Robotics and Autonomous Systems 56 (2008) 493–502

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Conceptual spatial representations for indoor mobile robots☆

H. Zender [a,*], O. Martínez Mozos [b], P. Jensfelt [c], G.-J.M. Kruijff [a], W. Burgard [b]

[a] German Research Center for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Campus D32, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[b] Albert-Ludwigs-Universität Freiburg, Department of Computer Science, Georges-Köhler-Allee, Geb. 079, D-79110 Freiburg, Germany
[c] Royal Institute of Technology (KTH), Centre for Autonomous Systems, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Available online 25 March 2008

Keywords:
Spatial representation
Conceptual map
Mapping
Service robots
Mobile robots

ABSTRACT

We present an approach for creating conceptual representations of human-made indoor environments using mobile robots. The concepts refer to spatial and functional properties of typical indoor environments. Following different findings in spatial cognition, our model is composed of layers representing maps at different levels of abstraction. The complete system is integrated in a mobile robot endowed with laser and vision sensors for place and object recognition. The system also incorporates a linguistic framework that actively supports the map acquisition process, and which is used for situated dialogue. Finally, we discuss the capabilities of the integrated system.

1. Introduction

Recently, there has been an increasing interest in service robots, such as domestic or elderly care robots, whose aim is to assist people in human-made environments. 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 similarly 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 are 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 which includes objects like a couch or a television set. Thus representing the space in a way similar to humans needs to also account for the way linguistic references to spatial entities are established in situated natural language dialogues. In addition, a spatial knowledge representation for robotic assistants must address the issues involved with safe and reliable navigation control. Only then robots can be deployed in semi-structured environments, such as offices, where they have to interact with humans in everyday situations.

The specific problem we focus on in this article is how, given innate (possibly human-like) concepts a robot may have of spatial organization, the robot can autonomously build an internal representation of the environment by combining these concepts with different low-level sensory systems. This is done by creating a conceptual representation of the environment, in which the concepts represent spatial and functional properties of typical human-made indoor environments.

In order to meet both of the aforementioned requirements – robust robot control and human-like conceptualization – we propose a spatial representation that contains maps at different levels of abstraction. This stepwise abstraction from raw sensor input not only produces maps that are suitable for reliable robot navigation, but also yields a level of representation that is similar to a human conceptualization of spatial organization. Furthermore, this model provides a richer semantic view of an environment that permits the robot to do spatial categorization rather than only instantiation.

Our approach has been integrated into a system running on a mobile robot. This robot is capable of conceptual spatial mapping in an indoor environment, perceiving the world through different

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

* Corresponding author. Tel.: +49 0 6813025004.
E-mail addresses: zender@dfki.de (H. Zender), omartine@informatik.uni-freiburg.de (O. Martínez Mozos), patric@nada.kth.se (P. Jensfelt), gj@dfki.de (G.-J.M. Kruijff), burgard@informatik.uni-freiburg.de (W. Burgard).

0921-8890/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2008.03.007

typical sensors like a laser range finder and a camera. Moreover, the robot 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 2 we present related work. Section 3 gives an overview of the components of our robotic system. After explaining the individual techniques that are used for evaluating the sensory input in Section 4, we describe our approach to a multi-layered conceptual spatial representation that bridges the gap between sensory input and human spatial concepts in Section 5. Then, the general principles of our robot's situated dialogue capabilities are introduced in Section 6. In Section 7, we discuss the integration of the complete system in a mobile robot. Finally, concluding remarks are given in Section 8.

2. Related work

An approach to endowing autonomous robots with a human-like conceptualization of space inherently needs to take into account research in sensor-based mapping and localization for robots as well as findings about human spatial cognition.

Research in cognitive psychology addresses the inherently qualitative nature of human spatial knowledge. Backed up by experimental studies, it is nowadays generally assumed that humans adopt a partially hierarchical representation of spatial organization [1,2]. The basic units of such a qualitative spatial representation are topological regions [3], which correspond to more or less clearly bounded spatial areas. The borders may be defined physically, perceptually, or may be purely subjective to the human. It has been shown that even in natural environments without any clear physical or perceptual boundaries, humans decompose space into topological hierarchies by clustering salient landmarks [4]. In our approach, topological areas are the primitive units of the conceptual map that is used for human–robot interaction and dialogue.

Aside from the functionality of the cognitive map, another relevant question from cognitive science is how people categorize spatial structures. Categories determine how people can interact with, and linguistically refer to entities in the world. Basic-level categories represent the most appropriate name for a thing or an abstract concept. The basic-level category of a referent is assumed to provide enough information to establish equivalence with other members of the class, while distinguishing it from non-members [5,6]. We draw from these notions when categorizing the spatial areas in the robot's conceptual map. We are specifically concerned with determining appropriate properties that allow a robot to both successfully refer to spatial entities in a situated dialogue between the robot and its user, and meaningfully act in its environment.

There are different cognitively inspired approaches to robot navigation. 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 additional abstraction layer that can be used for categorization of topological entities.

Recently, a number of methods originating in robotics research have been presented that construct multi-layered environment models. These layers range from metric sensor-based maps to abstract conceptual maps that take into account information about objects acquired through computer vision methods. Vasudevan et al. [9] suggest a hierarchical probabilistic representation of space based on objects. The work by Galindo et al. [10] 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), introduced by Beeson et al. [11], allows a mobile robot to describe the world using different representations, each with its own ontology. Compared to these approaches our system puts more emphasis on user interaction, both for collecting knowledge about the world and for communicating the robot's knowledge to its user. Moreover, our conceptual spatial representation is constructed through fusion of acquired, asserted, and both inferred and innate knowledge.

Additionally, several mobile robotics approaches extend metric maps of indoor environments with semantic information. The work by Diosi et al. [12] 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. [13] extract a topological semantic map from a metric one using supervised learning. Alternatively, Friedman et al. [14] use *Voronoi Random Fields* for extracting the topologies. In our system we use a similar approach to [13] for semantic place classification.

Moreover, a number of systems have been implemented that permit a robot to interact with humans in their environment. Rhino [15] and Robox [16] 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. Examples of robots with more elaborate dialogue capabilities are RoboVie [17], BIRON [18], Godot [19], WITAS [20] and Mel [21]. BIRON is endowed with a system that integrates spoken dialogue and visual localization capabilities on a robotic platform. This system differs from ours in the degree to which conceptual spatial knowledge and linguistic meaning are grounded in, and contribute to, situation awareness. In contrast, in our system information from dialogue and situated contexts can be combined during processing utterances [22]. Furthermore, whereas RoboVie and BIRON use finite state machines to model dialogue behavior, we combine information states [23], like Godot, together with a task-oriented perspective, as WITAS or Mel.

In [24] we present the cognitive architecture of our robotic system and give details of its dialogue capabilities. We furthermore discuss how these components are used for interactive map acquisition. Complementing this work the present article focuses on our method for representing the environment on several levels of abstraction. Details about the computer vision algorithms used for object detection, about the processing of sensory input from a laser scanner, and about the principles of knowledge processing in the conceptual map layer are given.

3. System overview

Following the research in spatial cognition and qualitative spatial reasoning on the one hand, and in mobile robotics and artificial intelligence on the other hand, we propose a spatial representation for indoor mobile robots that is divided into layers. These layers represent different levels of abstraction from sensory input to human-like spatial concepts.

This multi-layered spatial representation is the centerpiece of our integrated robotic system. It is created using information coming from different modalities, as shown in Fig. 1. The individual modalities range from low-level robot control and perception modules to a communication subsystem for spoken dialogue with the user. There are three main subsystems involved in constructing, maintaining, and using the spatial representation: the perception subsystem for evaluation of sensory input, the

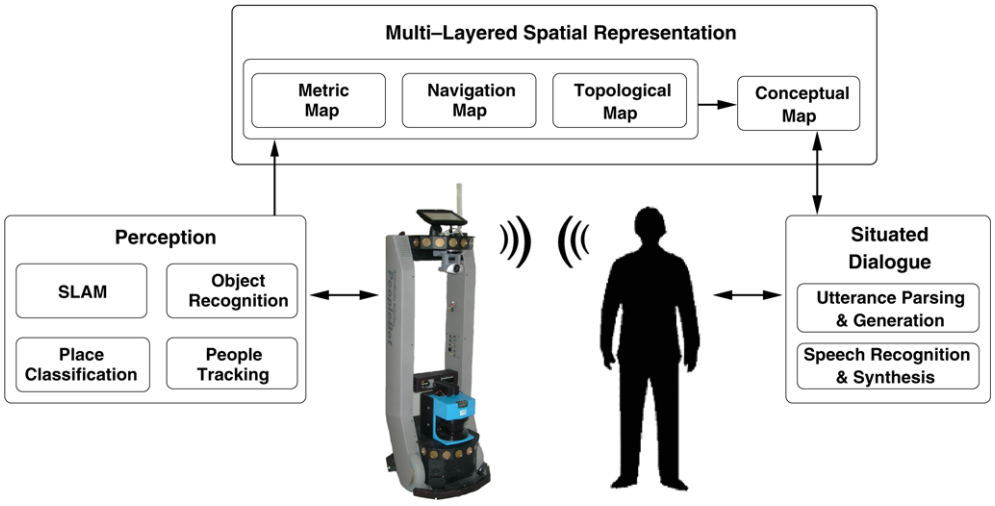

Fig. 1. Overview of the components of our robotic system.

communication subsystem for situated spoken dialogue, and the subsystem for multi-layered conceptual spatial mapping that bridges the gap between sensor-based maps and a human-like spatial representation. The main techniques used in the perception and communication subsystems and the structure of the multi-layered spatial representation that sits at the core of our system are explained in more detail in the following sections.

4. Perception

The perception subsystem gathers information from the laser range scanner and from a camera. Different techniques are used for evaluation of the sensory input. The laser data is processed and used to create the low-level layers of the spatial representation. At the same time the input from the laser scanner is used by a component for detecting and following people [25]. Finally, the images acquired by the camera are analyzed by a computer vision component for object recognition.

4.1. Simultaneous localization and mapping

To reach a high level of autonomy the robot needs the ability to build a map of its environment that can be used to safely navigate and stay reliably localized. To this end we use the Simultaneous Localization and Mapping (SLAM) technique described in [26]. In our system the SLAM module extracts geometric primitives from laser range scans and applies an Extended Kalman Filter (EKF) framework for the integration of feature measurements. The geometric features used in our approach are lines, which typically correspond to walls and other straight structures that appear as a line segment at the height of the laser scanner. Since walls are in most cases static, these invariant features of the environment are used to keep the robot localized. The line features are stored in a global metric map with an absolute frame of reference. Fig. 2 shows an example of a line map created using this method.

4.2. Place information

Apart from line features other features can be derived from the laser range data. These features are useful to semantically interpret the position at which they were detected. In our approach a laser scan can be interpreted as belonging to one of three semantic place classes: doorway, corridor, or room.

Doorways indicate the transition between different spatial regions. They are detected and added whenever the robot passes through an opening of door width. The width of the opening is selected so that it agrees with standard doorways in the environment. Information about the door opening, such as width and orientation, is stored along with the detected position of the doorway. A more complex door model as in [27,28] would allow more robust door detection but also puts constraints on how doors have to look to be recognized. Our only model assumption is that the door is a narrow opening which the robot passes through but we make no assumptions about having a swinging or sliding door leaf, having some special structure around the door, for example. This is an important consideration for a robot that has to operate in different environments. An alternative would be to use a learning approach such as in [29] where both visual features and the motion of the door is taken into account.

Fig. 2. The SLAM module creates a metric line map representing walls and other straight surfaces in the environment.

Corridors and rooms are classified according to the laser observation that the robot takes at that location. The main idea of this approach is to extract simple geometrical features from the laser scans and their polygonal approximation. All features are rotational invariant to make the classification of a pose dependent only on the (x, y)-position of the robot and not of its orientation. Examples for typical features are shown in Fig. 3. Features are then represented by weak hypotheses that are boosted into a strong classifier as presented in [30].

This approach is supervised, which means that the robot must be first trained in an indoor environment containing rooms and corridors. However, as we will see later, the training environment can be a different one from where the classifier is used, as it is able to generalize quite well.

4.3. Object recognition

Objects play an integral role in the conceptual map presented in this paper, as the information of recognized objects is used for inferring subconcepts (e.g. “kitchen” or “living room”) for rooms.

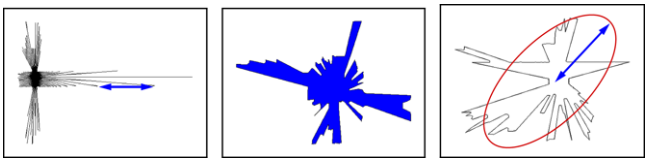

Fig. 3. Examples of features generated from laser data, namely the average distance between two consecutive beams, the perimeter of the area covered by a scan, and the major axis of the ellipse that approximates the polygon described by the scan. Here, the laser beams cover a 360° field of view.

Fig. 4. Two of the training images for object recognition.

Object recognition has been and still is a very active area of research. Recently, the so-called Scale Invariant Feature Transform or SIFT [31] was presented. It has been shown to give good results for object recognition. We have investigated two slightly different methods for the recognition of objects.

In the first method, SIFT features are extracted from all training images and put in one common KD-tree. Each SIFT feature is given a label that refers back to from what object it comes. During recognition, SIFT features are extracted from the new images. Each feature is matched to all training images at once using the KD-tree to perform fast matching. Each match with a feature in the KD-tree represents one vote for a certain object, namely the one contributing the corresponding SIFT feature. The output of this initial matching step is the list of objects that accumulated the most votes. In the second step we try to verify a match using the standard SIFT matching algorithm [31] against this list of most likely objects. The initial step allows us to prune the search space in order to avoid having to perform matching against all objects in the database. As we use a low resolution image (320 x 240 pixels) we are limited to using rather large objects, like the TV set and the flower from Fig. 4. Also note that since we do not segment the objects and thus do not tell the system what is foreground and background in the training image we are actually performing image recognition rather than object recognition. This means that unless the object in question is dominant enough in the image the system might still recognize the scene as that object even without it.

In an attempt to allow for the use of smaller objects, like cups, books, etc. we applied a second method which is based on the ideas presented in [32]. One of the key points here is that object recognition, when performed on a mobile platform, actually should be thought of as two separate processes, object detection and object recognition. The first step consists in finding out where the objects are. In a second step, the identity of the objects is verified. As reliable recognition of small objects at typical indoor distances (2–3 m) is not possible with the low resolution images, we also rely on the use of zoom. Fig. 5 shows the type of training image used in this approach along with a sequence of more and more zoomed in images of the same object.

To guide the search, we use an attention mechanism based on receptive field cooccurrence histograms (RFCH) [33]. RFCH provides us with a vote matrix for each object we search for. This vote matrix tells us where in the image the corresponding object is likely to be found, if at all. Fig. 6 shows an example of a test image and the corresponding vote matrix when looking for this object.

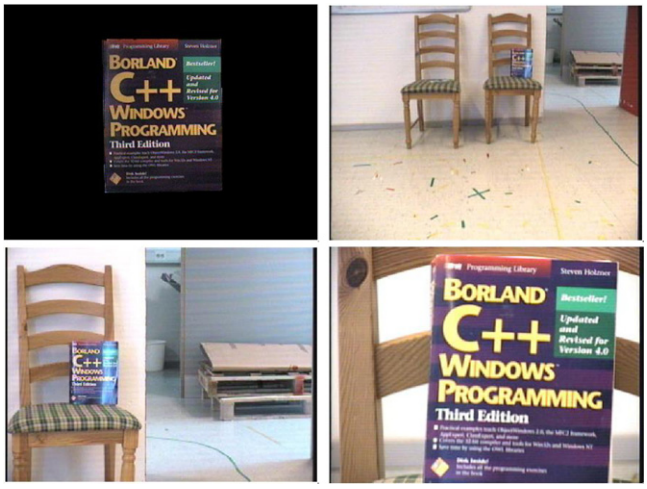

Fig. 5. The training image followed by an example of how zooming allows for a closer look of the object.

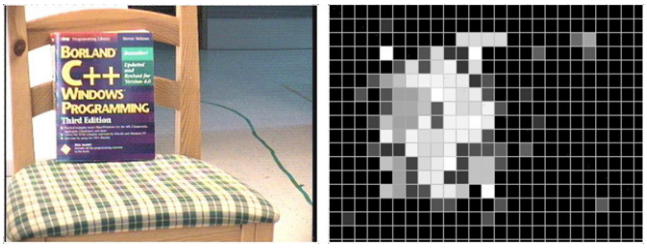

Fig. 6. Image and the corresponding vote matrix from RFCH that guides the zooming.

The lighter the color the higher the likelihood of a match. A more detailed description of this system can be found in [34].

With this method we can detect even quite small objects at a relatively long distance. The main problem with the method is that it is rather slow in its current form. The time to search a scene scales roughly linearly with the number of objects in the database. One improvement to this would be to use top–down information to cut the number of objects to search for. For example, if we know that we are in a living room we do not have to search for a coffee machine.

5. Multi-layered spatial representation

The sensors that a robot has are very different from the human sensory modalities. Yet if a robot is to act in a human-populated environment, and to interact with users that are not expert roboticists, it needs to understand its surroundings in terms of *human spatial concepts*. We propose a layered model of space at different levels of abstraction that range from low-level metric maps for robot localization and navigation to a conceptual layer that provides a human-like decomposition and categorization of space. Fig. 7 depicts the main layers of the conceptual spatial representation.

The lower layers of our model are derived from sensor input. Different methods are used to gradually construct more abstract representations. On the highest level of abstraction, we regard topological regions and spatially situated objects as the primitive entities of a spatial conceptualization that is compatible with human environment models. In order for a robot to meaningfully act in, and talk about, an environment, it must be able to assign human categories to spatial entities. Our work rests on the assumption that the basic-level categories of spatial entities in an environment are determined by the actions they afford.

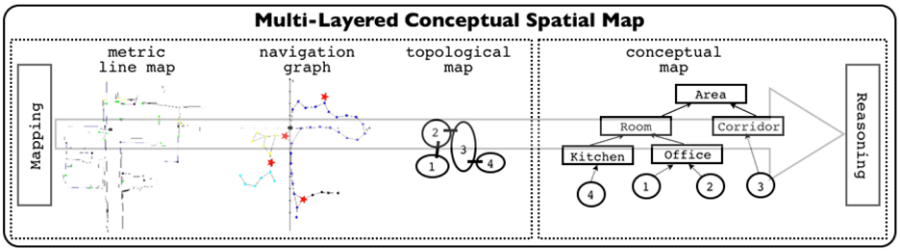

Fig. 7. Our multi-layered map, ranging from sensor-based maps to a conceptual abstraction.

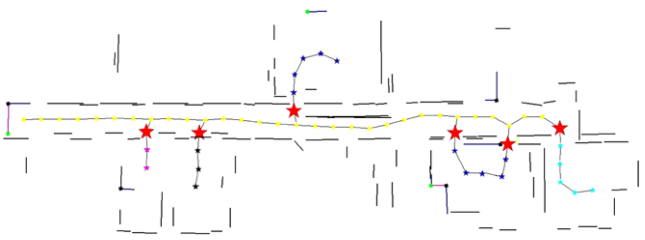

Fig. 8. The navigation map overlaid on the metric map. The navigation map is visually represented by the stars. Different colors represent different areas separated by doors, which are marked by bigger red stars.

Many types of rooms are designed in a way that their structure and spatial layout afford specific actions, such as corridors, or staircases. Other types of rooms afford more complex actions. These are in most cases provided by objects that are located there. For instance, the concept *living room* applies to rooms that are suited for resting. Having rest, in turn, can be afforded by certain objects, such as couches or TV sets. We thus conclude that besides basic geometric properties, such as shape and layout, the objects that are located in a room are a reliable basis for appropriately categorizing it.

Below, the individual layers of our spatial representation will be addressed more closely.

5.1. Metric map

At the lowest layer of our spatial model, we have a metric map. In this map, lines are the basic primitive to represent the boundaries of open space. The metric line map supports self-localization of the robot. It is maintained and used by the SLAM component, as described in Section 4.1.

As can be seen in Fig. 2, the line-based metric map gives a rather sparse description of the environment, not sufficient to fully support navigation actions. In comparison to an occupancy grid representation [35], the line-based map does not provide a description of the free space but only the part of the space that can be described by lines. Moreover, since the global coordinate system of the metric map is purely internal to the robot and since humans are not able to easily (i.e. without additional tools) evaluate quantitative spatial descriptions, the metric map alone does not provide a suitable common ground for human–robot dialogues.

5.2. Navigation graph

The next layer of our representation is composed of a navigation graph, which establishes a model of free space and its connectivity, i.e. reachability. It is based on the notion of a *roadmap* of *virtual free-space markers* as described in [36,37]. As the robot navigates through the environment, a marker or *navigation node* is dropped whenever the robot has traveled a certain distance from the closest existing node. Nodes are connected following the order in which they were generated. This order is given by the trajectory that the robot follows during the map acquisition process (see Fig. 8). The final graph serves for planning and autonomous navigation in the already visited part of the environment.

It is also in the navigation graph that the robot's spatial representation is augmented with semantic environment information. This is encoded by assigning navigation nodes one of three classes which can be considered to be present in every indoor environment. The classes are room, corridor, and doorway.

The approach presented in Section 4.2 for semantic classification assigns a label (corridor or room) to each pose of the robot during a trajectory. However, we are interested in classifying navigation nodes, which are dropped only when the robot has moved a certain distance (1 m in our case). Classifying only the exact location of the node into one class will ignore previous information about the labels of the poses leading to the mark. To use this information, we store the classification of the last N poses of the robot in a short term memory. This label history will be used to classify the node using a majority vote approach. In our experiments we obtained a significant improvement when using this approach for node classification.

Objects detected by the computer vision component are also stored on this level of the map. They are associated with the navigation node that is closest to their estimated metric position.

5.3. 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 which are separated by a node classified as a doorway. In Fig. 8, the topological segmentation is represented by the coloring of the nodes. This layer of abstraction corresponds to a human-like qualitative segmentation of an indoor space into distinct regions. In this view, the exact shape and boundaries of an area, as represented in the lower map layers, are abstracted to a coarse categorical distinction between rooms and corridors. In order to determine the category of an area, we take a majority vote approach of the classification results of all nodes in the given area. The topological areas, along with detected objects (Section 4.3), are passed on to the conceptual map, where they are represented as instances of their respective categories.

5.4. Conceptual map

On the highest level of abstraction, our system is endowed with a conceptual map. The nature of this map is two-fold. For one, it contains an innate conceptual ontology that defines abstract categories for rooms and objects and how they are related. Second, the information extracted from sensor data and given through situated dialogue about the actual environment is represented as tokens that instantiate abstract concepts. This division corresponds to the distinction between a TBox (terminological knowledge, i.e. concepts) and an ABox (assertional knowledge, i.e. instances) in traditional knowledge representation systems.

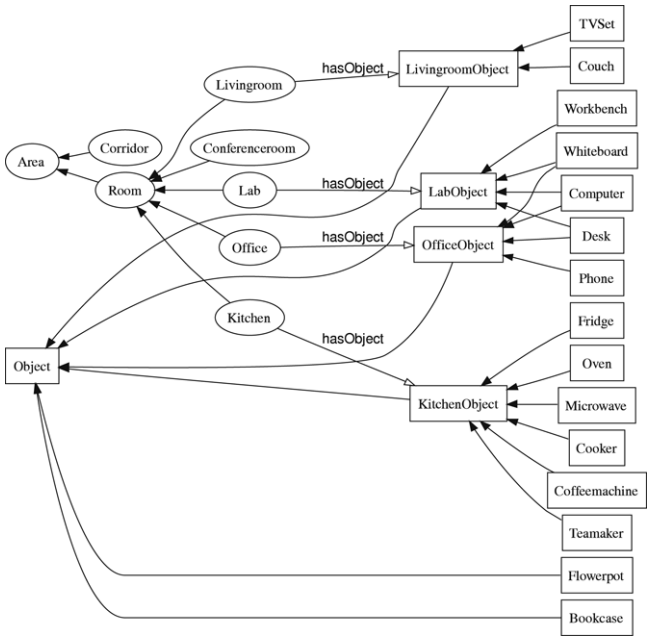

Fig. 9. Illustration of a part of the commonsense ontology of an indoor office environment. Edges with solid arrow heads denote the taxonomical 'is–a' relation.

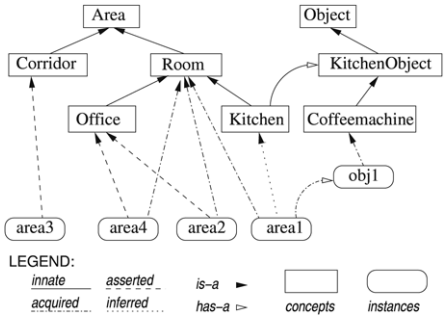

Fig. 10. Combining different types of knowledge in the conceptual map.

The conceptual knowledge is encoded in an OWL-DL ontology of an indoor office environment (see Fig. 9). It describes taxonomies (*is–a* relations) of room types and typical objects found in an office environment. In line with the way humans categorize space, our ontology defines room types on the basis of the objects they contain — represented by *has–a* relations. The conceptual ontology in the TBox constitutes *innate* knowledge, since it has been predefined and cannot be changed during run-time. However, while the robot operates in its environment, the sensors constantly *acquire* new information, which is then represented as instance knowledge in the ABox. Through situated dialogue the robot can obtain *asserted* knowledge from its user. A description-logic reasoner can then fuse this knowledge in order to *infer* new knowledge about the world that is neither given verbally nor actively perceived.

5.5. Spatial knowledge processing

Below, we describe the information processing principles for these individual types of knowledge in more detail. Fig. 10 shows an example of how spatial knowledge from different sources converges in the conceptual map.

5.5.1. 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 category *area*. Furthermore, as soon as reliable information about the semantic classification of an area is available (cf. Section 5.3), this is reflected in the conceptual map by assigning the area's instance a more specific category (*room* or *corridor*). Information about recognized objects stemming from the vision subsystem (cf. Section 4.3) 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 ABox. Moreover, the object's instance and the instance of the area where the object is located are related via the *hasObject* relation.

5.5.2. 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.

5.5.3. Innate conceptual knowledge

We have handcrafted an ontology (Fig. 9) that models conceptual commonsense knowledge about an indoor office environment. On the top level of the conceptual taxonomy, there are the two general concepts *area* and *object*. The concept *area* can be further partitioned into the concepts *room* and *corridor*. The basic-level categories, i.e. the subconcepts of *room*, are characterized by the *object* instances that are found there, as represented by the *hasObject* relation.

5.5.4. Inferred knowledge

Based on the knowledge representation in the ontology, our system uses a description-logic-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 *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 *living room*. Conversely, if an area is classified as *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 classifications 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 [38]. 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.

6. Situated dialogue

In this section, we discuss the functionality which enables a robot to carry out a natural language dialogue with a human.

A core characteristic of our approach is that the robot builds up a semantic representation for each utterance. The robot interprets it against the dialogue context, relating it to previously mentioned objects and events, and to previous utterances in terms

of "speech acts" (dialogue moves). Since dialogues in human–robot interaction are inherently situated, the robot also tries to ground the utterance content in the situated context — including past and current visuo-spatial contexts (reification of visuo-spatial references), and future contexts (notably, planned events and states). Below we highlight several aspects; for more details, we refer the reader to [22,24].

Speech recognition yields a string-based representation for spoken input, which is subsequently parsed using the Combinatory Categorial Grammar (CCG) parser of OpenCCG [39]. The parser analyzes the utterance syntactically and derives a semantic representation [40]. The semantic representation is a logical form in which propositions are assigned ontologically sorts, and related along typed relations (e.g. "Location", "Actor").

The logical forms yielded by the parser are interpreted further, both within the dialogue system and against information about the situated context. Objects and events in the logical form are related against the preceding context (co-reference resolution), as is the dialogue move of the utterance. The resulting dialogue model is similar to that proposed in [41,19]. The robot also builds up a temporal-aspectual interpretation for events, relating it to preceding events in terms of how they temporally and causally follow on each other [42]. In combination with the dialogue model this is closely related to [21].

7. System integration

Our approach has been implemented as an integrated system, running on an ActivMedia PeopleBot mobile robot platform. In this section, we discuss the integration of the components presented in the earlier sections. We focus on what integration brings us in terms of achieving a better understanding of sensory signals, i.e. one that is more complete and more appropriate for interacting with humans; particularly, given that sensory information usually only provides a partial, potentially noisy view of the environment.

For perceiving the environment, the robot is equipped with a SICK laser range finder, and a pan-tilt-zoom camera (cf. Fig. 1). As discussed in Section 4, the laser scanner is used for the metric map creation, for the semantic classification of places, and for people following. The PTZ camera is used for object detection. The software for controlling the robot, the individual components that contribute to the multi-layered spatial map, and the dialogue system run on a number of computers, including an on-board machine, interconnected using a wireless network. Additionally a speech recognition software connected to a bluetooth headset, and a text-to-speech engine connected to the robot's built-in speakers are used for spoken interaction between the robot and its user.

Below we discuss system integration and its effects on complementation and robustness, illustrated on several core capabilities in a "home tour scenario" [43,44]. During a guided tour, a user takes the robot around the house. The advantage of such a scenario is that the user can tell the robot where specific rooms and objects are, or instruct the robot to perform particular tasks. The robot concurrently and incrementally constructs an internal representation of the spatial organization of its environment. For this, the robot needs to be able to recognize areas, and boundaries between areas to build up organization. But it also needs to overcome the drawbacks that such a scenario brings. One obvious problem is for instance that the user constantly occupies large parts of the robot's field of view while guiding the robot around.

For each of these aspects, we present how the functionalities of the individual components interact to give rise to these capabilities, and how their integration improves robustness and completeness of interpretation.

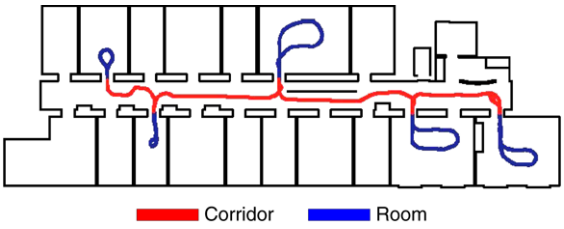

Fig. 11. Trajectory followed by the robot to train the classifier for distinguishing between corridors and rooms. The different places are depicted with distinct colors.

7.1. Place classification

First of all, in such a scenario where the robot continuously interacts with a user and is facing her/him most of the time, the information content of the laser sensor input suffers as the user occupies a large part of the field of view. Secondly, the original approach to laser-based place classification presented in [30] relies on laser observations covering a 360° field of view. However, our robot is equipped with only one laser covering 180° in front of the robot.

The first step we take for solving these problems is to simulate a rear-view laser scanner by ray tracing in a local occupancy grid. When classifying a pose of the robot during a trajectory, we simulate the rear beams by determining the end points of laser beams that hit some occupied cell in the occupancy grid. For each simulated beam that does not hit any object in the occupancy grid, we calculate its value using an interpolation between the values of their (known) neighboring beams at both sides.

The second step we take in order to achieve robust place classifications even while interacting with a user is to not try to classify every possible geometric coordinate inside an area, but instead to classify only the nodes in the navigation graph. For determining a robust classification of a navigation node we compute the majority vote of consecutive classifications of that node, as described in Section 5.2.

In order to test our method for classification of rooms and corridors, we used trajectories in different floors of the CAS building at KTH for training and evaluating the classifier. To train our classifier we used a trajectory of the 6th floor (Fig. 11). The robot was then moved to the 7th floor of the same building, which contained a similar structure. In this new floor, we classified two different trajectories in opposite directions. The classification rates of all the poses of the robot during its movement ranged from 93.18% to 96.8%.

7.2. Object recognition

The user's presence not only disturbs the laser-based place classification, but also the camera-based object recognition. In our case, the camera is mounted on a pan-tilt unit and could have been used to actively look for objects and build a metric map using visual information while following the user. However, most of the time the camera only sees the user and not the environment. Therefore, we opted for giving the user the possibility to instruct the robot to "have a look around."

For object recognition we tested the two different methods presented in Section 4.3. The first one uses unsegmented images and allows the system to expand its database of objects in a much easier way. It is a matter of making sure that the object is in the field of view of the camera and acquiring an image. The problem is that depending on how close the robot gets to the objects the background might dominate the image. In that case what the robot recognizes is not only the object but rather the whole scene including the object and its surroundings.

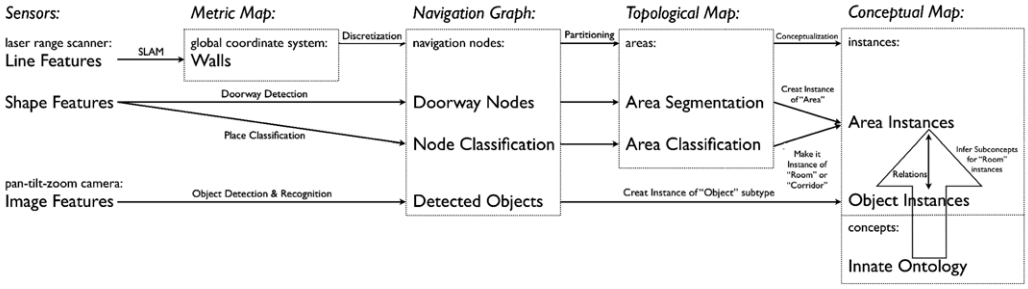

Fig. 12. Process diagram showing convergence on consistent interpretation across levels of spatial interpretation.

Additionally, we carried out experiments using the second approach of Section 4.3, which uses segmented images. This makes the acquisition of knowledge about new objects more complicated. In [32], it is suggested to use image differencing combined with morphological operations to segment the image. For this to work the user has to present the robot with two images, one with and one without the objects. While this might be feasible for small objects, it is clearly a problem for objects like a TV.

7.3. Conceptualizing areas

What does it mean to recognize an area? What *defines* an area? In SLAM-based approaches, the notion of area roughly corresponds to an "enclosed space." Observed linear structures are interpreted as walls, delineating an area. This yields a purely geometrical interpretation of the notion of area, based on its perceivable physical boundaries. Doorways are regarded as transitions between distinct topological areas. Although this is already a suitable level of abstraction, it is not yet sufficient for discriminating between areas. Another observation from human spatial cognition is that humans tend to categorize space not only geometrically, but also functionally. This functionality is often a result of the different objects inside an area, like home appliances or furniture, that afford these functions. In order to achieve a functional-geometric interpretation, a robot thus has to integrate its knowledge about distinct topological areas with its knowledge about the presence of certain objects.

Fig. 12 illustrates the way in which the modules in our system (cf. Section 4) contribute to the individual layers of our conceptual spatial map (cf. Section 5), and how additional pieces of knowledge are combined to achieve a more complete conceptualization of space.

In our experiment [44], this is illustrated when the user repeatedly asks the robot about where it thinks it is. At first, the only thing the robot knows is that the current area is classified as *room*, and answers "I am in a room." However, after the robot has recognized the couch and the TV set in the current room, its ontological reasoning capabilities can infer the appropriate subconcept *living room*. Hence the robot can produce an answer that contains more information: "I am in the living room."

Our approach thus not only creates a qualitative representation of space that is similar to the way humans perceive it. It also serves as a basis for successful dialogues by allowing the robot to successfully refer to spatial entities using natural language expressions [45].

Finally, experiments highlighted the need for non-monotonic reasoning, that is, knowledge must not be written in stone. As erroneous acquired or asserted knowledge will otherwise lead to irrecoverable errors in inferred knowledge. One solution to this would be to not maintain spatial conceptual knowledge in the ABox of the monotonic DL reasoner. Instead only the most recent, ideally reliable and stable, knowledge about area classes and object positions should be propagated to the conceptual map when it is needed, e.g. when generating or resolving linguistic referring expressions.

7.4. Recognizing transitions between areas

Boundaries segment space into different areas, i.e. different topological regions. Gateways, like doors, are a typical type of transitional boundary.

The door detector used in this work finds doors by looking for narrow openings in the laser data, as explained in Section 4.2. We tested this approach and it produced some false positives, as there are many gaps in the environment with similar width as doorways. To counteract this, we only accept openings that the robot itself passes through. This significantly reduces the number of false positives. We additionally used the method presented in [46,24] for clarification dialogue to handle the few false doors that are still found. The disadvantage with this door detection scheme is that the robot is unable to detect doors where it has not traveled so far. In some cases it would be of value to be able to detect doors further away to allow reasoning about unexplored space. An example of this is when the user refers to a room when walking down the corridor with the robot as explained in [38].

Boundaries need not be explicit, though. The transition from a hall to a corridor may just be indicated by a "pronounced" difference in geometrical shape (rectangular to elongated), whereas the segmentation of e.g. a living kitchen into dining and cooking areas is based purely on functional aspects. In our presented system we only partition navigation nodes into areas based on detected doorways. This is an obvious disadvantage in the mentioned cases where a space enclosed by walls and doors itself can be considered to consist of several, functionally and/or geometrically distinct, regions. We currently investigate how the knowledge about present objects and their alignment, along with the shift in probability distributions of the laser-based place classifier along a trajectory can serve as additional cues for segmenting space into regions.

8. 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 information needed for each level stems from different modalities, including a laser sensor, a camera, and a natural language processing system. The complete system was integrated and tested on a mobile robot including a framework for spoken interaction. An assessment of the integrated system shows that our approach is able to provide a high level of human–robot communication and conceptual representation.

References

[1] A. Stevens, P. Coupe, Distortions in judged spatial relations, Cognitive Psychology 10 (1978) 422–437.
[2] T. McNamara, Mental representations of spatial relations, Cognitive Psychology 18 (1986) 87–121.
[3] A.G. Cohn, S.M. Hazarika, Qualitative spatial representation and reasoning: An overview, Fundamenta Informaticae 46 (2001) 1–29.
[4] S.C. Hirtle, J. Jonides, Evidence for hierarchies in cognitive maps, Memory and Cognition 13 (1985) 208–217.

[5] R. Brown, How shall a thing be called?, Psychological Review 65 (1) (1958) 14–21.
[6] E Rosch, Principles of categorization, in: E. Rosch, B. Lloyd (Eds.), Cognition and Categorization, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1978, pp. 27–48.
[7] B. Kuipers, The Spatial semantic hierarchy, Artificial Intelligence 119 (2000) 191–233.
[8] B. Krieg-Brückner, T. Röfer, H.-O. Carmesin, R. Müller, A taxonomy of spatial knowledge for navigation and its application to the Bremen autonomous wheelchair, in: C. Freksa, C. Habel, K.F. Wender (Eds.), Spatial Cognition, in: Lecture Notes in Artificial Intelligence, vol. 1404, Springer Verlag, 1998, pp. 373–397.
[9] S. Vasudevan, S. Gachter, M. Berger, R. Siegwart, Cognitive maps for mobile robots an object based approach, in: Proc. of the IEEE/RSJ IROS 2006 Workshop: From Sensors to Human Spatial Concepts, Beijing, China, 2006.
[10] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernández-Madrigal, 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.
[11] P. Beeson, M. MacMahon, J. Modayil, A. Murarka, B. Kuipers, B. Stankiewicz, Integrating multiple representations of spatial knowledge for mapping, navigation, and communication, in: AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants, Stanford, CA, USA, 2007.
[12] A. Diosi, G. Taylor, L. Kleeman, Interactive SLAM using laser and advanced sonar, in: Proc. of the IEEE Int. Conference on Robotics and Automation, ICRA, Barcelona, Spain, 2005.
[13] O. Martínez Mozos, A. Rottmann, R. Triebel, P. Jensfelt, 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.
[14] S. Friedman, H. Pasula, 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.
[15] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun, Experiences with an interactive museum tour-guide robot, ArtificialIntelligence 114 (1–2).
[16] R. Siegwart, et al., Robox at expo.02: A large scale installation of personal robots, Robotics and Autonomous Systems 42 (2003) 203–222.
[17] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, R. Nakatsu, Robovie: An interactive humanoid robot, Int. J. Industrial Robotics 28 (6) (2001) 498–503.
[18] A. Haasch, S. Hohenner, S. Huewel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G.A. Fink, J. Fritsch, B. Wrede, G. Sagerer, Biron - the Bielefeld robot companion, in: Proceedings of the International Workshop on Advances in Service Robotics, Fraunhofer IRB Verlag, Stuttgart, Germany, 2004, pp. 27–32.
[19] J. Bos, E. Klein, T. Oka, Meaningful conversation with a mobile robot, in: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics EACL'03, Budapest, Hungary, 2003.
[20] O. Lemon, A. Bracy, A. Gruenstein, S. Peters, A multi-modal dialogue system for human–robot conversation, in: Proceedings of the Second Meeting of the North American Chapter of the Association of Computational Linguistics NAACL 2001, Pittsburg PA, 2001.
[21] C. Sidner, C. Kidd, C. Lee, N. Lesh, Where to look: A study of human–robot engagement, in: Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI, 2004, pp. 78–84.
[22] G. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, N. Hawes, Incremental, multi-level processing for comprehending visually situated dialogue in human–robot interaction, in: Proceedings of the Symposium on Language and Robotics, LANGRO 2007, Aveiro, Portugal, 2007.
[23] D. Traum, S. Larsson, The information state approach to dialogue management, in: J. van Kuppevelt, R. Smith (Eds.), Current and New Directions in Discourse and Dialogue, Kluwer Academic Publishers, 2003.
[24] G.-J.M. Kruijff, H. Zender, P. Jensfelt, H.I. Christensen, Situated dialogue and spatial organization: What, where ... and why?, International Journal of Advanced Robotic Systems 4 (2). Special section on Human and Robot Interactive Communication.
[25] H. Zender, P. Jensfelt, G.-J.M. Kruijff, Human- and situation-aware people following, in: Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication RO-MAN 2007, Jeju Island, Korea, 2007, pp. 1131–1136.
[26] J. Folkesson, P. Jensfelt, 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.
[27] P. Jensfelt, Approaches to mobile robot localization in indoor environments, Ph.D. Thesis, Signal, Sensors and Systems (S3), Royal Institute of Technology, SE-100 44 Stockholm, Sweden 2001.
[28] A. Tapus, G. Ramel, L. Dobler, R. Siegwart, Topology learning and recognition using Bayesian programming for mobile robot navigation, Intelligent Robots and Systems, 2004. IROS 2004. Proceedings. 2004 IEEE/RSJ International Conference on 4.
[29] D. Anguelov, D. Koller, E. Parker, S. Thrun, Detecting and modeling doors with mobile robots, in: Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on 4.
[30] O. M. Mozos, C. Stachniss, W. Burgard, Supervised learning of places from range data using adaboost, in: Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, 2005, pp. 1742–1747.
[31] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[32] S. Ekvall, D. Kragic, P. Jensfelt, Object detection and mapping for service robot tasks, Robotica, International Journal of Information, Education and Research in Robotics and Artificial Intelligence 25 (2) (2007) 175–187.
[33] S. Ekvall, D. Kragic, Receptive field cooccurrence histograms for object detection, in: Proc. IEEE/RSJ International Conference Intelligent Robots and Systems, IROS'05, 2005, pp. 84–89.
[34] D.G. López, Combining object recognition and metric mapping for spatial modeling with mobile robots, Master's Thesis, Royal Institute of Technology, July 2007.
[35] H.P. Moravec, Sensor fusion in certainty grids for mobile robots, AI Magazine 9 (1988) 61–74.
[36] J.C. Latombe, Robot Motion Planning, Academic Publishers, Boston, MA, 1991.
[37] P. Newman, J. Leonard, J. Tardós, J. Neira, Explore and return: Experimental validation of real-time concurrent mapping and localization, in: Proceedings of the 2002 IEEE International Conference on Robotics and Automation, ICRA 2002, Washington, DC, USA, 2002, pp. 1802–1809.
[38] E.A. Topp, H. Hüttenrauch, H. Christensen, 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, 2006.
[39] J. Baldridge, 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.
[40] J. Baldridge, G.-J.M. Kruijff, 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.
[41] N. Asher, Lascarides, Logics Of Conversation, Studies in Natural Language Processing, Cambridge University Press, 2003.
[42] G. Kruijff, M. Brenner, Modelling spatio-temporal comprehension in situated human–robot dialogue as reasoning about intentions and plans, in: AAAI Spring Symposium on Intentions in Intelligent Systems, 2007.
[43] E.A. Topp, 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.
[44] H. Zender, P. Jensfelt, Óscar Martínez Mozos, G.-J.M. Kruijff, W. Burgard, An integrated robotic system for spatial understanding and situated interaction in indoor environments, in: Proc. of the Twenty-Second Conference on Artificial Intelligence, AAAI-07, Vancouver, British Columbia, Canada, 2007, pp. 1584–1589.
[45] H. Zender, G.-J.M. Kruijff, Towards generating referring expressions in a mobile robot scenario, in: Language and Robots: Proceedings of the Symposium, Aveiro, Portugal, 2007, pp. 101–106.
[46] G.-J.M. Kruijff, H. Zender, P. Jensfelt, 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, pp. 282–288.

H. Zender is a PhD student researcher at the Language Technology Lab of the German Research Center for Artificial Intelligence (DFKI). His research interests are linguistic aspects of spatial cognition and spatial knowledge representations for human–robot interaction. He received his Diploma degree in Computational Linguistics from Saarland University in 2006.

O. Martínez Mozos is a Ph.D. student at the lab of Autonomous Intelligent Systems headed by Wolfram Burgard at the University of Freiburg in Germany. His areas of interest lie on mobile robotics, artificial intelligence, and pattern recognition. In 2005, he received a M.Sc. in applied Computer Science at the University of Freiburg. In 1997 he completed a M.Eng. in Computer Science at the University of Alicante in Spain.

P. Jensfelt is an assistant professor at the Centre for Autonomous Systems at the Royal Institute of Technology, Stockholm, Sweden. He received his M.Sc. in Engineering Physics in 1996 and Ph.D. in Automatic Control in 2001. His research interests include mapping and localization, mobile robotics, and system integration.

G.-J. Kruijff is a Senior Researcher at the DFKI Language Technology Lab, where he leads efforts in the area of "cognitive systems." His research focuses on developing "talking robots". He is particularly interested in developing theories and has implemented architectures for cognitive robots to understand, and produce, situated dialogue with human users — in other words, how do we make talking robots? He has over 90 refereed conference papers and articles in human–robot interaction, and formal and computational linguistics. He is a member of IEEE.

W. Burgard is a professor at the Department of Computer Science at the University of Freiburg, where he heads the lab for Autonomous Intelligent Systems. He studied Computer Science at the University of Dortmund and received his Ph.D. degree in Computer Science from the University of Bonn in 1991. His research focuses on mobile robotics and system integration.