Situated dialogue and understanding spatial organization: Knowing what is where and what you can do there

Geert-Jan M. Kruijff, Hendrik Zender
Language Technology Lab
DFKI GmbH
Saarbrücken, Germany
{gj,hendrik.zender}@dfki.de

Patric Jensfelt, Henrik I. Christensen
Center for Autonomous Systems
Royal Institute of Technology (KTH)
Stockholm, Sweden
{patric,hic}@nada.kth.se

Abstract— **The paper presents an HRI architecture for human-augmented mapping. Through interaction with a human, the robot can augment its autonomously learnt metric map with qualitative information about locations and objects in the environment. The system implements various interaction strategies observed in independent Wizard-of-Oz studies. The paper discusses an ontology-based approach to representing and inferring 2.5-dimensional spatial organization, and presents how knowledge of spatial organization can be acquired autonomously or through spoken dialogue interaction.**

I. Introduction

More and more robots find their way into environments where their primary purpose is to interact with humans to help and solve a variety of service-oriented tasks. Particularly if such a service robot is mobile, it needs to have an understanding of the spatial and functional properties of the environment in which it operates. The problem we address is how a robot can acquire an understanding of the environment so that it can autonomously operate in the environment, and talk about it with a human. We present an architecture that provides the robot with this ability through a combination of human-robot interaction and autonomous mapping techniques. The architecture captures various functions that independently conducted Wizard-of-Oz studies have observed to be necessary for such a system.

The main issue we must solve is how we can establish a correspondence between how a human perceives spatial and functional aspects of an environment, and what the robot autonomously learns as a map. Most existing approaches to robot map building, or Simultaneous Localization And Mapping (SLAM), use a metric representation of space. Humans, though, have a more qualitative, topological perspective on spatial organization [1]. We adopt an approach in which we build a multi-level representation of the environment, combining metrical maps and topological graphs (as an abstraction over metrical information), like [2]. We extend these representations with structural descriptions that capture aspects of spatial and functional organization. The robot obtains these descriptions either through interaction with a human, or through inference combining its own observations (*I see a coffee machine*) with ontological knowledge (*Coffee machines are usually found in kitchens, so this is likely to be a kitchen!*). We store objects in the spatial representations, and so associate the functionality of a location with that of the functions of the objects present there.

Following [3], [4] we talk about *Human-Augmented Mapping* (HAM) to indicate the active role that human-robot interaction plays in the robot's acquisition of qualitative spatial knowledge. In §II we discuss various observations that independently performed Wizard-of-Oz studies have made on typical interactions for HAM scenarios, and we indicate which we will be able to handle. In §III we present our approach to spatial represention and the structural descriptions it uses to encode knowledge about spatial and functional aspects of the environment. We discuss its implementation in an HRI architecture in §IV, illustrating it on examples in §IV and §V. The paper closes with conclusions.

II. Observations on HAM

Various Wizard-of-Oz studies have investigated the nature of human-robot interaction in HAM. [4] discuss a study into how a human presents a familiar indoor environment to a robot, to teach the robot more about the spatial organization of that environment. [5] study the different types of dialogues found when interacting with a robot wheelchair. Below we discuss several important insights these studies yield.

The experimental setup in [4] models a typical guided-tour scenario. The human tutor guides the robot around and names places and objects. One result of the experiment is that tutors employ many different strategies to introduce new locations. Besides naming whole rooms ("This is the kitchen" referring to the room itself) or specific locations in rooms ("This is the kitchen" referring to the cooking area), a frequently observed strategy was to name specific locations by the objects found there ("Here is the coffee machine"). Any combination of these individual strategies could be found during the experiments. Moreover, it has been found that subjects only name those objects and locations that they find interesting or relevant, thus *personalizing* the representation of the environment that the robot constructs.

In the study presented in [5] the subjects were seated in a robot wheelchair and were asked to guide it around using verbal commands. This setup has a major impact on the data collected. The tutors must use verbal commands containing deictic references in order to steer the robot. First of all, the perspective of the human tutor is identical to the one of the robot. Deictic references can thus be mapped one-to-one to the robot's frame of reference. One interesting finding is that people tend to name areas that are only passed by. This can either happen in a 'virtual tour' when giving route directions or in a 'real guided-tour' ("Here to the right of me is the door to the room with the mailboxes."). A robust conceptual mapping system must therefore be able to handle information about areas that have not yet been visited.

In §III we discuss how we deal with the above findings, combining information from dialogue and ontologies.

III. Spatial Organization

If we want a robot to be able to understand and talk about spatial organization, we must close the gap between the different ways humans and robots think of spatial organization. We discuss our approach to representing the spatial and functional aspects of an environment at multiple levels of abstraction, closing this gap. *Spatial aspects* cover the organization of an environment in terms of connected areas and gateways. We associate *functional aspects* with an area on the basis of objects present in it. Through dialogue we can build, query, and clarify these representations, and we point out how they are used in carrying out tasks.

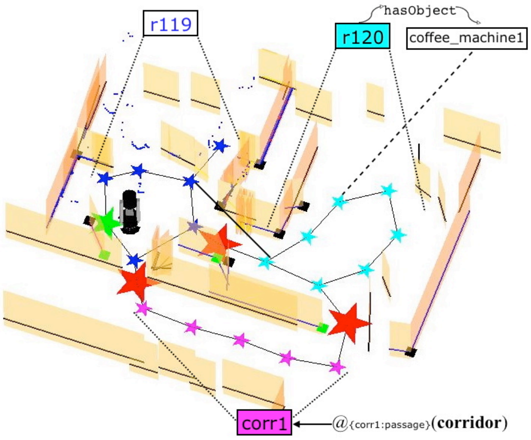

Fig. 1. Multi-level environment representation

A. Representing the environment

We represent the spatial organisation of an (indoor) environment at three levels (Figure 1). At the lowest level, we have a *metric map*, capturing observed spatial structures in the environment, e.g. walls. This map does not explicitly represent free space like an occupancy grid does. Therefore, when driving around the environment, the robot constructs a *route graph* on top of the metric map to indicate places it can go to. A route graph is a connected graph in which nodes represent areas and gateways between areas, and edges indicate accessibility. To anchor the route graph in the metric map, we associate a metrical coordinate with each route graph node. Finally, we create a *conceptual map* by abstracting over the route graph. We subsume sets of route graph nodes into areas. The boundaries of an area are constrained by occurrences of gateways, such as doors, that can be observed from laser range data. The conceptual map is then a connected graph from nodes that represent entire areas, and gateways between areas, with edges again representing accessibility. This map is a first approximation of a topological perspective on metric data.

We use the conceptual map as a qualitative level for describing the environment. It is at this level that we associate structural descriptions with areas. Here, we first focus on the nature of these structural descriptions, and how we can use ontological knowledge of objects and areas to enrich such descriptions. Below we show how we obtain structural descriptions through human-robot interaction.

A structural description is an ontologically richly sorted, relational structure, formalized in a description logic-like framework [6]. Example 1 illustrates such a description.

(1) "The office of GJ, having one desk"

$$
\begin{aligned}
&@_{\{r1:\mathsf{room}\}}(\textbf{office}\\
&\quad \&\ \langle Realization\rangle\textbf{concrete}\\
&\quad \&\ \langle Owner\rangle(g1 : person\ \&\ \textbf{GJ})\\
&\quad \&\ \langle HasObject\rangle(d1 : furniture\ \&\ \textbf{desk}\\
&\qquad\quad \&\ \langle Delimitation\rangle\textbf{unique}\\
&\qquad\quad \&\ \langle Quantification\rangle\textbf{specific\_singular}))
\end{aligned}
$$

The structural description in Example 1 consists of several, related *elementary predicates* (EPs). One type of EP represents an identifiable spatial aspect as a proposition with a handle: $@_{\{r1:\mathsf{room}\}}(\textbf{office})$ means that $r1$ is an **office**, which is a *room*. Another type of EP states relations between aspects as modal relations, e.g. $@_{\{r1:\mathsf{room}\}}\langle Owner\rangle(g1 : person\ \&\ \textbf{GJ})$ means the owner of the room $r1$ is a *person* called GJ. Within each EP we can have semantic features, e.g. room $r1$ has a concrete realization.

A structural description captures what the robot knows about an area. This knowledge need not always be complete. The robot may have observed only part of an area and the objects therein, and as we already pointed out in §II, humans need not necessarily convey complete information about a room either. The robot thus needs to be able to create a more complete structural description on the basis of only partial information. For this, we use ontological knowledge of spatial and functional aspects. §IV-D provides explanations and an example of the method used.

B. Human-Augmented Mapping

In a typical HAM scenario, a human tutor takes the robot on a guided tour of the environment. He or she then presents and introduces *locations* ("This is the kitchen") and *objects* ("This is the coffee machine"). The issue here is how we can use this information to augment our spatial representation.

From language processing, we obtain a representation of the semantics of an utterance. Depending on the kind of utterance (e.g. question, command, assertion), we decide in what modalities we need to process this content further. A prototypical utterance in a HAM scenario makes an assertion about the kind of location the current area is. In this case, we create a structural description from the semantics of the utterance, and try to update the conceptual map with it. If the conceptual map does not yet contain a description for the current area, we use the description we just obtained.

Else, if the area has a description, and what we just got is inconsistent with that description, the robot points this out:

(2) H.1 "This is the kitchen."
R.2 $\langle New\ area:\ @_{\{k1:\text{location}\}}\textbf{kitchen}\rangle$ "Okay"

(3) H.1 "This is the corridor."
R $\langle Robot\ has\ not\ spotted\ gateway\ to\ the\ corridor\rangle$
R.2 "I am sorry. I thought this was the hallway."
H.3 "No, this is the corridor."

If the human makes an assertion about an object, we take several steps. First, the vision system learns a model of the object, labelling the model with the structural description for the object [7]. Next, we anchor the occurrence of the object and its description at the different levels of the spatial representation: in the route graph (at the node nearest its position), in the conceptual map (adding more information to the structural description of the area) and in the ontological representation (an instance of the object's type is created and related to the individual that represents the current area). By using the same structural description for an object as label for its visual model and as pointer in the spatial representation, we can maintain associations across these representations.

The realization feature provides a way to treat assertions of the human tutor about objects and locations that the robot has not yet observed (cf. §II). We deal with this by marking areas and objects as either *concrete* or *abstract* (Example 1). Individuals in the ontology that are not anchored to either a location in the topological graph or an object perceived by the visual recognition are marked as abstract. Only entities that the robot has actively perceived are considered concrete. We can use these abstract entities as cue for the robot to explore unknown places during autonomous exploration.

C. Answering questions about locations and objects

Given the robot's conceptual map, we can at any given time ask the robot about where it thinks it is. If a structural description of the current room has been given before, the robot retrieves this information from the conceptual map. The description of the area is then returned to the dialogue system, which generates a proper utterance to convey the given information. If the conceptual map does not contain a description for the current area, we rely on ontological reasoning (cf. §IV-D) to infer the type of the current location.

If asked about the location of an object, we retrieve occurrences of the desired object. We then generate a structural description of the room where that object can be found, and provide this description to the dialogue system to convey it.

(4) H.1 "Where is the dishwasher?"
R.2 "It is in the kitchen."

D. Clarification

Existing dialogue-based approaches to HAM usually implement a *master/slave* model of dialogue: the human speaks, the robot listens (e.g. [8]). However, situations naturally arise in which the robot needs to take the initiative, e.g. to clarify an issue with the human. This is one form of *mixed-initiative* interaction, enabling a robot to recognize when help is needed from a human, and learn from this interaction [9]. Situations that may require clarification are e.g. uncertainty in automatic classification: Doors provide important knowledge about spatial organization, but are difficult to recognize robustly and reliably. Clarification dialogues can help to improve the quality of the spatial representation the robot constructs, and to increase the robot's robustness in dealing with uncertain information.

We have extended an approach to processing clarification questions in multi-modal dialogue systems. For space reasons, we refer the reader to [10] for technical details. The basic idea is to allow for any modality to raise an issue, formulated as a *(clarification) question* about objects ("What is this thing near me?") or about the truth of a proposition ("Is there a door here?"). When a modality raises an issue, mediation stores the issue on a list of open issues, and requests another modality to help resolving the issue. For example, when mapping is unsure about the presence of a door in a given location, an issue is raised which is then addressed through interaction with the human: "Is there a door here?" "No." Once dialogue analysis has related the answer to the question, both are sent back to the mediator to inform the modality that raised the issue.

E. Carrying out tasks

Guiding the robot around an environment is only one step in working with a service robot. The main purpose of a service robot, and of most domestic robots, is to carry out tasks requested by a human. The multi-level represention of the environment we build up provides an important basis for that. We can combine knowledge about what objects are needed to perform particular actions, with the knowledge where they are.

(5) H.1 "Could you get me a coffee?"
R $\langle Robot\ infers:\ \text{coffee-coffee machine}\rangle$
R $\langle$*retrieves location of coffee machine*$\rangle$
R.2 "Yes."
R $\langle carries\ out\ task\rangle$

The retrieval of the location where the object can be found is performed by first doing a lookup in the conceptual map. If that fails, ontology reasoning is applied (§IV-D). For more on how we deal with requests, see [11].

IV. Implementation

We have implemented the approach of §III in a distributed architecture which integrates different sensorimotoric and cognitive modalities. We have developed the architecture to enable an ActivMedia PeopleBot to move about in an indoor environment, and have a situated dialogue with a human about various visual and spatial aspects of a situation.

Figure 2 shows the relevant aspects of the architecture. We have subsystems for communication, spatial localization & mapping, and visual processing. We use a BDI-subsystem (Belief, Desire, Intention) to mediate between subsystems. By this we mean that beliefs provide a common ground

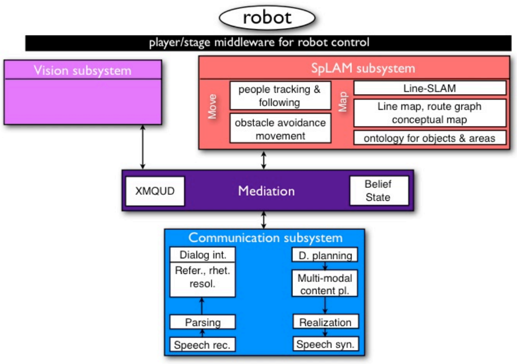

Fig. 2. The architecture

between different modalities, rather than being a layer on top of the different modalities. Beliefs provide a means for cross-modal information fusion, in its minimal form by co-indexing references to information in individual modalities [12]. In mediation we decide what modalities should further process linguistically conveyed information, and how to handle requests for clarifying issues that have arisen.

A. The communication subsystem

The communication subsystem consists of several components for the analysis and production of natural language. It has been implemented as a distributed architecture using the Open Agent Architecture [13]. On the analysis side, we use the Nuance speech recognition engine[1] with a domain-specific speech grammar. The string-based output of Nuance is then parsed with OpenCCG[2]. OpenCCG uses a combinatory categorial grammar [14] to yield a representation of the linguistic meaning for the recognized string/utterance [6]. In dialogue analysis we relate the linguistic meaning of an utterance to the current dialogue context, in terms of how it rhetorically and referentially relates to preceding utterances, yielding an updated model of the dialogue context [15], [8].

To produce flexible, contextually appropriate interaction we use several levels of dialogue planning. Based on a need to communicate, arising from the current dialogue flow or from another modality, the dialogue planner establishes a communicative goal. We then plan the content to express this goal, possibly in a multi-modal way using non-verbal (pose, head moves) and verbal means. During planning we can inquire the models of the situated context (e.g. dialogue context, visually scene) to ensure the plan is contextually appropriate. We realize verbal content using the OpenCCG realizer, which generates a string for the utterance, and then synthesize this string using text-to-speech[3].

B. Spatial Localization & Mapping

The subsystem for SLAM uses a feature based representation where the main features are lines, typically corresponding to walls in the environment. The underlying feature representation is flexible and other types of features can easily be incorporated [16]. The basis for integrating the feature observations is the extended Kalman filter (EKF).

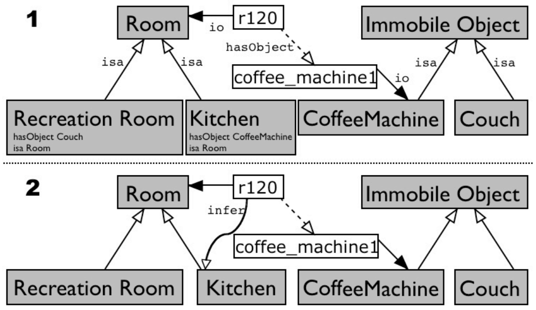

Fig. 3. Ontological reasoning

A feature based map is rather sparse and only captures structures that fit the predefined feature description (e.g. lines). One cannot distinguish free space from areas where the structures do not fit the feature model. For this we use a technique as in [17] and build a route graph (cf. §III) while the robot moves around. When the robot has moved a certain distance, a node is placed in the graph at the current position of the robot. Whenever the robot moves between two nodes, these are connected in the graph.

We build the conceptual map automatically from the route graph by labeling the nodes into different areas and thus partitioning it. Our strategy rests on the simple observation that the robot passes a door to move between rooms. Whenever the robot passes a door a node marked as a door is added to the navigation graph and consecutive nodes are given a new area label. Currently, door detection is simply based on detecting when the robot passes through a narrow opening. The fact that the robot has to pass through an opening removes many false doors that would result from simply looking for narrow openings. However, this alone will still lead to some false doors in cluttered rooms. We use *loop closing* to spot inconsistencies [10] arising from falsely recognized doors, and then trigger a clarification dialogue.

C. Vision

The vision subsystem provides visual scene understanding based on three cues: identity, color, and size of objects in the scene. We use an implementation of SIFT (Scale Invariant Feature Transform) features [18] and visual codebooks [19] to recognize object identity, and bounding boxes to establish size and color. The subsystem maintains a qualitative interpretation of the spatial organization of objects in the scene, based on topological and projective spatial relations [20].

D. Ontological reasoning

The ontological representation is part of the conceptual map. We use ontological reasoning to fuse knowledge about types and instances of types in the world. We have built a common-sense ontology of an indoor (office) environment as an *OWL ontology*[4], having *classes*, *individuals* (instances

[1] http://www.nuance.com
[2] http://openccg.sf.net
[3] http://mary.dfki.de
[4] http://www.w3.org/TR/owl-guide

of classes) and *properties* (binary relations between individuals). The ontology covers types of locations and typical objects. A priori, as the robot has not yet learnt anything, the ontology does not contain any individuals. We create individuals as the robot discovers its environment. For each new area, a new instance of class `Room` is created. When the robot is in a room, and is shown or visually detects an object, we create a new instance of the corresponding `Object` subclass, and relate the object's instance and the room's instance using a `hasObject` property.

We use the RACER/JRacer system[5] to reason over *TBoxes* (terminological knowledge/classes in our ontology) and *ABoxes* (assertional knowledge/instances). We use *assertions* about instances and relations to represent knowledge that the robot learns as it discovers the world. This includes explicit introductions by the tutor or autonomously acquired information. We do not change the TBox at runtime.

If the conceptual map does not contain a structural description that is relevant for the current task (cf. §III-C and §III-E) we try to infer the missing information. We use ABox retrieval functions as a first reasoning attempt. The reasoner checks if it can infer that an instance is consistent with the given description. If so, this instance is taken. Else, we use TBox reasoning as a second attempt to resolve uncertainties, e.g. when the robot has not been shown explicitly the occurence of a relevant object. The robot can thus make use of its a priori knowledge about typical occurences of objects and use this as a basis for autonomous planning.

Figure 3 shows an example of how partial information can be fused. If a `Room` instance (**r107**) exists that has an instance of a `CoffeeMachine` (**coffee_machine1**), the reasoner can infer that **r107** satisfies the necessary and sufficient conditions for being an instance of the (more specific) class `Kitchen`. (Note that Figure 3 only shows a small portion of our common-sense ontology.)

E. Interactive people following

For following the tutor we use a laser range based people tracking software [21] that uses a Bayesian filtering algorithm. The people tracker derives robust tracking information of dynamic objects within the robot's perceptual range. Given the tracking data, the people following module calculates appropriate motion commands that are sent to the robot control system to follow the tutor's trajectory, while preserving a socially appropriate distance to the tutor when standing still. The system is *interactive* in that it actively gives the tutor feedback about its state. A pan-tilt-unit with a stereo vision device is moved to always point to the tutor, thus giving a gaze-feedback such that he or she is aware that the robot is actually following the tutor and nobody else. Also, should the people tracker lose track of the tutor, we provide simple verbal grounding feedback (i.e. "Oops!") to quickly inform the tutor. This gives the tutor the possibility to immediately react and wait for the robot to recover. Once the person is found again, which typically takes about a second, another grounding feedback (i.e. "Ah!") is given to the tutor who can then proceed. The visual grounding feedback provided by the gaze helps detecting false recovery attempts quickly.

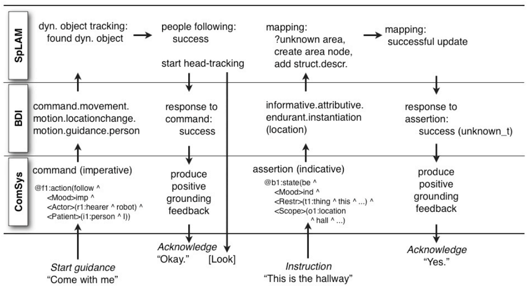
SpLAM
BDI
ComSys
dyn. object tracking: found dyn. object
people following: success
start head-tracking
mapping: ?unknown area, create area node, add struct.descr.
mapping: successful update
command.movement. motion.locationchange. motion.guidance.person
response to command: success
informative.attributive. endurant.instantiation (location)
response to assertion: success (unknown_t)
command (imperative)
@f1:action(follow ^ <Mood>imp ^ <Actor>(r1:hearer ^ robot) ^ <Patient>(i1:person ^ I))
produce positive grounding feedback
assertion (indicative)
@b1:state(be ^ <Mood>ind ^ <Restr>(t1:thing ^ this ^ ...) ^ <Scope>(o1:location ^ hall ^ ...)
produce positive grounding feedback
Start guidance
"Come with me"
Acknowledge
"Okay."
[Look]
Instruction
"This is the hallway"
Acknowledge
"Yes."

Fig. 4. Sample interaction

V. Examples

Figure 4 gives a flow diagram for a typical interaction between the robot and a human: First, the human instructs the robot to follow, and then tells the robot more about an aspect of the environment. The communication subsystem analyses "Come with me" as an *imperative*, expressing a command, and sends the content representation for the utterance's semantics to the BDI subsystem. Here we establish the complex ontological type of the utterance, indicating a guidance command, and decide to mediate the content with the SpLAM subsystem. This mediation triggers a new process. Dynamic object tracking has found a dynamic object, which we can interpret as a person. Thereupon, the SpLAM subsystem replies that the command is succesfully being executed, which then results in the communication subsystem producing a positive feedback: "Okay."

The *indicative* "This is the hallway" is interpreted as an assertion attributing a type description to the current location. The BDI subsystem therefore mediates this content (again) to the SpLAM subsystem. There, we create a new area node in the topological graph, and create a structural description for **hallway** for that node. One outcome of this process is that we can -through mediation- inform the communication subsystem of the successful augmentation of the map, upon which we again generate positive grounding feedback. Another outcome is that the identifier of the structural description for **hallway** is co-indexed with the identifier for the discourse referent for **hallway**, to form a belief at the mediation level.

VI. Experience

To round off, we briefly describe our experience with the implemented system. There are a couple of principal behaviors we need for HAM. If we want a human to guide a robot around an environment, then the robot must be able to (a) follow the human, (b) use information it gets from the human to augment its map, (c) take the initiative to ask the human for clarification; and (d) we need to be able to verify, and correct, what the robot has (not) understood. Where is the system successfull, and where is it not?[6]

[5] `http://www.racer-systems.com`

[6] Videos are available at `http://www.dfki.de/cosy/www/media`.

a) Although people tracking/following works fairly smoothly, the robot tends to loose track when the human e.g. passes around a corner. We are now studying how to predict the *path* where a tracked human is going, to overcome this problem and to reduce misclassifications of static objects as dynamic (due to laser data noise). We have also found that having a notion of what human behavior to expect is important: when a human moves to open a door, the robot should not follow the human behind the door, but go through it. The robot needs to reason over functionality of regions/objects in the environment to raise such expectations. We are now studying how we can use ontology reasoning to project functionality into the environment, and combine path prediction with functionality-related action recognition.

b) A question here is not just whether the robot can use information from the human - there is also the issue of how easy or difficult it is for the human to convey that information to the robot in the first place. In our grammar, we have *lexical families* that specify different types of syntactic structures and the meaning they convey, and *lexical entries* specifying how words belong to specific lexical families. This way we can specify many ways in which one can convey the same information (*synonymy*). Dialogue can thus be more flexible, as there is less need for the human to know and give the precise formulation (controlled language).

c) Clarification often concerns aspects of the environment which need to be explicitly referred to, e.g. "Is there a door *here*?". The difficulty lies in generating deictic references with a robot with a limited morphology. Although we can generate spatial referring expressions, non-verbal means would be preferable. However, body- and head-pose are not be distinctive enough. We may thus have to drive to a place (the "HERE") to make the deictic reference explicit, while avoiding disturbing the interaction.

d) Because we have reliable speech recognition (recognition rate is >90%), misunderstanding is primarily a semantic issue. This raises two main questions. First, how does the human understand that the robot understood what was said, without asking the robot? Various systems have the robot repeat what it has just heard. We have not done this; the robot only indicates whether it has understood ("yes"/"okay"/"no"). We have not experienced problems with this, but we are investigating now more explicit non-verbal cues for grounding feedback (e.g. gaze). Second, we need to study what types of misunderstanding may occur in HRI for HAM, and to what extent they may have a *relevant* effect on the robot's behavior. This is an issue we now investigate.

VII. Conclusions

We presented an HRI architecture for human-augmented mapping. We discussed the multi-level representations we build of the environment, including spatial organization and functional aspects (based on functions of objects present in areas). The system uses autonomous mapping, visual processing, human-robot interaction, and ontological reasoning to construct structural descriptions with which the multi-level representations are annotated. The approach has been fully implemented, and helps bridging the gap between robot and human conceptions of space. We showed its functionality, inspired by independently performed Wizard-of-Oz studies, on several running examples. For future research we want to study more detailed spatial organizations of regions and objects within rooms, to create 3-dimensional representations.

References

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

[2] B. Kuipers, "The spatial semantic hierarchy," *Artificial Intelligence*, vol. 119, pp. 191–233, 2000.

[3] E. Topp and H. Christensen, "Tracking for following and passing persons," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005, pp. 70–76.

[4] E. A. Topp, H. Hüttenrauch, H. I. Christensen, and K. Severinson Eklundh, "Acquiring a shared environment representation," in *Proc. Human-Robot Interaction (HRI'06)*, Salt Lake City, UT, 2006, pp. 361–362.

[5] H. Shi and T. Tenbrink, "Telling rolland where to go: HRI dialogues on route navigation," in *Proc. Workshop on Spatial Language and Dialogue*, Delmenhorst, Germany, 2005.

[6] J. Baldridge and G. Kruijff, "Coupling CCG and hybrid logic dependency semantics," in *Proc. ACL 2002*, Philadelphia, PA, 2002, pp. 319–326.

[7] G.-J. M. Kruijff, J. Kelleher, G. Berginc, and A. Leonardis, "Structural descriptions in human-assisted robot visual learning," in *Proc. Human-Robot Interaction (HRI'06)*, Salt Lake City, UT, 2006.

[8] J. Bos, E. Klein, and T. Oka, "Meaningful conversation with a mobile robot," in *Proc. Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, 2003.

[9] D. Bruemmer and M. Walton, "Collaborative tools for mixed teams of humans and robots," in *Proc. of the Workshop on Multi-Robot Systems*, Washington, D.C., 2003.

[10] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, "Clarification dialogues in human-augmented mapping," in *Proc. Human-Robot Interaction (HRI'06)*, Salt Lake City, UT, 2006.

[11] S. Wilske and G.-J. M. Kruijff, "Service robots dealing with indirect speech acts," in *Proc. Intelligent Robots and Systems (IROS'06)*, Bejing, China, 2006.

[12] I. Gurevych, R. Porzel, E. Slinko, N. Pfleger, J. Alexandersson, and S. Merten, "Less is more: Using a single knowledge representation in dialogue systems," in *Proc. HLT-NAACL WS on Text Meaning*, Edmonton, Canada, 2003.

[13] A. Cheyer and D. Martin, "The open agent architecture," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1, pp. 143–148, March 2001.

[14] J. Baldridge and G.-J. M. Kruijff, "Multi-modal combinatory categorial grammmar," in *Proc. EACL'03*, Budapest, Hungary, 2003.

[15] N. Asher and A. Lascarides, *Logics of Conversation.* Cambridge University Press, 2003.

[16] 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.

[17] 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 International Conference on Robotics and Automation (ICRA'02)*, Washington D.C., USA, 2002, pp. 1802–1809.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *Int. Jnl. Computer Vision*, 2004, pp. 91–110.

[19] M. Fritz, B. Leibe, B. Caputo, and B. Schiele, "Integrating representative and discriminant models for object category detection," in *Proc. International Conference on Computer Vision (ICCV05)*, Beijing, China, 2005.

[20] J. D. Kelleher and G.-J. M. Kruijff, "A context-dependent model of proximity in physically situated environments," in *Proc. Second ACL-SIGSEM workshop on the Linguistic Dimensions of Prepositions*, Colchester, Essex, 2005.

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