinli, bwrede, mhanheid\}@techfak.uni-bielefeld.de
\textsuperscript{5,6}School of Computer Science and Communication, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden
email:\{topp, hehu\}@csc.kth.se

Abstract— An important goal for research on service robots is the cooperation of a human and a robot as team. A service robot in a domestic environment needs to build a representation of its future workspace that corresponds to the human user’s understanding of these surroundings. But it also needs to apply this model about the “where” and “what” in its current interaction to allow communication about objects and places in a human-adequate way. In this paper we present the integration of a hierarchical robotic mapping system into an interactive framework controlled by a dialog system. The goal is to use interactively acquired environment models to implement a robot with interaction aware behaviors. A major contribution of this work is a three-level hierarchy of spatial representation affecting three different communication dimensions. This hierarchy is consequently applied in the design of the grounding-based dialog, laser-based topological mapping, and an objects attention system. We demonstrate the benefits of this integration for learning and tour guiding in a human-comprehensible interaction between a robot and its user in a home-tour scenario. The enhanced interaction capabilities are crucial for developing a new generation of robots that will be accepted not only as service robots but also as robot companions.

I. Introduction

With the progress in research and product development humans and robots get more and more close to each other. Specialized service robots (e.g., vacuum cleaners) have already entered people’s homes and are used on a regular basis. We suggest a more general view on service robots by working with the idea of a robot companion, a service robot that is assigned to a particular working environment and has a limited number of users. This is based on the idea of the “Cogniron – The cognitive robot companion” [1] project which builds the framework of the work presented here.

In order for such a robot to be accepted by human users who share their physical environment with the robot, it needs to behave in a socially acceptable way. However, the term ‘socially acceptable’ refers to a rather broad range of capabilities and addresses virtually all capabilities of the robot. One of the competences a socially acceptable robot should have, has been identified by Dautenhahn [2] as Interaction Awareness. This term is defined as the state of an agent who is “able to perceive important structural and/or dynamic aspects of an interaction that it observes or that it is itself engaged in”. However, in order for a robot to become interaction aware, it needs to be aware of its physical environment and whereabouts in order to be able to engage in communications about it. We call this spatial awareness and define it as the capability of the robot to perceive and understand its spatial environment and to draw inferences for its own behavior from this knowledge.

Spatial awareness plays a role in such different activities as task execution, verbal interaction or non-verbal interaction. For example, the multi-modal interaction between humans takes differences between large and small objects into account. When referencing a small object, a human will likely point to it or take it in his hands whereas when showing a large object such as a big furniture, a human communication partner will only face the object but less likely point to it. When referencing a room, no deictic gesture at all will be expected because the interlocutors are likely to already be inside the room.

We therefore propose an approach to spatial representation that provides the possibility to consider such information. While standard approaches use metric representations as a basis for localization and navigation functions of robots we extend such a representation by a human oriented representation that is gained during a human-augmented mapping process which has already been proposed in previous work [3] to obtain a shared environment representation for a service robot and its user. The scenario tackled by our approaches is the “robot home tour” in which a user can ask a robot to follow her to present the home environment in a guided tour. Hence, the scenario allows to give information to the robot that is considered important by its user during this tour.

Based on this scenario, our aim is to develop a representation that facilitates robot navigation on the one hand and communication with the human user on the other hand. As will be shown in this paper, our analyses indicate that a representation is needed that contains (1) a deictic dimension in order to facilitate (complementary non-verbal) communication, (2) a semantic dimension with a spatial ontology in order to integrate world-knowledge in the interaction and (3) a topological dimension as a basis for the physical movements of the robot.

This paper presents our effort to join the robotic mapping approach through a hierarchical environment model and an attentional hierarchy with a multi-modal dialog and
interaction control system. Joining these efforts converges into a robot unveiling situation-aware behavior with a focus on spatial representations. We show the link from the concepts and the handling of different situations that can occur during a “guided tour” scenario to the dialog and propose a communicational framework for the actual integration into a working system on software level. The environment modeling and the robotic mapping subsystem are thus used to provide the overall system with the ability to localize in a (for humans) topological sense and thus communicate its whereabouts appropriately.

The remainder of this paper is organized as follows. In section II we motivate our hierarchy from similar investigations in Cognitive Science and Robotics and give examples for how hierarchy can manifest itself in human-robot interaction. Section III explains how this hierarchical model is implemented for the robotic mapping process and in section IV we present the dialog model and its link to the environment model. With section V we present the integrated robotic system and outline experimental runs performed in a real world apartment. Conclusions are drawn in section VI where we also give some ideas for future work.

II. Hierarchical Environment and Map Representation

In order to model spatial awareness in a robot we need to augment a robotic map with information given by the user, while communication is facilitated with the help of an adequate generic environment model. This model needs to provide appropriate concepts (or categories) to be filled with the information given by the user. Since human encoded knowledge of spatial relations and layouts is most likely represented in a partial hierarchy [4], we suggest to use a hierarchical approach to model the link between robotic mapping and human adequate communication. Other hierarchical approaches to environment representations have been used to model exploration strategies [5] or way-finding routines for robots [6], mostly in the context of autonomous systems.

Based upon findings on human representation of spatial knowledge [4] that indicate a hierarchical representation on the one hand and observations on how users reference different spatial entities during an interaction on the other hand we use the three dimensions interaction, topology and semantics to define a spatial hierarchy.

A. Interaction related hierarchy

When interacting with the robot users tend to use different communicative means to refer to objects, areas, or rooms. While the verbal specification does not vary much (i.e. “This is X”, “and here is X” etc.) the non-verbal deixis is very different in the three cases. An object is shown by pointing to it or taking it by hand. Hence, in order to resolve the reference of the deictic gesture, an image based analysis needs to be performed. Thus, the representation of objects should also be based on image data.

In contrast, when showing an area such as “the dining table” it can be observed that the user will not use very specific deictic gestures but rather referencing the area by facing it. Therefore a reference to an area may best be detected by image data, supported by a wide range sensor like a laser range finder. Similarly, the representation of areas should also be based on these modalities.

When showing a room, no information about a specific direction is necessary as the robot can assume that it is already inside the room. Thus, in order to be able to classify rooms the system needs to represent rooms differently. Generally, rooms tend to exhibit a specific spatial layout which can best be represented with a laser range finder. Again, this should be reflected in the representation, in order to have correct representations of rooms as well as the references to them.

B. Topology related hierarchy

Based on this interaction-inspired hierarchy we define the basic three entities as follows:

A room is a part of an indoor environment. In most cases this corresponds directly to what is commonly referred to as a “room”, clearly delimited, e.g., by walls. With an area we denote a specific part - but not geometrical point - of a room like the space around the sink at the kitchen where, e.g., tasks like dishwashing can be performed by the robot. Areas describe in consequence a certain workspace for the service robot to allow tasks such as “Pick up my glasses from the table, please”. For the robot this task can be translated to “going to a suitable position close to the table” (into an area close to the table) and there determine the best position and orientation to pick up the glasses as requested.

We understand an object as an item without a fixed location, i.e., it can easily be moved (by either the robot and the human or both) and thus needs to be expected to be in different places over time. Objects are furthermore taught to a robot by users to allow their incorporation into service tasks. Tasks such as finding, carrying, transporting, manipulating often have such a target-object for the robot’s manipulation capabilities. We do not represent objects directly in the basic mapping due to their attribute of a changing location. Instead we incorporate objects in a separate Object Attention System [7] to handle their perceptual detection, reasoning as well as the referencing in command and control exchanges with users. This information is combined with the representations of rooms and areas in a so called environment model, which is a component of the overall robotic localization and mapping system.

C. Semantic related hierarchy

In order to couple the interaction related hierarchy with the interaction information we need to introduce semantic world-knowledge in a linguistic lexicon. We do this by using a spatial ontology consisting of the three layers room, area and object. The goal is to correctly analyze the semantic content of verbal utterances in the form of “This is X” with co-occurring deixis. Therefore, each verbal object-label that can be specified via the speech-based dialog is assigned to one of these layers. For example, the label ‘kitchen’ will be associated with the hierarchy level ‘room’ whereas ‘cup’ will be associated with ‘object’.
III. Acquiring a Topological Representation

The robotic localization and mapping system relies on different, hierarchically organized components. Fig. 1 gives a schematic overview of this mapping subsystem. Based on a combination of metric and topological methods the system is able to report its whereabouts in a human-comprehensive way like “beside the cupboard at the living room” instead of a pair of coordinates. Therefore concepts and topological information are used while exact navigation can still be enabled through an underlying metric mapping system. For this paper the most relevant level of this mapping hierarchy is the highest level, the environment model (the white rectangle in the figure), as described in the previous section. A central question for the topological/conceptual localization the subsystem has to provide is how to map the robotic sensory data into the representation based on concepts. This is also reflected in the “Region segmentation” in Fig. 1, which allows to decide, in which region (room) the robot is currently located. In the following three approaches to answer the representation question are given.

A. Representing Rooms

In recent work [8] we propose a laser range data feature based description of rooms (regions) that can be used for the segmentation of the environment into the necessary topological nodes (guided by the user’s information to form the nodes corresponding to the spatial concepts). Additionally the feature based description allows to facilitate the localization in terms of rooms to a certain degree; in cases the underlying metric localization process becomes uncertain , the feature description can be used to generate hypotheses for localization in the topological sense. This approach is similar to a method for learning of places used by Mozos et al. [9].

The features used are the results of a statistical analysis of a laser range data set taken at a point specified by the user. Using a small number of features (the length and width of the data set in the directions of the principal component vectors as well as their ratio to each other and the covered area (mass) of the set) allows to keep a very concise description for relevant rooms.

B. Representing Areas

The chosen option to represent areas with the help of available sensor readings is to use a “view” representation of the important part of the room, taken from a pose from which the robot observed the area. A challenge is that laser range data sometimes is insufficient to create a unique representation for each area. Thus, the pose (position and orientation) relative to the surrounding room is also taken into consideration to get an estimate of “where to go”, but the exact position can be reached or described within a certain margin. Another option to represent such a view is of course to use images of the particular observed scene and use a similarity measure of the appearance to determine the border of an area [10]. So far we concentrate on the combination of laser range data and pose representation. In the rare case that both cues, laser range data and pose, fail the Dialog is used to ask a human in the robots surrounding for assistance, learning information about its current position to avoid this situation next time.

C. Representing Objects

An object representation created by the Object Attention System incorporates the object location relative to the robot, visual information, auditory information, and descriptive information from the user’s utterance. The object location is estimated by deictic gesture recognition [7], which computes referenced positions as cylindrical coordinates relative to the robot. This results the robot to align its camera to the specified coordinates taking an image and also recording the current sound. The descriptive information is taken from the user utterance like the object color “blue” from “This is a blue cup”. As no objects are defined or trained previously due to the varying appearance, this description is the only information used to detect the object in the taken image, cutting the region out, and finally storing it. Optional descriptions in size (huge, small, ...) and shape (long, round, ...) are also taken into account when an utterance is processed by the integrated Dialog System.

IV. Dialog System

In general a dialog system is responsible for carrying out interactions with the user including transferring user commands to the robot control system and reporting task execution results to the user. During the conversation a dialog system should be able to regulate the initiative distribution, handle miscommunication, draw inferences between dialog participants’ contributions and organize and maintain the discourse. To enable these abilities we implemented a powerful dialog model [11] for BIRON based on grounding [12].

We represent dialog participants’ contributions as exchanges, i.e., pairs of contributions. They achieve the state ‘grounded’ only if the acceptance of the presentation is available which depends on the communication success (e.g., if the speech input is clearly understood) and the robot task execution status. These exchanges are organized in a stack which represents the ungrounded discourse up to the current state. The grounding status of the whole stack is dependent on the status of the individual exchanges and the relations between them. We introduced 4 types of such relations (default, support, correct and delete) and they can also have local effects on their previous exchanges. According to the execution results of the robot control system the dialog...
system (DLG) formulates contributions for the robot. Each contribution of both the user and the robot is categorized in terms of its roles, i.e., if it initiates an exchange of a certain relation to the previous exchange or if it is the acceptance of an existing one. According to this role, either a new exchange is pushed onto the stack or an old one (or a group of old ones) is popped because it reaches the status ‘grounded’. All the popped exchanges are collected into a vector which records the complete dialog history. We thus model the grounding process using an augmented push-down automaton which exhibits local flexibility in contrast to conventional approaches ([13], [14]). The implemented system enables a mixed-initiative dialog style and a well-organized discourse maintaining mechanism. This system could facilitate smooth conversation concerning localization tasks in HRI by flexibly handling complex clarification questions based on the messages that will be sent by the Topological Localization.

A. Grounding based spatial mapping

In the following we explain how the dialog system works with the example of a localization-related task (as shown in Fig. 2(a)). The user starts by naming the location (U1) but is not understood possibly due to speech recognition problems. This presentation creates a new exchange Ex 1. Without the need to consult the rest of the robot system, the dialog system immediately starts a clarification question (R1), i.e., it creates a new exchange Ex 2 with the grounding relation ‘support’ to Ex 1. When the user answers the robot’s question (U2) Ex 2 is popped from the stack as it is now grounded. Since Ex 2 has a support relation to the previous not-understood exchange Ex 1, Ex 1 is updated with the newly collected information that the user names the location with ‘kitchen’.

The dialog system then tries to provide acceptance for Ex 1 by sending the command SetLocName with the parameter ‘kitchen’ to the Localization. Once the dialog system receives a positive result about the successful operation SetLocName: kitchen the status of Ex 1 is changed to grounded and Ex 1 is popped from the stack with an acceptance being issued to the user (R2). The stack, i.e., the currently ungrounded discourse, is now empty.

After the map building process, the robot is able to answer questions like “Where are you?” (U3) which creates a new exchange Ex 3. To provide the acceptance for this exchange, the dialog system sends the command GetLocName to the Localization which then successfully delivers the name of the location ‘kitchen’. Thus, the dialog system can ground the current Ex 3 and pop it from the stack while informing the user about the current location (R3).

In the same way the Object Attention System is linked to the dialog, which uses the integrated semantic world-knowledge (see II-C) to distinguish between the three layers room, area, and object.

B. Grounding spatial inconsistencies

However, experience with user studies shows that this process of human augmented mapping happens rarely without ambiguities. For example, different people may name rooms differently or a certain kind of area in a room, e.g. a “cupboard”-area, may occur in different rooms. For the first kind of ambiguity the robot needs to ask the user whether this is an alternative name for a location or if it made a mistake in localization or object classification. Thus the task of labeling rooms alone depends on three parameters that act in interplay:

- Lu: the label that the human user proposes for the current room. Users can name the same room differently for various reasons (e.g., by mistake, different users, at different time).
- Lc: robot’s perception of the current room, i.e., what the robot thinks where it is. The robot can be wrong in its perception, or it can also have registered a wrong label.
- Lm: The robot’s knowledge of other rooms.

The need for interaction arises when these three parameters do not correlate in a logical manner. For example, the user proposes the name A for the current room (Lu = A) while the robot “believes” that it is currently in an unlabeled place (Lc = new) and it already knows a room with the name A (A ∈ Lm). In this situation, possibly the user wants a second room to have the name A, or the robot’s perception is wrong and it is actually in the known room A. However, if the robot does not know another room with the name A (A ∉ Lm) then it is more likely that this is a normal operation of the user to label a room, which was previously unknown to the robot, with the name A. This would be a situation that hardly needs further negotiation between the user and the robot. To resolve such ambiguities the robot should be able to first initiate clarification questions and then

---

**Fig. 2.** Interplay of dialog and topological mapping.

---

**Diagram:**

- Ex 1: This is the kitchen.
- R1: I beg your pardon?
- U2: This is the kitchen.
- R2: Oh, you really have a nice kitchen!
- Ex 2: Where are you?
- U3: I’m in your wonderful kitchen.

**Table:**

<table>
<thead>
<tr>
<th>Case Name</th>
<th>Parameter combination</th>
<th>User Intention recognized through clarification</th>
<th>Operation of Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial labeling</td>
<td>A new A</td>
<td>label A</td>
<td>label A</td>
</tr>
<tr>
<td>repeated labeling</td>
<td>A A A</td>
<td>A</td>
<td>relocate to A</td>
</tr>
<tr>
<td>double labeling</td>
<td>A new A</td>
<td>user wants a second room to be labeled as A</td>
<td>add A</td>
</tr>
<tr>
<td>correction</td>
<td>B A A</td>
<td>user wants to correct Lc</td>
<td>relocate to A</td>
</tr>
<tr>
<td>relocate</td>
<td>B A A, B</td>
<td>user wants to rename the current location</td>
<td>add B</td>
</tr>
<tr>
<td>lost</td>
<td>any unknown any</td>
<td>asks the user to take the robot back to an known location for re-orientation</td>
<td>relocate to B</td>
</tr>
</tbody>
</table>

**Figure captions:**

(a) Dialog example (U: User, R: Robot).

(b) Use cases in situated spatial exploration.
relabel its entries in the environment representation or re-localize according to the user’s answer which requires a close cooperation between the dialog system and the Topological Localization. Figure 2(b) summarizes cases that occur most frequently in labeling tasks. For cases that require ambiguity resolution the relationship between the intention of the user (recognized by the dialog system through clarification) and the corresponding operation of the Topological Localization is specified. This clarification request behavior is comparable to very young children who expect only one label for the same object or area and need additional explanation otherwise. Besides ambiguities in the labeling process, ambiguities for areas or objects with the same name during data request should be solved. The topological map representation provides the information for the robot to discriminate a brush at the bathroom (for teeth cleaning) from a brush at the hallway for tidying up the room. So a fetch and carry task like ”Please, fetch me the brush” might be solved differently according to the recent position of the robot, or at least the robot will suggest the according brush via the dialog. For a robot already knowing its operating area a comfortable and more human-like way of guiding can be established.

Human navigation often takes relations to landmarks (corresponding to our ”areas”) into account, e.g., ”The kitchen is to the right side of the stairs.” instead of ”Go 5 meters ahead, turn 90 degrees clockwise, go 2 meters ahead.” With the hierarchical approach that allows to have areas within rooms such landmarks can be represented and used for navigation. In our scenario the robot can guide an interaction partner from one way point to another, with each way point being an entry on its topological map by following a person and politely asking her to turn around if she has gone into the wrong direction or guiding her to the next way point (”landmark”) according to the environment representation.

V. Interaction Awareness on the Integrated System

In the following we present more specific use cases in which our integrated system will prove its benefit and performance. Therefore we will explore areas in the real world, which are initially unknown to the robot BIRON.
B. Experimental Runs

Using a modular approach for easily extending the abilities of the robot towards more cooperative interactions enables us to investigate the arising questions concerning HRI in our daily surrounding. In order to test the robot in realistic environments an apartment (layout shown in Fig. 5(a)) has been rent permanently in which the interactive acquisition of the spatial representation is carried out by users in a guided tour (Fig. 5(b)). The approaches presented in this paper were successfully integrated and tested in this setting with different users introducing rooms, areas, and objects. The camera shots in Fig. 5 are taken from a recorded interactive trial with a user\(^1\). In this run, the user guided the robot from the living room to the kitchen passing the dining room and introduced the different rooms and referenced an object using a pointing gesture. Whenever the robot drew the conclusion that a new model of the room had to be acquired, it carried out a 360 degree turn to build a laser-based representation as outlined in section III (see Fig. 5(c) for the dining room). If learning of a new object is triggered by gesture and verbal input, the object attention system takes control and directs the robot’s pan-tilt camera to the reference object region (see Fig. 5(d)). After this tour, the robot has built up the hierarchical representation of the parts of the apartment shown to the robot. It is now able to answer questions regarding its current position (see exemplary dialog excerpt in Fig. 2(a)) which provides the information to relate objects to rooms.

VI. Conclusion

In this paper we presented a first integration of a localization and mapping system with a spoken dialog system for a joint exploration during a tutoring situation.

A major contribution of this work is a three-level hierarchy of spatial representation affecting three different communication dimensions. This hierarchy is consequently applied in the design of grounding-based dialog laser-based topological mapping, and an objects attention system. We demonstrated the benefits of this integration for learning and tour guiding in a human-comprehensible interaction between a robot and its user in a home-tour scenario with a real robot. The enhanced interaction capabilities are crucial for developing a new generation of robots that will be accepted not only as service robots but also as robot companions.

VII. Acknowledgments

This work has been supported by the European Union within the ‘Cognitive Robot Companion’ (COGNIRON) project (FP6-IST-002020) [1].

VIII. References


\(^1\)The complete video can be downloaded from http://www.techfak.uni-bielefeld.de/~mhanheid/biron1.avi