From Labels to Semantics: An Integrated System for Conceptual Spatial Representations of Indoor Environments for Mobile Robots

Óscar Martínez Mozos* Patric Jensfelt† Hendrik Zender‡ Geert-Jan M. Kruijff‡ Wolfram Burgard*
* University of Freiburg, Department of Computer Science, Freiburg, Germany
† Royal Institute of Technology, Center for Autonomous Systems, Stockholm, Sweden
‡ German Research Center for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany
*{omartine, burgard}@informatik.uni-freiburg.de, †patric@nada.kth.se, ‡{zender, gj}@dfki.de

Abstract— **We present an integrated approach for creating conceptual representations of human-made environments using mobile robots. The concepts represent spatial and functional properties of typical indoor environments. Our model is composed of layers which represent maps at different levels of abstraction. The complete system was integrated in a service robot which is endowed with laser and vision sensors for place and object recognition. It also incorporates a linguistic framework that actively supports the map acquisition process and is used for situated dialogue. In the experiments we show how the robot acquires the conceptual information and how it is used for situational and functional awareness.**

I. Introduction

Recently, there has been an increasing interest in robots whose aim is to assist people in human-like environments, such as domestic or elderly care robots. In such situations, the robots will no longer be operated by trained personnel but instead have to interact with people from the general public. Thus an important challenge lies in facilitating the communication between robots and humans.

One of the most intuitive and powerful ways for humans to communicate is spoken language. It is therefore interesting to design robots that are able to speak with people and understand their words and expressions. If a dialogue between robots and humans is to be successful, the robots must make use of the same concepts to refer to things and phenomena as a person would do. For this, the robot needs to perceive the world similar to a human.

An important aspect of human-like perception of the world is the robot's understanding of the spatial and functional properties of human-made environments, while still being able to safely act in it. For the robot, one of the first tasks will consist in learning the environment in the same way as a person does, sharing common concepts like, for instance, "corridor" or "living room". These terms can be used not only as labels but as semantic expressions that relate them to some complex object or objective situation. For example, the term "living room" usually implies a place with some particular structure, and includes objects like a couch or a television set. Moreover, a spatial knowledge representation for robotic assistants must address the issues involved with safe and reliable navigation control, with representing the space in a way similar to humans, and finally, with the way linguistic references to spatial entities are established in situated natural language dialogues.

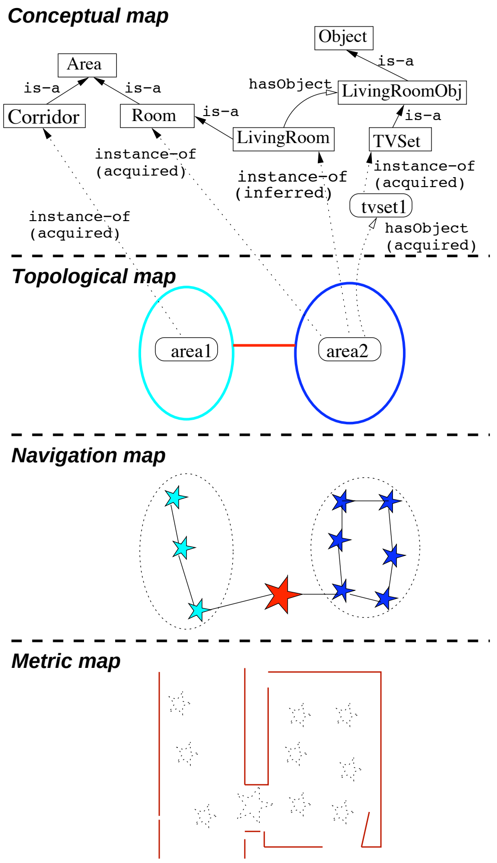

Fig. 1. An example of a layered spatial representation. Solid arrows indicate innate knowledge from the ontology. Dotted arrows refer to knowledge from the environment: asserted, acquired or inferred.

In this work we present an integrated approach for creating conceptual representations of human-made environments using mobile robots. The concepts represent spatial and functional properties of typical indoor environments. Our model is composed of layers containing maps at different levels of abstraction as shown in Fig. 1. The lower layers contain a metric map, a navigation map and a topological map, each of which plays a role in navigation and self-localization of the robot. On the topmost level of abstraction, the conceptual map provides a richer semantic view of the spatial organization, containing *acquired*, *asserted* and both *inferred* and *innate* conceptual-ontological knowledge about the environment. This model permits the robot to do spatial categorization rather than only instantiation.

The complete multi-layered representation is created in a semi-supervised map acquisition process, which is actively supported by a linguistic framework. This has been integrated into a cognitive system for mobile robots that is capable of conceptual spatial mapping in an indoor environment and that is endowed with the necessary abilities to conduct a reflected, situated dialogue about its environment.

The rest of the paper is organized as follows. In Section II, we present some related work. Section III describes our multi-layered conceptual spatial representation. The map acquisition process is outlined in Section IV. Situated dialogue is introduced in Section V. Section VI discusses how to achieve a notion of situational awareness using our conceptual representations. In Sections VII and VIII, we present implementation details and results respectively from an experimental evaluation of the integrated system. Finally, some concluding remarks are given in Section IX.

II. Related Work

Several approaches on mobile robotics extend metric maps of indoor environments with semantic information. The work by Diosi *et al.* [1] creates a metric map through a guided tour. The map is then segmented according to the labels given by the instructor. Martinez Mozos *et al.* [2] extract a topological semantic map from a metric one using supervised learning. Alternatively, Friedman *et al.* [3] use *Voronoi Random Fields* for extracting the topologies. In our system we use a similar approach to [2] for semantic classification.

Research in spatial representations has yielded different multi-layered environment models. Vasudevan *et al.* [4] suggest a hierarchical probabilistic representation of space based on objects. The work by Galindo *et al.* [5] presents an approach containing two parallel hierarchies, spatial and conceptual, connected through anchoring. Inference about places is based on objects found in them. Furthermore, the *Hybrid Spatial Semantic Hierarchy* (HSSH) is introduced by Beeson *et al.* [6]. This representation allows a mobile robot to describe the world using different representations each with its own ontology. Compared to these approaches our implementation uses human augmented mapping for collecting information. The communication with the robot is made entirely using natural language and dialogues. Moreover our conceptual representation comes from the fusion of acquired, asserted, and both inferred and innate knowledge.

There are more cognitively inspired approaches to robot navigation for conveying route descriptions from a technically naive user to a mobile robot. These approaches need not necessarily rely on an exact global self-localization, but rather require the execution of a sequence of strictly local, well-defined behaviors in order to iteratively reach a target position. Kuipers [7] presents the *Spatial Semantic Hierarchy* (SSH). Alternatively, the *Route Graph* model is introduced by Krieg-Brückner *et al.* [8]. Both theories propose a cognitively inspired multi-layered representation of the "map in the head", which is at the same time suitable for robot navigation. Their central layer of abstraction is the topological map. Our approach differs in that it provides an abstraction layer that can be used for reference resolution of topological entities.

A number of systems have been implemented that permit a robot to interact with humans in their environment. Rhino [9] and Robox [10] are robots that work as tour-guides in museums. Both robots rely on an accurate metric representation of the environment and use limited dialogue to communicate with people. The robot BIRON [11] is endowed with a system that integrates spoken dialogue and visual localization capabilities on a robotic platform similar to ours. This system differs from ours in the degree to which conceptual spatial knowledge and linguistic meaning are grounded in, and contribute to, situational awareness.

III. Multi-Layered Conceptual Mapping

The aim of this work is to generate spatial representations that enable a mobile robot to conceptualize human-made environments similar to the way humans do. These concepts correspond to spatial and functional properties of typical indoor environments. Following findings in cognitive psychology [12], we assume that topological areas are the basic spatial units suitable for situated interaction between humans and robots. We also proceed from the assumption that the way people refer to a place is determined by the functions people ascribe to that place and that the linguistic description of a place leads people to anticipate the functional properties or affordances of that place. At the same time, the constructed maps must allow for safe navigation and reliable self-localization of the robot. Considering these ideas, our final representation model is divided into layers, each representing a different level of abstraction. Each individual layer is important for the overall system because each layer serves a specific purpose. Starting from sensory input (laser scanner and odometry), a metric map and a navigation map representing traveled routes are constructed. On the basis of detected doorways, a topological partitioning of the navigation map is maintained. All these layers play a crucial role for the robot control systems. The conceptual map provides a conceptual abstraction layer of the lower layers. In this layer, spatial knowledge, innate conceptual knowledge and knowledge about entities in the world stemming from other modalities, such as vision and dialogue, are combined to allow for symbolic reasoning and situated dialogue. Fig. 1 depicts the four layers of the conceptual spatial representation.

A. Metric Map

The first layer of our model (Fig. 1, bottom) contains a metric representation of the environment in an absolute frame of reference. The geometric primitives consist of lines extracted from laser range scans. Such lines typically correspond to walls and other flat structures in the environment. The complete metric map is created by a mobile robot using *Simultaneous Localization and Mapping* (SLAM) techniques. The metric map is created online as the robot navigates around the environment based on the

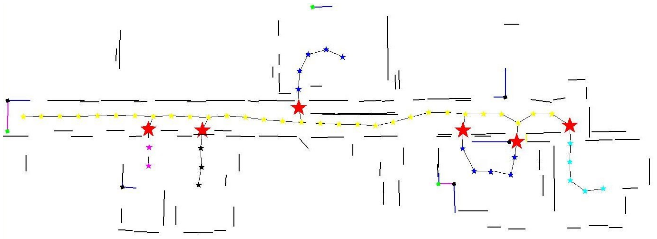

Fig. 2. The metric map is represented by lines. The navigation map is visually represented by the stars. Different colors represent different areas separated by doors, which are marked by bigger red stars.

same framework as in Folkesson *et al.* [13], which uses general representations for features that address symmetries and constraints in the feature coordinates to be added to the map with partial initialization. The number of dimensions for a feature can grow with time as more information is acquired. The basis for integrating the feature observations is the extended Kalman filter (EKF). An example metric map created using this method is shown in Fig. 2.

B. Navigation Map

The second layer contains the navigation map represented by a graph. This representation establishes a model of free space and its connectivity, i.e. reachability, and is based on the notion of a *roadmap* of *virtual free-space markers* [14], [15]. As the robot navigates through the environment, a marker (navigation node) is dropped whenever the robot has traveled a certain distance from the closest existing marker. The graph serves for planning and autonomous navigation in the known part of the environment.

We distinguish between two kinds of navigation nodes: place nodes and doorway nodes. Doorway nodes indicate the transition between different places and represent possible doors. They are detected and added whenever the robot passes through a narrow opening. Later, the status (open/closed) of a known door can be monitored using the laser scanner. Additionally, doorway nodes are assigned information about the door opening such as width and orientation.

Each place node is classified into one of two semantic labels, namely `Corridor` or `Room`, following the approach by Martinez Mozos *et al.* [2]. This method classifies the position of the robot based on the current scan obtained from the range sensor. The approach uses the AdaBoost algorithm to boost simple geometrical features into a strong classifier. Examples for typical features extracted from scans obtained in an office environment are shown in Fig. 3. The approach is supervised, which means that the robot must first be trained in an environment containing the semantic labels. As shown in [2] the training process does not have to be carried out in the same environment as the testing.

The approach for semantic classification assigns a label to each pose of the robot. To increase the robustness of the method, we classify each place node using the majority vote of the classification of the poses close to it. As explained before, a node is added when the distance to the previous node is greater than a threshold. We use this fact to store the classification of the last N poses of the robot in a buffer previous to adding the node. We then compute the majority vote of these last N poses and assign the final classification to the corresponding node.

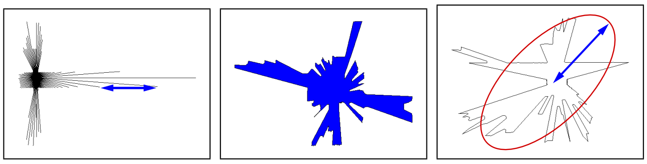

Fig. 3. Examples of features generated from laser data, namely the average distance between two consecutive beams, the perimeter of the area covered by a scan, and the mayor axis of the ellipse that approximates the polygon described by the scan. The laser beams cover a 360° field of view.

C. Topological Map

The topological map divides the set of nodes in the navigation graph into areas. An area consists of a set of interconnected nodes (cf. Fig. 2). In this view, the exact shape and boundaries of an area are irrelevant. The set of nodes is partitioned on the basis of the door detection mechanism explained in the previous section. This approach complies with previous studies [12], [16], which state that humans segment space into regions that correspond to more or less clearly defined spatial areas. The borders of these regions may be defined physically, perceptually, or may be purely subjective to the human. Walls in the robots environment are the physical boundaries of areas. Doors are a special case of physical boundaries that permit access to other areas.

D. Conceptual Map

The conceptual map provides the link between the low-level maps and the communication system used for situated human-robot dialogue by grounding linguistic expressions in representations of spatial entities, such as instances of rooms or objects. It is also in this layer that knowledge about the environment stemming from other modalities, such as vision and dialogue, is anchored to the metric and topological maps.

Based on the work by Zender [17], our system is endowed with a commonsense OWL[1] ontology of an indoor environment (see Fig. 4) that describes taxonomies (*is-a* relations) of room types and typical objects found therein through *has-a* relations. These conceptual taxonomies have been handcrafted and cannot be changed online. However, instances of the concepts are added to the ontology during run-time. Through fusion of *acquired* and *asserted* knowledge – gathered in an interactive map acquisition process (cf. Section IV) – and through the use of the *innate conceptual* knowledge, a reasoner[2] can *infer* information about the world that is neither given verbally nor actively perceived. This way linguistic references to spatial areas can be generated.

[1] `http://www.w3.org/TR/owl-guide/`
[2] `http://www.racer-systems.org`

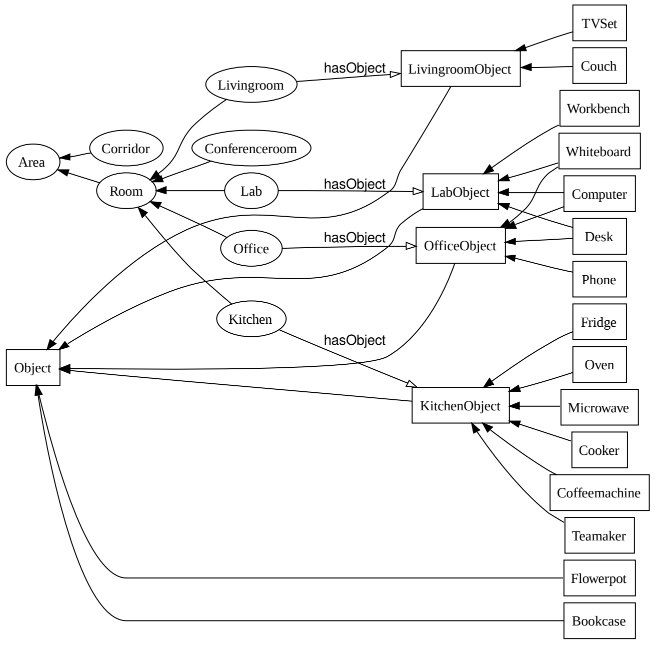

Fig. 4. Illustration of a part of the commonsense ontology of an indoor office environment. Solid arrows denote the taxonomical `is-a` relation.

1) Acquired Knowledge: While the robot moves around constructing the metric and topological maps, our system derives higher-level knowledge from the information in these layers. Each topological area, for instance, is represented in the conceptual map as an ontological instance of the type `Area`. Furthermore, as soon as reliable information about the semantic classification of an area is available, this is reflected in the conceptual map by assigning the area's instance a more specific type of either `Room` or `Corridor`. Information about recognized objects stemming from the vision subsystem is also represented in the conceptual map. Whenever a new object in the environment is recognized, a new instance of the object's type, e.g. `Couch`, is added to the ontology. Moreover, the object's instance and the instance of the area where the object is located are related via the `hasObject` relation. This process is shown in Fig. 1.

2) Asserted Knowledge: During a guided tour with the robot, the user typically names areas and certain objects that he or she believes to be relevant for the robot. Typical assertions in a guided tour include "You are in the corridor," or "This is the charging station." Any such assertion is stored in the conceptual map, either by specifying the type of the current area or by creating a new object instance of the asserted type and linking it to the area instance with the `hasObject` relation.

3) Innate Conceptual Knowledge: We have handcrafted an ontology (Fig. 4) that models conceptual commonsense knowledge about an indoor office environment. On the top level of the conceptual taxonomy, there are the two base concepts `Area` and `Object`. `Area` can be further partitioned into `Room` or `Corridor`. The basic-level subconcepts of `Room` are characterized by the instances of `Object` that are found there, as represented by the `hasObject` relation.

4) Inferred Knowledge: Based on the knowledge representation in the ontology, our system uses a description-logics based reasoning software that allows us to move beyond a pure labeling of areas. Combining and evaluating acquired and asserted knowledge within the context of the innate conceptual ontology, the reasoner can infer more specific categories for known areas. For example, combining the acquired information that a given topological area is classified as a room and contains a couch with the innate conceptual knowledge given in our commonsense ontology, it can be inferred that this area can be categorized as being an instance of `LivingRoom`. Conversely, if an area is classified as a corridor and the user shows the robot a charging station in that area, no further inference can be drawn. The most specific category the area instantiates will still be `Corridor`.

Our method allows for multiple possible classification of any area because the main purpose of the reasoning mechanisms in our system is to facilitate human-robot interaction. The way people refer to the same room can differ from situation to situation and from speaker to speaker, as reported by Topp *et al.* [18]. For example, what one speaker prefers to call the kitchen might be referred to as the recreation room by another person. Since our aim is to be able to resolve all such possible referring expressions, our method supports ambiguous classifications of areas.

IV. Interactive Map Acquisition

The multi-layered representation is created using an enhanced method for concurrent semi-supervised map acquisition, i.e. the combination of a user-driven supervised map acquisition process with autonomous exploration discovery by the robot. This process is based on the notion of *Human-Augmented Mapping*, as introduced by Topp and Christensen [19]. We additionally use a linguistic framework that actively supports the map acquisition process and is used for situated dialogue about the environment (see Section V).

The map can be acquired during a so-called guided tour scenario in which the user shows the robot around and continuously teaches the robot new places and objects. During such a guided tour, the user can command the robot to follow him or instruct the robot to perform navigation tasks. Our system does not require an initial complete guided tour. It is also possible to incrementally teach the robot new places and objects at any time the user wishes. With every new piece of information, the robot's internal representations become more complete. Still, the robot can always perform actions in, and conduct meaningful dialogue about, the aspects of its environment that are already known to it.

Whenever the user gives an assertion about areas in the environment or objects found therein, the robot updates the conceptual map with the asserted information. The concurrent constructions of the metrical map and the topological abstraction level propagate information in a bottom-up manner. Together with the laser-based area classification, these pieces of information lead to an update of the conceptual map with acquired knowledge.

Following the approach by Kruijff *et al.* [20], the robot can also initiate a clarification dialogue if it detects an inconsistency in its spatial representation, illustrating the mixed-initiative capabilities of the dialogue system.

V. Situated Dialogue

In this section, we will present the linguistic methods used for natural language dialogue with a robot. We will also address the role of dialogue for supervised map acquisition and task execution.

On the basis of a string-based representation that is generated from spoken input through a speech recognition software, the Combinatory Categorial Grammar (CCG) parser of OpenCCG[3] [21] analyzes the utterance syntactically and derives a semantic representation in the form of a Hybrid Logics Dependency Semantics (HLDS) logical form [22]. The dialogue system mediates the content from the speech input to the mapping or navigation subsystem in order to initiate the desired action of the robot or to collect pieces of information necessary to generate an answer. The generated answer string is then generated by the OpenCCG realizer and sent to a text-to-speech engine. The complete dialogue system is described in more detail in Kruijff *et al.* [23].

In the experiment of Section VIII, the user guides the robot around using a set of commands for initiating and stopping the interactive people following process and for instructing the robot with navigation commands to move near around. During this tour, the user augments the robot's internal map with assertions about the environment. In order to grasp the robot's understanding of its environment, the user has the possibility to ask the robot questions about the environment. The following examples contain HLDS representations of typical utterances in our scenario example:

(1) HLDS logical form of the utterance "This is the charging station."

$$
\begin{array}{l}
@_{\{B1:\mathsf{state}\}}(\mathbf{be}\\
\quad \& \ \langle Mood\rangle\mathbf{indicative}\\
\quad \& \ \langle Restr\rangle(T6 : thing \ \& \ \mathbf{this}\\
\qquad \& \ \langle Proximity\rangle\mathbf{proximal})\\
\quad \& \ \langle Scope\rangle(C3 : thing \ \& \ \mathbf{chargingstation}\\
\qquad \& \ \langle Delimitation\rangle\mathbf{unique}\\
\qquad \& \ \langle Number\rangle\mathbf{singular}))
\end{array}
$$

(2) HLDS logical form of the utterance "I am in a living room."

$$
\begin{array}{l}
@_{\{B9:\mathsf{state}\}}(\mathbf{be}\\
\quad \& \ \langle Mood\rangle\mathbf{indicative}\\
\quad \& \ \langle Restr\rangle(R2 : person \ \& \ \mathbf{I})\\
\quad \& \ \langle Scope\rangle(I4 : region \ \& \ \mathbf{in}\\
\qquad \& \ \langle Plane\rangle\mathbf{horizontal}\\
\qquad \& \ \langle Positioning\rangle\mathbf{static}\\
\qquad \& \ \langle Dir : Anchor\rangle(L1 : loc \ \& \ \mathbf{livingroom}\\
\qquad\quad \& \ \langle Delimitation\rangle\mathbf{existential}\\
\qquad\quad \& \ \langle Number\rangle\mathbf{singular})))
\end{array}
$$

(3) HLDS logical form of the utterance "Follow me!"

$$
\begin{array}{l}
@_{\{F3:\mathsf{action}\}}(\mathbf{follow}\\
\quad \& \ \langle Mood\rangle\mathbf{imperative}\\
\quad \& \ \langle Actor\rangle(R7 : hearer \ \& \ \mathbf{robot})\\
\quad \& \ \langle Patient\rangle(I2 : speaker \ \& \ \mathbf{I}))
\end{array}
$$

VI. Situational and Functional Awareness

We currently investigate how the information encoded in the multi-layered conceptual spatial representation can be used for a smarter, human- and situation-aware behavior. As one aspect of this, the robot should exploit its knowledge about objects in the environment to move in a way that allows for successful interaction with these objects. For instance, when following a person, the robot should make use of its knowledge about doors in the environment, such that it recognizes when the person wants to perform an action with the door. As actions that are performed in a doorway or with the door itself potentially require a wide space, e.g. for swinging or sliding open the door, for letting people pass, or for stepping past the door opening to grab the door handle, it is crucial that the robot adjusts its actions accordingly. A failure to understand such a situation could, for example, lead the robot to a position where it traps the user in the doorway that he or she was trying to close. In the experiment presented in this paper (see Section VIII), we opt for the robot to increase the distance it keeps to the user when it detects that the user approaches a door and to decrease it again when it detects that the user left the area. In this way, as the robot does not stop tracking and following the person, the people following behavior stays smooth and intuitive.

VII. System Integration Details

The complete system was implemented and integrated in an ActivMedia PeopleBot mobile platform (Fig. 5, left). The robot is equipped with a SICK laser range finder, which is used for the metric map creation, people following, and for the semantic classification of places. The place classification is based on a 360° field of view (Section III-B). However our robot has only one laser at the front covering a restricted 180° field of view. To solve this problem we follow the approach in [2] and maintain a local map around the robot, which permits us to simulate the rest of the beams covering the rear part of the robot. Additionally, a camera is used only for object detection. The detection systems uses SIFT features for finding typical objects like a television set, a couch or a bookcase. We recognize instances of objects and not categories [24]. The objects must be shown previously to the robot and learned by it (Fig. 5, right).

The communication with people was completely done using spoken language. The user can talk to the robot using a bluetooth headset and the robot replies using a set of speakers mounted on the mobile platform.

As an additional tool, we use an online viewer for the metric and navigation maps. The output of this program is composed of the lines extracted by our SLAM implementation extended to 3D planes to facilitate the visualization. The

[3] `http://openccg.sourceforge.net`

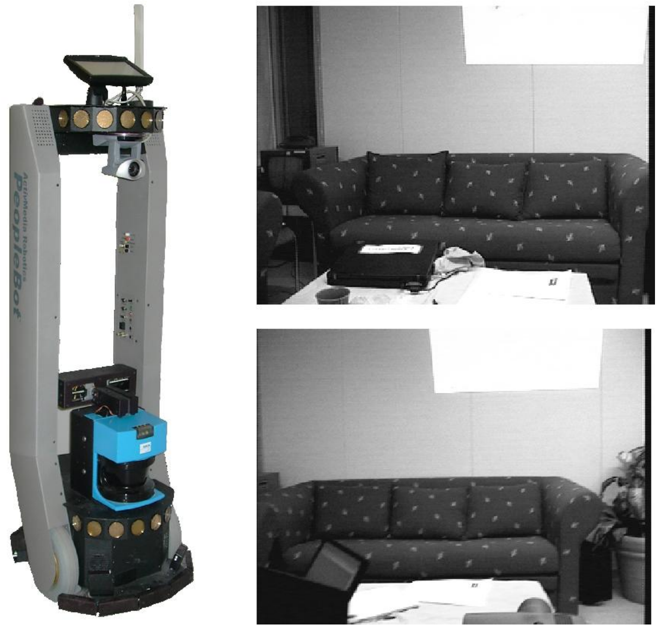

Fig. 5. The left image shows the robot used during the experiment. The right images depict examples for for object detection: training couch image (top), detected couch image (bottom).

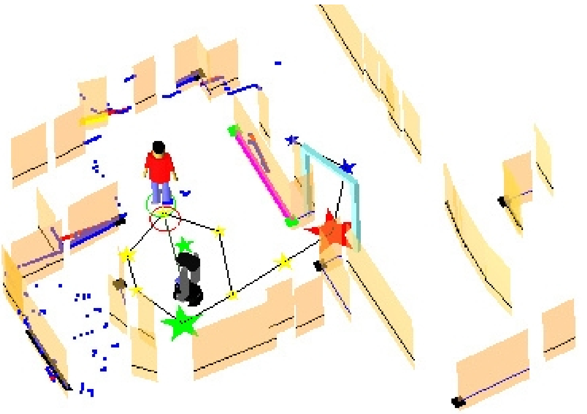

Fig. 6. Snapshot of the online viewer using during the experiment. The stars indicate the nodes in the navigation map. Blue for corridor, yellow for room, red for doorways and green for the actual position of the robot. Additionally, lines are extended to 3D planes and simulated doorways are drawn for facilitating the visualization. The person is drawn in the position detected by the people following software.

viewer shows the different nodes and edges used to construct the navigation map. Nodes corresponding to doorways are drawn bigger and with red color and with an associated doorframe (Fig. 6). Finally, the robot and the user are constantly shown in the positions where they are localized. The localization of the robot is calculated using SLAM [13], while the pose of the person is estimated using people tracking methods based only on laser readings [25].

The robot, being equipped with an onboard computer (850 MHz) connected to two built-in loudspeakers, runs the Player software[4] for control and access of the hardware, and the speech synthesis software[5]. The rest of the system runs on five laptops (1.8 GHz) interconnected using a wireless network. The first laptop is placed aboard the robot platform. It is connected to the onboard computer via an Ethernet crossover cable and to the rest of the system using its wireless adapter. This laptop runs the software for navigation, SLAM and people tracking. A second laptop runs the Windows operating system and is used for the real time speech recognition[6]. It is also placed on the robot platform in order to ensure a reliable bluetooth connection to the headset that recorded the user's voice commands. The recognized speech strings are sent to a third laptop, which runs the real-time dialogue processing and conceptual mapping subsystems. The fourth computer constantly classifies the current pose of the robot into a semantic class based on laser data. The last computer handles the viewer tool for debugging purposes. The communication between the different processes is established in a mixed environment using TCP/IP sockets and an OAA[7] framework. Fewer computers could have been used, but the setup was convenient as it allowed each subsystem developer to have his own computer.

VIII. Experiments

In order to show all the functionalities explained in the previous sections, we carried out an experiment at the 7th floor of the CAS building at the Royal Institute of Technology in Stockholm. In this experiment the robot, together with a user, goes through different situations (or episodes). The complete experiment was carried out non-stop, i.e. we did not stop the robot or restart the system at any moment. The duration of the complete experiment was of approximately 6 minutes. Each of the episodes is explained in detail in the next sections and a video is available on the Internet[8]. The experiment was thought of as a test, and for this reason we "forced" some artificial situations to simulate possible real ones (e.g. the false doorway of Section VIII-B). A similar experiment was carried out in which the robot interacts constantly with the user and the environment for more than 30 minutes during a demo in the CoSy project[9]. In this case, the robot was presented to an audience while explaining its actions. Some of the episodes were repeated to clarify some questions. The robot again run with no interruptions or system problems. This led us to think that our implementation is quite robust and maybe can serve as basis for a long term service robot.

The idea of the experiments is to show how the robot learns its environment while interacting with a tutor. However, some previous knowledge is needed during this process. First, the robot needs an ontology representing the general knowledge about the environment. For this purpose, we use the ontology depicted in Fig. 4. Furthermore, the classification of places is based on previous general knowledge about the geometry of rooms and corridors, which is encoded in a

[4] http://playerstage.sourceforge.net/
[5] http://www.cstr.ed.ac.uk/projects/festival/
[6] http://www.nuance.com
[7] http://www.ai.sri.com/~oaa/
[8] http://www.dfki.de/cosy/www/media
[9] http://www.cognitivesystems.org

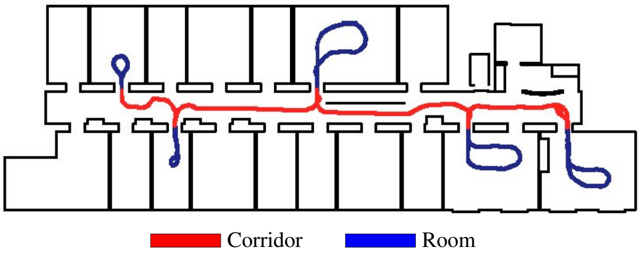

Fig. 7. Trajectory followed by the robot to train the classifier for distinguishing between corridor and room. The different places are depicted with distinct colors.

classifier based on laser readings as explained in Section III-B. The classifier is trained using examples of corridors and rooms from real environments as the one shown in Fig. 7. These two kinds of knowledge are independent of the environment used for testing, in the sense that the robot does not need to be physically present in the test environment to acquire the information. Finally, the robot has to recognize different objects, such as couches or TV sets, using vision. Because we do instance recognition rather than categorization, the objects we want to recognize must be presented to the robot before running the experiment. For this purpose, we position the robot in front of these objects, acquire a training image and label it with the corresponding term, which is added to a small database of objects and also included in the language systems for its posterior use.

A. *Episode 1: Waking Up*

The experiment starts in the corridor, where the robot is positioned close to the charging station. The user activates the robot and tells it that it is located at the charging station. The user then asks the robot to follow him. The robot drops markers (navigation nodes), which are classified as corridor. Then the person followed by the robot enters a room through a doorway. The door is recognized and the corresponding node is set. From this point the next nodes will be classified as a new area and correctly labeled as room.

B. *Episode 2: Clarification Dialogues*

In this episode we want to show the utility of the clarification dialogues. As explained in Section III-B, our door detection is simply based on detecting when the robot passes through a narrow opening. However, this alone will still lead to some false doors in cluttered rooms. Assuming that there are few false negatives in the detection of doors, we get great improvements by enforcing that it is not possible to change room without passing through a door. For example, while moving around in a room the robot may detect a narrow passage and falsely assume that a door was passed, putting a door label on that particular node. The robot continues to move around in the room and eventually reaches the nodes from before adding the false door. These nodes will then have different room labels, that is, the room has changed without passing a door. If this happens, an inconsistency is found and a clarification dialogue with the user is triggered.

Fig. 8. The user asks the robot: “Where is the charging station?”.

To test the former situation we put a bucket close to a table in the room creating an illusion of a doorway when using only the laser as sensor. The robot passes through this false doorway and comes back to a previously visited node. At this point the robot infers that there is an inconsistency in the map and initializes a clarification dialogue asking if there was a door previously. The user denies this fact and the map is updated accordingly. A more detailed explanation of the complete process of clarification dialogues for a similar situation is presented in Kruijff *et al.* [20].

C. *Episode 3: Inferring New Concepts*

In this episode we test how the robot infers new categorizations of places when discovering new objects. The goal is to use our SIFT-based object detector together with the laser-based place classification to detect simple objects and places. Then, using the inference on the office ontology as explained in Section III-D, the robot is able to come up with more specific concepts.

While staying in the room, the robot is asked for the current place and it answers with the indefinite description “a room”, which is inferred from the navigation nodes in the area. A majority vote among the nodes in the area is used in case the node classification is not unanimous. Then the robot is asked to look around. This command activates the vision-based object detection capabilities of the robot. The robot moves and detects a couch, and then a television set. After that, the user asks the robot for the name of the place. Because of the inference over the detected objects and places, the robot categorizes the place as a `Livingroom`. Note that previous to the detection of objects the same place was categorized as a `Room`. As a further test of the robot’s classification it is asked where the charging station is located and correctly answers “it is in a corridor” (Fig. 8).

D. *Episode 4: Situational and Functional Awareness*

This episode shows the social capabilities of our robot. The robot must behave accordingly to the current situation, which in our case, is the opening of a door by the user (see Section VI).

Continuing with the experiment, the user asks the robot to follow him while he approaches a doorway. The robot knows from the navigation map where the doorway is and keeps a long distance to the user when he is near to the door. It then continues following the user by again decreasing its distance to him when he has passed the door. This action implies a certain degree of knowledge about social behavior, which is important if the goal is to create a robot that will live together with people.

E. Episode 5: Going to Objects

Finally, we show how the navigation map is used by the robot to come back to previously visited places.

After the door opening situation, the robot is asked to go to the television. The robot then navigates to the node where the television was observed. This functionality permits the user to command the robot to places without the need of giving concrete coordinates. It is also more powerful in the sense that the user may not know the concrete name of the place, but he can remember it as ‘the room with a television". After that, the robot is commanded to go to the charging station. Again the robot follows the navigation map until it positions itself on the station, thus finishing the experiment.

IX. Conclusions

We presented an integrated approach for creating conceptual representations of human-made environments where the concepts represent spatial and functional properties of typical office indoor environments. Our representation is based on multiple maps at different levels of abstraction. The complete system was integrated and tested in a service robot which includes a linguistic framework with capabilites for situated dialogue and map acquisition. The experiments show that our system is able to provide a high level of human-robot communication and certain degree of social behavior.

Acknowledgment

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project “CoSy” FP6-004250-IP.

References

[1] A. Diosi, G. Taylor, and L. Kleeman, “Interactive SLAM using laser and advanced sonar,” in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, Barcelona, Spain, April 2005.

[2] O. Martínez Mozos, A. Rottmann, R. Triebel, P. Jensfelt, and W. Burgard, “Semantic labeling of places using information extracted from laser and vision sensor data,” in *IEEE/RSJ IROS Workshop: From Sensors to Human Spatial Concepts*, Beijing, China, 2006.

[3] S. Friedman, H. Pasula, and D. Fox, “Voronoi random fields: Extracting the topological structure of indoor environments via place labeling,” in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, 2007.

[4] S. Vasudevan, S. Gachter, M. Berger, and R. Siegwart, “Cognitive maps for mobile robots an object based approach,” in *In Proc. of the IEEE/RSJ IROS 2006 Workshop: From Sensors to Human Spatial Concepts*, Beijing, China, 2006.

[5] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernández-Madrigal, and J. González, “Multi-hierarchical semantic maps for mobile robotics,” in *Proc. of the IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, Edmonton, Alberta, Canada, 2005.

[6] P. Beeson, M. MacMahon, J. Modayil, A. Murarka, B. Kuipers, and B. Stankiewicz, “Integrating multiple representations of spatial knowledge for mapping, navigation, and communication,” in *Interaction Challenges for Intelligent Assistants*, ser. AAAI Spring Symposium, Stanford, CA, USA, 2007.

[7] B. Kuipers, “The Spatial Semantic Hierarchy,” *Artificial Intelligence*, vol. 119, pp. 191–233, 2000.

[8] B. Krieg-Brückner, T. Röfer, H.-O. Carmesin, and R. Müller, “A taxonomy of spatial knowledge for navigation and its application to the Bremen autonomous wheelchair,” in *Spatial Cognition*, ser. Lecture Notes in Artificial Intelligence, C. Freksa, C. Habel, and K. F. Wender, Eds. Springer Verlag, 1998, vol. 1404, pp. 373–397.

[9] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, “Experiences with an interactive museum tour-guide robot,” *ArtificialIntelligence*, vol. 114, no. 1–2, 2000.

[10] R. Siegwart and et al., “Robox at expo.02: A large scale installation of personal robots,” *Robotics and Autonomous Systems*, vol. 42, pp. 203–222, 2003.

[11] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. J. A. Kröse, “BIRON, where are you? - enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization,” in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006, pp. 934–940.

[12] T. McNamara, “Mental representations of spatial relations,” *Cognitive Psychology*, vol. 18, pp. 87–121, 1986.

[13] J. Folkesson, P. Jensfelt, and H. Christensen, “Vision SLAM in the measurement subspace,” in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA’05)*, 2005, pp. 30–35.

[14] J. C. Latombe, *Robot Motion Planning*. Boston, MA: Academic Publishers, 1991.

[15] P. Newman, J. Leonard, J. Tardós, and J. Neira, “Explore and return: Experimental validation of real-time concurrent mapping and localization,” in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, Washington, D.C., USA, 2002, pp. 1802–1809.

[16] S. C. Hirtle and J. Jonides, “Evidence for hierarchies in cognitive maps,” *Memory and Cognition*, vol. 13, pp. 208–217, 1985.

[17] H. Zender, “Learning spatial organization through situated dialogue,” Master’s thesis, Dept. of Computational Linguistics, Saarland University, Saarbruecken, Germany, 2006.

[18] E. A. Topp, H. Hüttenrauch, H. Christensen, and K. Severinson Eklundh, “Bringing together human and robotic environment representations – a pilot study,” in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October 2006.

[19] E. A. Topp and H. I. Christensen, “Tracking for following and passing persons,” in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, Alberta, Canada, 2005.

[20] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, “Clarification dialogues in human-augmented mapping,” in *Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI 2006)*, Salt Lake City, UT, USA, 2006.

[21] J. Baldridge and G.-J. M. Kruijff, “Multi-modal combinatory categorial grammmar,” in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, 2003.

[22] ——, “Coupling CCG and hybrid logic dependency semantics,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, PA, USA, 2002, pp. 319–326.

[23] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, “Situated dialogue and spatial organization: What, where. . . and why?” *International Journal of Advanced Robotic Systems, special section on Human and Robot Interactive Communication*, vol. 4, no. 2, March 2007.

[24] D. Lowe, “Distinctive image features from scale-invariant keypoints,” in *Int. Journal of Computer Vision*, vol. 60, no. 2, 2004, pp. 91–110.

[25] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers, “People tracking with a mobile robot using samplebased joint probabilistic data association filters,” *International Journal of Robotics Research*, vol. 22, no. 2, pp. 99–116, 2003.