EU FP7 CogX
ICT-215181
May 1 2008 (52months)

DR 3.3:
Spatial entities for HRI and functional understanding of space

M. Zillich[1], K. Zhou[1], M. Vincze[1], N. Hawes[4], G. Horn[4], K. Sjöö[2], A. Aydemir[2], P. Jensfelt[2], H. Zender[3], G.-J. Kruijff[3]

[1]*TUW, Vienna* [2]*KTH, Stockholm* [3]*DFKI GmbH, Saarbrücken*
[4]*BHAM, Birmingham*

⟨zillich@acin.tuwien.ac.at⟩

Due date of deliverable:	Month 48
Actual submission date:	May 28, 2012
Lead partner:	TUW
Revision:	final
Dissemination level:	PU

WP3 deals with qualitative spatial cognition, i.e. the acquisition of spatial (room level) knowledge and reasoning within that knowledge to support efficient and robust task execution in an environment that presents incomplete and uncertain information, as well as to support human robot interaction (HRI) for communicating these tasks. Over the 4 years of CogX project we developed increasingly powerful enabling technologies to support the kind of reasoning required for a cognitive system that reflects on its knowledge and identifies gaps and accordingly opportunities for exploration. We furthermore integrated these enabling technologies into a framework for multi-layered conceptual spatial mapping which forms part of the CAST framework instantiation in the Dora demonstrator.

The present report deals with two bodies of work. First, an integrated model for representing spatial knowledge for situated action and human-robot interaction, and second a set of methods for functional understanding of space. These latter include segmentation and labelling of a geometric map of the environment, where the segmentation is based on functional definitions of the different room concepts, as well as identifying functional spatial regions within a room from spatial relations of objects in the room. Furthermore two methods for augmenting object search with higher level information, using either web searches to extract Common Sense about Object Locality (CSOL) or 3D context learned from a large set of labelled 3D training images, such as collected in the newly established project Kinect@Home.

1 Tasks, objectives, results 1
1.1 Planned work 1
1.2 Actual work performed 2
1.2.1 Task 3.4: Establishing reference to spatial entities for human-robot interaction 2
1.2.2 Task 3.5: Functional understanding of space 2
1.3 Relation to state-of-the-art 4

2 Annexes 7
2.1 H. Zender, "Multi-Layered Conceptual Spatial Mapping – Representing Spatial Knowledge for Situated Action and Human-Robot Interaction . . . 7
2.2 K. Sjöö, "Semantic map segmentation using function-based energy maximization" 8
2.3 N. Hawes et al., "Towards a Cognitive System That Can Recognize Spatial Regions Based on Context" 9
2.4 K. Zhou et. al, "Web Mining Driven Semantic Scene Understanding and Object Localization" 10
2.5 K. Zhou et. al, "Web Mining Driven Object Locality Knowledge Acquisition for Efficient Robot Behavior" 11
2.6 A. Aydemir and P. Jensfelt, "Exploiting and modeling local 3D structure for predicting object locations" 12
2.7 A. Aydemir et. al, "Kinect@Home: Crowdsourcing a Large 3D Dataset of Real Environments" 13

References 14

Executive Summary

Over the 4 years of the CogX project we developed a large body of work related to spatial cognition. Work was driven by the need of cognitive systems to deal with uncertain and incomplete information and reason with that knowledge to support efficient and robust task execution as well as communicating these tasks to the robot. Accordingly we developed various probabilistic methods (e.g. for room categorisation, for planning over uncertain information in large domains) and integrated these into a comprehensive framework as demonstrated in the Dora scenario.

This report deals with two aspects within this larger context of spatial cognition. First, a model for representing spatial knowledge for situated action and human-robot interaction, addressing *Task 3.4 Establishing reference to spatial entities for human-robot interaction.* The problems here are that the robot is faced with changing and incomplete spatial information about the environment, and needs to communicate the semantics of this spatial information at different levels of abstraction in a natural way, to support situated human-robot interaction. We developed the enabling techniques, such as room categorisation and reasoning about typical objects present in a room, and integrated these into a comprehensive probabilistic framework, enabling planning and task execution with uncertain and incomplete information.

Secondly, we present work related to *Task 3.5: Functional understanding of space.* Here we present a method that uses learned spatial relations between objects in the room together with analogy to define functional regions such as “the front of the room”. A complementary method uses information provided by the web rather than learning by the system for segmenting and labelling a geometric map of the environment, where the segmentation is based on functional definitions of the different room concepts, based on the definition in the Oxford online dictionary, defining e.g. a kitchen as a room where food is cooked. We furthermore use knowledge from the web to extract Common Sense about Object Locality (CSOL). For this we calculate the likelihood of finding objects at certain locations from search query results such as “the cup was on the table” or “the mug was on the shelf”, and use these locations to direct search for these objects. A complementary approach is independent of room category and uses surrounding 3D structure (termed 3D context) to direct search for a given object, avoiding the need to explicitly detect supporting surfaces such as shelves. This 3D context is learned from labelled 3D training data. To collect a wide variety of different typical indoor scenes, we initiated the Kinect@Home project (`http://www.kinectathome.com`), where users can upload 3D image sequences, where special care had to be taken to handle the enormous amount of point cloud data using special compression techniques.

Role of spatial cognition in CogX

Spatial cognition here serves two roles: First as the process of abstracting raw metric spatial information into semantically meaningful information to support task planning and execution with uncertain information situated and to support human robot interaction. Secondly, as top down context information for object search, e.g. for a cup on a kitchen counter.

Contribution to the CogX scenarios and prototypes

The work presented here is mainly used in the Dora scenario, where the robot recognises different room types (based on functionality) and uses these to communicate with the user. Also object search at room level, e.g. for fetch and carry tasks, is most associated with the Dora scenario.

1 Tasks, objectives, results

1.1 Planned work

Task 3.4: Establishing reference to spatial entities for human-robot interaction. *The goal is to investigate, in the context of human-robot interaction, how the robot can refer to objects based on their spatial relations and how to learn this.*

Task 3.5: Functional understanding of space. *The goal is to investigate how to gain knowledge about the function of space by analyzing spatial models over time.*

Task 3.4 originally had a focus on learning spatial relations between objects in a scene and using these for human robot interaction (HRI). The actual work performed in this task then concentrated more on the room level, building a hierarchy of spatial concepts for HRI, which turned out to be more relevant to work in the scenarios. Task 3.5 aimed at learning from analysis over time. Instead we chose to learn from large corpora on the web, which is a promising route of research especially when requiring large amounts of training data.

The work presented in this deliverable contributed to the following of the CogX objectives:

- 2. Specific representations of beliefs about beliefs for the specific cases of dialogue, manipulation, maps, mobility and some types of vision. [WPs 2,3,6]
- 3. Representations of how actions will alter the belief state of the cognitive system, and those of other agents, as represented in the first two objectives, i.e. models of the effects of actions on beliefs about space, categorical knowledge, action effects, dialogue moves etc. [WPs 1,2,3,4,5,6]
- 7. Methods for perception and spatial modelling that enable a robot to identify gaps in its spatial models (e.g. maps) and to extend them so as to support natural communication with humans. [WP 3]
- 11. A robotic implementation of our theory able to complete a task involving mobility, interaction and manipulation, in the face of novelty, uncertainty, partial task specification, and incomplete knowledge. [WPs 2,3,6,7]

We address objectives 2, 3 and 7 by providing a multi-layered conceptual spatial mapping framework that on top of metric and topological maps represents probabilistic knowledge about room categories and relations between rooms and objects found in them. We also provide the planning techniques required to deal with this kind of uncertain information in large planning domains. Objective 11 is addressed by demonstrating the validity of our approaches in numerous experiments in the Dora scenario.

1.2 Actual work performed

1.2.1 Task 3.4: Establishing reference to spatial entities for human-robot interaction

Intelligent autonomous robots that efficiently collaborate with humans in everyday tasks must have the capabilities to engage in *situated human-robot interaction*. This implies that they must be able to understand their spatial environment and its semantics in a way that is compatible to the way their human users do. If they are furthermore expected to conduct *situated spoken dialogues*, their spatial conceptualization must be expressible in natural language. On the other hand, however, intelligent mobile robots must be endowed with navigation capabilities that take into account the specific sensors and actuators the robot is equipped with.

The kinds of autonomous mobile robots that we consider in CogX ultimately operate in dynamic, large-scale environments. These environments are subject to change and cannot be apprehended as a perceptual whole. At the same time, the robots have the possibility to alter the world around them, and to perform actions that allow them to extend their own knowledge. For this to be successful, their knowledge representation must be able to deal with *changing* and *incomplete information*.

In [44] (Annex 2.1) we present a consolidated and integrated approach to *multi-layered conceptual spatial mapping* that addresses the aforementioned challenges. In this approach, spatial knowledge is represented at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations. We also discuss reasoning methods that can be performed using such spatial conceptual knowledge in order to overcome the problem of *partial information at the sensory-symbol interface*, as well as the bootstrapping of ontological knowledge from available linguistic and commonsense databases, and how such knowledge can be quantified in order to support probabilistic action planning for more efficient robot behaviour in human-oriented environments.

The work presented here summarises the underlying representations for reference resolution in spatial contexts reported previously in DR.6.4, Annex 2.1.

1.2.2 Task 3.5: Functional understanding of space

When interacting with people, human level concepts such as room labels are very important. In [33] (Annex 2.2) we present a method for simultaneously segmenting and labeling a geometric map of the environment. The segmentation is based on commonsense definitions from the Oxford online dictionary – for example, a kitchen is defined as “That room or part of a house in which food is cooked; a place fitted with the apparatus for cooking.” We note that the definitions are crucially bound up with aspects of function – e.g., what ultimately makes something a kitchen is that food can be cooked there – and consequently we posit concrete numerical interpretations of these functional apects. Combining these values into an energy function which is then maximized, we produce a function-sensitive segmentation of space. It is also shown how the segmentation can adjust to accommodate referring expressions. For example, if the human were to mention the “kitchen next to the corridor” when speaking to the system it would be able to use this as an indication that the segmentation needs to produce at least one kitchen and at least one corridor, next to each other.

In the work discussed in [21] (Annex 2.3) we define spatial regions (such as the front of a room) by functional use, but this time derived from spatial relations of objects in the room (such as chairs all pointing in a certain direction). We present a cognitive system able to learn context-dependant spatial regions by combining qualitative spatial representations, semantic labels, and analogy and evaluate it against human annotations of real world scenes.

In the work on object search previously reported in CogX we used the assumption that objects are often to be found on tables or other supporting surfaces. This assumption was taken for granted and hard-coded into parts of the system. Starting with our work in [19], and also DR.6.4, Annex 2.2, [44] (DR.3.3, Annex 2.1) and [1] (DR.3.4, Annex 2.1), we showed how this common sense knowledge can be extracted from web queries in a probabilistic fashion, which significantly improves the performance of visual search. There we employed knowledge like “cups are likely to be located in kitchens” in a visual search task using a planner switching between continual symbolic planning and decision theoretic planning, which was capable of dealing with the uncertain information (cups are not always in kitchens after all) as well as the large planning domain. In the work presented here in [47, 48] (Annexes 2.4 and 2.5) we expanded on the way in which those queries are formed. Additionally to the image search engine employed in our previous work we also employed a web text mining technique using sequential pattern retrieval to extract Common Sense about Object Locality (CSOL) for linking the search of objects with their potential localities. We calculate the object location belief $OLB(O, L)$ of finding object O at location L by searching for patterns like 'object' + '*be*' + 'on' + … + 'location', such as “the cup was on the table”. We use specific databases like the Open Mind Indoor Common Sense database (OMICS)[1] or generic web searches on google, yahoo or bing. The result is a probability distribution over locations

[1] `openmind.hri-us.com`, Honda Research Institute USA

an object is most likely to be found. These locations then map to constraints for the visual search task. Experiments using an indoor mobile robot for an Active Visual Search (AVS) task (e.g. for a cup or can) demonstrate the benefits in terms of reduced search time.

The above approach exploits spatial relations between objects (supporting surfaces and objects on them in that case) to perform the search more efficiently. One of the bottlenecks with this is that we rely heavily on the perception system to categorize objects. Unless finding the larger supporting object is easy it might not help enough in finding the small objects on it. One strand of work therefore investigated ways to build models for calculating the likelihood of finding objects not based on the detection of other objects but by surrounding 3D structure (we call this the 3D context) which gives strong cues as to what objects could be found there. So, instead of learning that cups are on tables, we learn that the local surrounding of a cup is typically planar and horizontal. This results in a more flexible model presented in [3] (Annex 2.6).

When working on the 3D context we initially gathered a dataset from the different sites within CogX (reported on last year in DR.3.2). We soon realized that if we are serious about understanding real-world spaces we need to have data from such environments and data from robot labs gathered by roboticists across Europe might not be all that representative. We have therefore started an effort (`http://www.kinectathome.com`) to gather a large dataset of data from Microsoft's new sensor, the Kinect. We are working on the final details for the launch of this and plan to announce it widely at the end of the summer. The idea behind this effort was presented in [2] (Annex 2.7).

1.3 Relation to state-of-the-art

The work reported in Annex 2.1 builds upon and extends the author's previous research on *multi-layered conceptual spatial mapping* [45, 46] in the tradition of approaches like the *(Hybrid) Spatial Semantic Hierarchy* by Kuipers *et al.* [24, 25, 5], the *Route Graph* model by Krieg-Brückner *et al.* [43, 23], Buschka and Saffiotti's *hybrid maps* [8], as well as *multi-hierarchical semantic maps* for mobile robots by Galindo *et al.* [18, 17].

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.* [39] suggest a hierarchical probabilistic representation of space based on objects. The work by Galindo *et al.* [18, 17] presents an approach containing two parallel hierarchies, spatial and conceptual, connected through anchoring. Inference about places is based on objects found in them. This approach is based on the Multi-AH-graph model by Fernandez and Gonzalez [14]. The work by Diosi *et al.* [11] creates a metric map through a guided tour. The map is then segmented into discrete rooms according to the labels given by the instructor. Furthermore, the *Hybrid Spatial Semantic Hierarchy* (HSSH), introduced by Beeson *et al.* [5], allows a mobile robot to describe the world using different representations, each with its own ontology.

More recently, Pronobis *et al.* [32] have presented a refined approach to multi-layered mapping, in which, inter alia, the representations of the lower map layers were re-defined, and a probabilistic inference engine is used for reasoning with the discrete symbols in the conceptual map layer.

Lemaignan *et al.* [26] present a similar approach to endowing robots with spatial representations that allow them to act in and talk about their environment. Their framework has the advantage of providing a kind of *theory of mind* that allows the robot to reason about the perspective of its interlocutor in order to disambiguate and ground natural-language instructions. While our approach addresses the specific challenges involved when engaging in dialogues about spatial environments that are larger than what can be perceived at once, their approach focusses on adequate reasoning techniques for shared visual scenes, like, e.g. tabletop scenarios.

With the availability of affordable 3D sensors and appropriate techniques for using them for robotic mapping purposes, a number of approaches for building layered representations of 3D space have been proposed recently. The KnowRob-Map framework [36] combines low-level metric costmaps, maps of 3D point clouds, and ontological knowledge bases into a semantic environment model of places, object locations, and afforded actions. Pangercic *et al.* [30] use natural-language task instructions from the WWW to construct a Description Logics-based knowledge base for tabletop scenarios. Tenorth *et al.* [35] present a framework that allows mobile service robots to use multiple web-based knowledge sources (including OMICS, WordNet and an internet image search engine) in order to perform everyday manipulation tasks. While these approaches are especially useful for (mobile) manipulation in human-oriented environment (e.g., kitchens [6]), our approach has a stronger focus on human-robot interaction and situated human-robot dialogues.

Viswanathan *et al.* [40, 41] propose another approach that makes use of existing commonsense knowledge resources. They use the LabelMe dataset to train an automated place classifier that relies on the presence of detected objects to infer which other objects are likely to occur nearby and which kind of place (e.g., kitchen or office) is seen in the scene.

Given a discretization of space, for example in the form of a Voronoi diagram, Diosi et al.[10] and Milford et al. [29] let a user impose labels for different locations. In [28] metric features are used to classify regions, while [38] utilize spatial relations between objects. Friedman et al. [15] use a graph-based approach in which place classification is based on potentials defined on nodes in a graph. The model is more local, and learned as opposed to specified by functional criteria as in our work. In the work by Friedman et al. the world is segmented into either belonging to the class of corridor or room, but no distinction is made between different rooms or corridors. In our work in Annex 2.2 we identify the individual areas as well as label them.

Knowledge acquisition from the web or sharing databases have been adopted to supply a large corpus of training data [13] for visual recognition, to build 3D models for robot manipulation [22], improve visual object recognition [27], to complete qualia structures describing an object [9], to guide robot planning for specific tasks such as table setting for a meal [31], and even more ambitiously to fill knowledge gaps when an indoor robot is executing sophisticated tasks [42]. [19] showed how web queries revealing probabilistic knowledge about the most likely room locations of various objects significantly improves search for a given object in a robotic system able to plan with uncertain knowledge. In the work presented in Annexes 2.4 and 2.5 we expand on the way in which those web queries are performed and incorporate queries from image as well as text databases.

The work closest to our work on using the 3D shape context (Annex 2.6) to predict object locations is probably [37] where low-level features are extracted from the whole image for context driven attention and object detection. We make use of the 3D information and propose a conceptually simple method to capture and exploit this information.

Work presented in Annex 2.3 created representations of spatial regions that may be referenced by humans in task descriptions, e.g. the instruction for the robot to "go to the *front of the classroom*". These regions are defined using Qualitative Spatial Relations based on the objects present in a room and their configuration. Whilst mobile robots exist which can determine the type of a room from the objects found in it [20, 16], these works only concern themselves with the types of whole rooms, and cannot represent subregions within them. This is also true for those robotic systems which use some elements of QSR [4]. The need for an autonomous system to ground references to human-generated descriptions of space has been recognised in domains where a robot must be instructed to perform a particular task, however existing systems are restricted to purely geometrically-defined regions [34, 12, 7], rather than the qualitatively-defined, functional regions in our work.

2 Annexes

2.1 H. Zender, “Multi-Layered Conceptual Spatial Mapping – Representing Spatial Knowledge for Situated Action and Human-Robot Interaction

Bibliography H. Zender. “Multi-Layered Conceptual Spatial Mapping – Representing Spatial Knowledge for Situated Action and Human-Robot Interaction.” in In Y. Amirat, A. Chibani, and G. P. Zarri, editors, *Bridges Between the Methodological and Practical Work of the Robotics and Cognitive Systems Communities – From Sensors to Concepts*, Intelligent Systems Reference Library. Springer Verlag, Berlin/Heidelberg, Germany, 2012 (to appear).

Abstract In this book chapter, we present the principle of multi-layered conceptual spatial mapping. In multi-layered conceptual spatial mapping, spatial knowledge is represented at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations. It addresses the diverse needs involved in representing spatial knowledge for situated action and human-robot interaction. We give an overview of relevant topics in human cognition that need to be taken into account when designing robotic systems that are supposed to act for and among humans. We then describe different existing individual mapping techniques that can be integrated into a multi-layered conceptual spatial map, with a special emphasis on ontological reasoning techniques that can be employed at the highest level of abstraction in order to link the internal robotic spatial representations to human-compatible concepts and symbols.

Relation to WP Abstracting from raw metric sensor data to a spatial representation that is meaningful in a situated human robot dialogue (Task 3.4) is a crucial capability for any cognitive robot, as demonstrated in the Dora scenario,

2.2 K. Sjöö, "Semantic map segmentation using function-based energy maximization"

Bibliography K. Sjöö, "Semantic map segmentation using function-based energy maximization", In Proc. of the International Conference on Robotics and Automation (ICRA), 2012

Abstract This work describes the automatic segmentation of 2-dimensional indoor maps into semantic units along lines of spatial function, such as connectivity or objects used for certain tasks. Using a conceptually simple and readily extensible energy maximization framework, segmentations similar to what a human might produce are demonstrated on several real-world datasets. In addition, it is shown how the system can perform reference resolution by adding corresponding potentials to the energy function, yielding a segmentation that responds to the context of the spatial reference.

Relation to WP The work presented in this paper details one possibility to abstract from metric floor plans into functionally relevant spatial regions (Task 3.5), thus feeding into the multi-layered conceptual spatial map described in the work in Annex 2.1.

2.3 N. Hawes et al., "Towards a Cognitive System That Can Recognize Spatial Regions Based on Context"

Bibliography N. Hawes, M Klenk, K. Lockwood, G.S. Horn and John D. Kelleher, "Towards a Cognitive System That Can Recognize Spatial Regions Based on Context", Proceedings of the 26th National Conference on Artificial Intelligence (AAAI), 2012

Abstract In order to collaborate with people in the real world, cognitive systems must be able to represent and reason about spatial regions in human environments. Consider the command "go to the front of the classroom". The spatial region mentioned (the front of the classroom) is not perceivable using geometry alone. Instead it is defined by its functional use, implied by nearby objects and their configuration. In this paper, we define such areas as context-dependent spatial regions and present a cognitive system able to learn them by combining qualitative spatial representations, semantic labels, and analogy. The system is capable of generating a collection of qualitative spatial representations describing the configuration of the entities it perceives in the world. It can then be taught context-dependent spatial regions using anchor points defined on these representations. From this we then demonstrate how an existing computational model of analogy can be used to detect context-dependent spatial regions in previously unseen rooms. To evaluate this process we compare detected regions to annotations made on maps of real rooms by human volunteers.

Relation to WP This paper presents a new approach to representing regions of space whose presence and shape are dependent on spatial context, i.e. the objects present in a scene and their configuration. Regions of this nature are of particular relevance to this WP because they represent an approach to building functional models of space (Task 3.5) without explicitly representing human activity, and they are a type if regions that humans may make reference to when talking to a robot (Task 3.4).

2.4 K. Zhou et. al, “Web Mining Driven Semantic Scene Understanding and Object Localization”

Bibliography K. Zhou, K. M. Varadarajan, M. Zillich, M. Vincze, “Web Mining Driven Semantic Scene Understanding and Object Localization”, IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 2824-2829, 2011

Abstract Knowledge acquisition from the Internet for robotic applications has received widespread attention recently. It has turned out to be an important supplementary or even a complete replacement to conventional robotic perception. In this paper, we investigate state-of-the-art online knowledge acquisition systems for robotic vision applications and present a framework for further fusion and tighter integration. Boot-strapped by an interconnected process wherein modules for object detection and supporting structure detection co-operate to extract cross-correlated information, a web text mining technique using sequential pattern retrieval is introduced for linking the search of objects with their potential localities. Experiments using an indoor mobile robot for an Active Visual Search (AVS) task demonstrate the benefits of our coherent framework for visual representation and knowledge acquisition from the Internet.

Relation to WP One of the reasons for the importance of knowing about the semantics of space is that it allows to formulate expectations of what to find there, where the semantics of a space is related to the function it provides (Task 3.5). In the above work we use information from the web to identify typical object locations.

2.5 K. Zhou et. al, “Web Mining Driven Object Locality Knowledge Acquisition for Efficient Robot Behavior”

Bibliography K. Zhou, M. Zillich, M. Vincze, “Web Mining Driven Object Locality Knowledge Acquisition for Efficient Robot Behavior”, submitted to the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012

Abstract As an important information resource, visual perception has been widely employed for various indoor mobile robots. The common-sense knowledge about object locality (CSOL), e.g. a cup is usually located on the table top rather than on the floor and vice versa for a trash bin, is a very helpful context information for a robotic visual search task. In this paper, we propose an online knowledge acquisition mechanism for discovering CSOL, thereby facilitating a more efficient and robust robotic visual search. The proposed mechanism is able to create conceptual knowledge with the information acquired from the largest and the most diverse medium – the Internet. Experiments using an indoor mobile robot demonstrate the efficiency of our approach as well as reliability of goal-directed robot behaviour.

Relation to WP One of the reasons for the importance of knowing about the semantics of space is that it allows to formulate expectations of what to find there, where the semantics of a space is related to the function it provides (Task 3.5). In the above work we use information from the web to identify typical object locations.

2.6 A. Aydemir and P. Jensfelt, “Exploiting and modeling local 3D structure for predicting object locations”

Bibliography A. Aydemir and P. Jensfelt, “Exploiting and modeling local 3D structure for predicting object locations”, submitted to the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012

Abstract In this paper, we argue that there is a strong correlation between local 3D structure and object placement in everyday scenes. We call this the 3D context of the object. In previous work, this is typically hand-coded and limited to flat horizontal surfaces. In contrast, we propose to use a more general model for 3D context and learn the relationship between 3D context and different object classes. This way, we can capture more complex 3D contexts without implementing specialized routines. We present extensive experiments with both qualitative and quantitative evaluations of our method for different object classes. We show that our method can be used in conjunction with an object detection algorithm to reduce the rate of false positives. Our results support that the 3D structure surrounding objects in everyday scenes is a strong indicator of their placement and that it can give significant improvements in the performance of, for example, an object detection system. For evaluation, we have collected a large dataset of Microsoft Kinect frames from five different locations, which we also make publicly available.

Relation to WP Similar to Annex 2.4 this work deals with object search, where in this case the local 3D context around an object encodes local functional understanding (Task 3.5), e.g. a door handle being attached to the vertical door blade next to the door frame.

2.7 A. Aydemir et. al, “Kinect@Home: Crowdsourcing a Large 3D Dataset of Real Environments”

Bibliography A. Aydemir, D. Henell, P. Jensfelt and R. Shilkrot, “Kinect@Home: Crowdsourcing a Large 3D Dataset of Real Environments”, AAAI Spring Symposium 2012: Wisdom of the Crowd

Abstract We present Kinect@Home, aimed at collecting a vast RGB-D dataset from real everyday living spaces. This dataset is planned to be the largest real world image collection of everyday environments to date, making use of the availability of a widely adopted robotics sensor which is also in the homes of millions of users, the Microsoft Kinect camera.

Relation to WP The crowd-sourcing project presented in this work provides (amongst others) the training data for the learning mechanism in Annex 2.6.

References

[1] Alper Aydemir, Moritz Göbelbecker, Andrzej Pronobis, Kristoffer Sjöö, and Patric Jensfelt. Plan-based object search and exploration using semantic spatial knowledge in the real world. In *Proc. of the European Conference on Mobile Robotics (ECMR'11)*, Örebro, Sweden, September 2011.

[2] Alper Aydemir, Daniel Henell, Patric Jensfelt, and Roy Shilkrot. Kinect@home: Crowdsourcing a large 3d dataset of real environments. In *AAAI Spring Symposium 2012: Wisdom of the Crowd*, 2012.

[3] Alper Aydemir and Patric Jensfelt. Exploiting and modeling local 3d structure for predicting object locations. In *submitted to Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'12)*, 2012.

[4] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011.

[5] Patrick Beeson, Matt MacMahon, Joseph Modayil, Aniket Murarka, Benjamin Kuipers, and Brian Stankiewicz. Integrating multiple representations of spatial knowledge for mapping, navigation, and communication. In *Interaction Challenges for Intelligent Assistants*, Papers from the AAAI Spring Symposium, Stanford, CA, USA, 2007. AAAI.

[6] Nico Blodow, Cosmin Goron, Zoltan-Csaba Marton, Dejan Pangercic, Thomas Rühr, Moritz Tenorth, and Michael Beetz. Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4263–4270, San Francisco, CA, USA, September 2011.

[7] M. Brenner, N. Hawes, J. Kelleher, and J. Wyatt. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 2072–2077, Hyderabad, India, 2007.

[8] Pär Buschka and Alessandro Saffiotti. Some notes on the use of hybrid maps for mobile robots. In *Proceedings of the 8th International Conference on Intelligent Autonomous Systems (IAS)*, Amsterdam, The Netherlands, March 2004.

[9] Philipp Cimiano and Johanna Wenderoth. Automatically learning qualia structures from the web. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, DeepLA '05, pages 28–37, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[10] A. Diosi, G. Taylor, and L. Kleeman. Interactive slam using laser and advanced sonar. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 1103–1108. IEEE, 2005.

[11] Albert Diosi, Geoffrey Taylor, and Lindsay Kleeman. Interactive SLAM using laser and advanced sonar. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, Barcelona, Spain, April 2005.

[12] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA'09)*, Kobe, Japan, May 2009.

[13] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*, volume 2, pages 1816–1823, October 2005.

[14] Juan-Antonio Fernández and Javier González. *Multi-Hierarchical Representation of Large-Scale Space – Applications to Mobile Robots*, volume 24 of *International Series on Microprocessor-Based and Intelligent Systems Engineering.* Kluwer Academic Publishers, Dordrecht / Boston / London, 2001.

[15] S. Friedman, H. Pasula, and D. Fox. Voronoi random fields: Extracting the topological structure of indoor environments via place labeling. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 35, 2007.

[16] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. A. Fernandez-Madrigal, and J. Gonzalez. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, pages 2278 – 2283, August 2005.

[17] Cipriano Galindo, Juan-Antonio Fernández-Madrigal, and Javier González. *Multiple Abstraction Hierarchies for Mobile Robot Opera-*

tion in Large Environments, volume 68 of *Studies in Computational Intelligence*. Springer Verlag, Berlin/Heidelberg, Germany, 2007.

[18] Cipriano Galindo, Alessandro Saffiotti, Silvia Coradeschi, Pär Buschka, Juan-Antonio Fernández-Madrigal, and Javier González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-05)*, pages 3492–3497, Edmonton, Canada, August 2005.

[19] M. Hanheide, C. Gretton, R. Dearden, N. Hawes, J. Wyatt, A. Pronobis, A. Aydemir, M. Goebelbecke, and H. Zender. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 2011.

[20] Marc Hanheide, Nick Hawes, Jeremy Wyatt, Moritz Göbelbecker, Michael Brenner, Kristoffer Sjöö, Alper Aydemir, Patric Jensfelt, Hendrik Zender, and Geert-Jan M. Kruijff. A framework for goal generation and management. In *Proceedings of the AAAI'10 Workshop on Goal-Directed Autonomy*, 2010.

[21] Nick Hawes, Matthew Klenk, Kate Lockwood, Graham S. Horn, and John D. Kelleher. Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the 26th National Conference on Artificial Intelligence (AAAI'12)*, 2012.

[22] Ulrich Klank, Muhammad Zeeshan Zia, and Michael Beetz. 3d model selection from an internet database for robotic vision. In *IEEE International Conference on Robotics and Automation*, pages 2406 –2411, May 2009.

[23] Bernd Krieg-Brückner, Udo Frese, Klaus Lüttich, Christian Mandel, Till Massokowski, and Robert J. Ross. Specification of an ontology for Route Graphs. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, and Interaction*, volume 3343 of *Lecture Notes in Artificial Intelligence*, pages 390–412. Springer Verlag, Heidelberg, Germany, 2005.

[24] Benjamin Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119:191–233, 2000.

[25] Benjamin Kuipers, Joseph Modayil, Patrick Beeson, Matt MacMahon, and Francesco Savelli. Local metrical and global topological maps in the Hybrid Spatial Semantic Hierarchy. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA 2004)*, New Orleans, LA, USA, April 2004.

[26] Séverin Lemaignan, Raquel Ros, E. Akin Sisbot, Rachid Alami, and Michael Beetz. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2):181–199, 2012.

[27] Marcin Marszalek and Cordelia Schmid. Semantic Hierarchies for Visual Object Recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[28] O. Martinez Mozos, R. Triebel, P. Jensfelt, A. Rottmann, and W. Burgard. Supervised semantic labeling of places using information extracted from sensor data. *Robotics and Autonomous Systems*, 55(5):391–402, 2007.

[29] M. Milford, R. Schulz, D. Prasser, G. Wyeth, and J. Wiles. Learning spatial concepts from ratslam representations. *Robotics and Autonomous Systems*, 55(5):403–410, 2007.

[30] Dejan Pangercic, Rok Tavcar, Moritz Tenorth, and Michael Beetz. Visual scene detection and interpretation using encyclopedic knowledge and formal description logic. In *Proceedings of the International Conference on Advanced Robotics (ICAR).*, Munich, Germany, June 2009.

[31] Dejan Pangercic, Rok Tavcar, Moritz Tenorth, and Michael Beetz. Visual scene detection and interpretation using encyclopedic knowledge and formal description logic. In *Proceedings of the International Conference on Advanced Robotics (ICAR).*, Munich, Germany, June 22 - 26 2009.

[32] Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Adrian N. Bishop, and Patric Jensfelt. Representing spatial knowledge in mobile cognitive systems. In *11th International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada, August 2010.

[33] Kristoffer Sjöö. Semantic map segmentation using function-based energy maximization. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA'12)*, May 2012.

[34] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI'11)*, 2011.

[35] Moritz Tenorth, Ulrich Klank, Dejan Pangercic, and Michael Beetz. Web-enabled Robots – Robots that Use the Web as an Information Resource. *Robotics & Automation Magazine*, 18(2):58–68, 2011.

[36] Moritz Tenorth, Lars Kunze, Dominik Jain, and Michael Beetz. KNOWROB-MAP – Knowledge-Linked Semantic Object Maps. In *Proceedings of the 10th IEEE-RAS International Conference on Humanoid Robots*, pages 430–435, Nashville, TN, USA, December 2010.

[37] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[38] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart. Cognitive maps for mobile robotsan object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, 2007.

[39] Shrihari Vasudevan, Stefan Gachter, Viet Nguyen, and Roland Siegwart. Cognitive maps for mobile robots – an object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, May 2007.

[40] Pooja Viswanathan, David Meger, Tristram Southey, James J. Little, and Alan K. Mackworth. Automated spatial-semantic modeling with applications to place labeling and informed search. In *CRV '09: Proceedings of the 2009 Canadian Conference on Computer and Robot Vision*, pages 284–291, Washington, DC, USA, 2009. IEEE Computer Society.

[41] Pooja Viswanathan, Tristram Southey, James J. Little, and Alan K. Mackworth. Automated place classification using object detection. In *Proceedings of the Seventh Canadian Conference on Computer and Robot Vision (CRV 2010)*, Ottawa, Canada, 2010.

[42] Markus Waibel, Michael Beetz, Raffaello D'Andrea, Rob Janssen, Moritz Tenorth, Javier Civera, Jos Elfring, Dorian Gálvez-López, Kai Häussermann, J.M.M. Montiel, Alexander Perzylo, Björn Schieŝle, Oliver Zweigle, and René van de Molengraft. RoboEarth - A World Wide Web for Robots. *Robotics & Automation Magazine*, 18(2), 2011.

[43] Steffen Werner, Bernd Krieg-Brückner, and Theo Herrmann. Modelling navigational knowledge by Route Graphs. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl F. Wender, editors, *Spatial Cognition II*, volume 1849 of *Lecture Notes in Artificial Intelligence*, pages 295–316. Springer Verlag, Heidelberg, Germany, 2000.

[44] Hendrik Zender. Multi-layered conceptual spatial mapping – representing spatial knowledge for situated action and human-robot interaction. In Yacine Amirat, Abdelghani Chibani, and Gian Piero Zarri, editors, *Bridges Between the Methodological and Practical Work of the Robotics and Cognitive Systems Communities – From Sensors to Concepts*, Intelligent Systems Reference Library. Springer Verlag, Berlin/Heidelberg, Germany, to appear 2012.

[45] Hendrik Zender and Geert-Jan M. Kruijff. Multi-layered conceptual spatial mapping for autonomous mobile robots. In Holger Schultheis, Thomas Barkowsky, Benjamin Kuipers, and Bernhard Hommel, editors, *Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems – Papers from the AAAI Spring Symposium*, Technical Report SS-07-01, pages 62–66, Menlo Park, CA, USA, March 2007. AAAI, AAAI Press.

[46] Hendrik Zender, Óscar Martínez Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June 2008.

[47] Kai Zhou, Karthik Mahesh Varadarajan, Michael Zillich, and Markus Vincze. Web mining driven semantic scene understanding and object localization. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Phuket, Thailand, Dec 2011.

[48] Kai Zhou, Michael Zillich, and Markus Vincze. Web mining driven object locality knowledge acquisition for efficient robot behavior. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (submitted)*, Vilamoura, Algarve, Portugal, Oct 2012.

Multi-Layered Conceptual Spatial Mapping

Representing Spatial Knowledge for Situated Action and Human-Robot Interaction

Hendrik Zender

Abstract In this chapter, we present the principle of multi-layered conceptual spatial mapping. In this approach, spatial knowledge is represented at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations. The approach addresses the diverse needs involved in representing spatial knowledge for situated action and human-robot interaction. In the beginning of this chapter, we give an overview of relevant topics in human cognition. We then describe existing robotic mapping techniques that can be integrated into a multi-layered conceptual spatial map, with a special emphasis on ontological reasoning techniques that can be employed at the highest level of abstraction to link the robot's spatial representations to human-compatible concepts and symbols. We conclude with a discussion of how ontological knowledge can be bootstrapped from available linguistic and commonsense databases, and how such knowledge can be quantified in order to support probabilistic action planning for more efficient robot behavior.

1 Introduction

If we want intelligent autonomous robots to efficiently collaborate with humans in everyday tasks, they must have the capabilities to to engage in *situated human-robot interaction*. This implies that they must be able to understand their spatial environment and its semantics in a way that is compatible with the way their human users do. If they are furthermore expected to conduct *situated spoken dialogues*, their spatial model must be expressible in natural language. The problem is complicated by the fact that intelligent mobile robots must be endowed with navigation capabilities that take into account the specific sensors and actuators the robot is equipped with. Such autonomous, interactive, mobile robotic systems must thus have access

H. Zender
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, e-mail: zender@dfki.de

to low-level spatial representations that are suitable for fine-grained control, while at the same time their representations must afford a human-compatible spatial understanding. The challenge is to establish such qualitative representations on the basis of low-level maps that are built from sensor input.

The kinds of autonomous robots that we consider in this work operate in dynamic, large-scale environments. These environments are subject to change and cannot be apprehended as a perceptual whole. At the same time, the robots have the possibility to alter the world around them, and to perform actions that allow them to extend their own knowledge. For this to be successful, their knowledge representation must be able to deal with *changing* and *incomplete information*.

In the following, we will discuss an approach to *multi-layered conceptual spatial mapping* that addresses the aforementioned challenges. In this approach, spatial knowledge is represented at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations. We will also discuss reasoning methods that can be performed using such spatial conceptual knowledge in order to overcome the problem of *partial information at the sensory-symbol interface*.

Two different instantiations of the multi-layered conceptual spatial mapping approach have been implemented in the two integrated robotic systems “Dora” [93, 36, 35, 34] and its predecessor the “CoSy Explorer” [48, 100, 37, 80] shown in Figure 1. The principles discussed here will be illustrated with examples from these integrated systems.

After an introduction to the challenges involved in representing spatial knowledge for situated action and human-robot interaction in Section 2, we discuss similar approaches and other related work in Section 3. Section 4 presents an overview of relevant topics in human cognition that need to be taken into account when designing robotic systems that are supposed to interact with humans. In Section 5 we describe an approach to multi-layered conceptual spatial mapping that makes use of different existing techniques for representing spatial knowledge on different levels of abstraction. Section 6 continues with a discussion of symbolic reasoning techniques that can be used at the highest level of abstraction, in order to link the internal robotic spatial representations to human-compatible concepts and symbols. In Section 7 we discuss how ontological knowledge can be bootstrapped and disambiguated using available resources for linguistic and commonsense knowledge (WordNet and Open Mind Indoor Common Sense Project, respectively). We furthermore describe how a search engine (Bing image search) can be leveraged to quantify commonsense facts in order to support probabilistic action planning for more efficient robot behavior. We conclude in Section 8 with a reflection of the presented approach.

2 Motivation

Both the robotics community and the cognitive sciences are concerned with research question of *spatial understanding* and its connection to acting and interacting. While

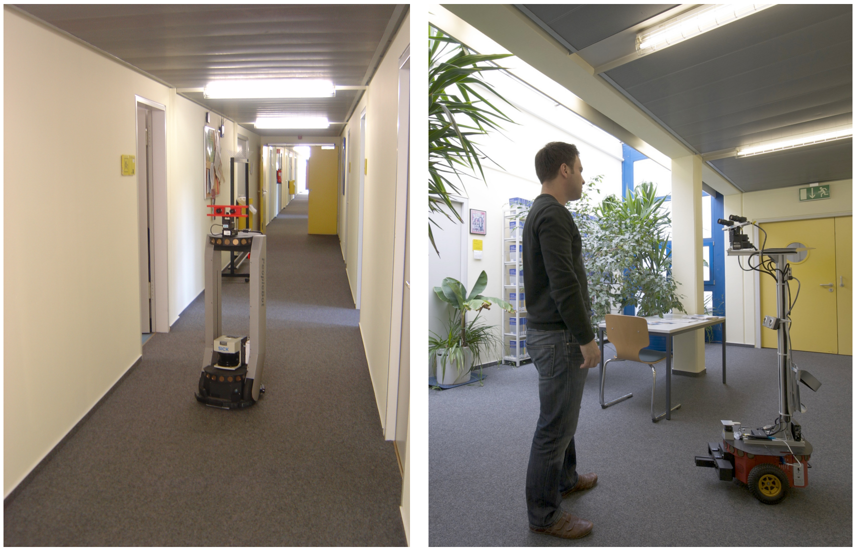


Fig. 1: Autonomous mobile robots – the CoSy Explorer (left) and Dora (right) – operating in an office building.

there has been a lot of progress on the individual, and unrelated, aspects of autonomous robot mapping and navigation on the one hand, and on human spatial cognition on the other, more intelligent and more interactive robots that are supposed to act as assistants or companions for their human users need to bridge that gap.

Unless such a robot is equipped with a form of external localization – such as robots acting in instrumented environments (which, in turn, are faced with their own challenges [20]) – it must be equipped with sensors that allow it to perceive its surroundings. In the simplest case, such sensors are only used to prevent the robot from hitting an obstacle[1] or to enable the robot to move to a fixed target position.[2] This, however, does not amount to much spatial understanding other than a robot-centric frame of reference that captures the here-and-now. An understanding of larger spatial structures requires that the robot at least be able to represent – i.e., remember and retrieve – landmarks that are outside the currently observable part of space.

We are driven by the research question of *spatial understanding* and its connection to acting and interacting in indoor environments. We want to endow autonomous robots with the capability to conduct spatially *situated dialogues*. For this robots must be able to understand space in terms of concepts that can be expressed

[1] For instance, the e-puck educational robot is equipped with eight infrared (IR) proximity sensors, which measure the presence of nearby obstacles [64].

[2] The iRobot Roomba autonomous vacuum cleaner has the capability to find its way to a docking station by sensing the IR signals that the station emits.

in, and resolved from natural language. As soon as human-robot interaction is required, a further spatial *abstraction* from the robot's sensory perception to human-compatible symbols becomes key.

We start from the assumption that the environment is not instrumented in order to facilitate the mapping problem. The kinds of environments that we are interested in are indoor spaces that are designed by humans for humans – and that are intuitively and easily *understood* by humans. This includes ordinary and everyday indoor office environments or apartments that are populated by humans working and living there. We call this class of environments that are made and designed by humans for being used and populated by humans *human-oriented environments*. Figure 1 demonstrates examples of different human-oriented environments in which autonomous agents have to operate. In order to provide some intuition about what a robot's external perception is like, Figure 2 shows how a robot's sensors (cameras and laser range finders) observe its environment.

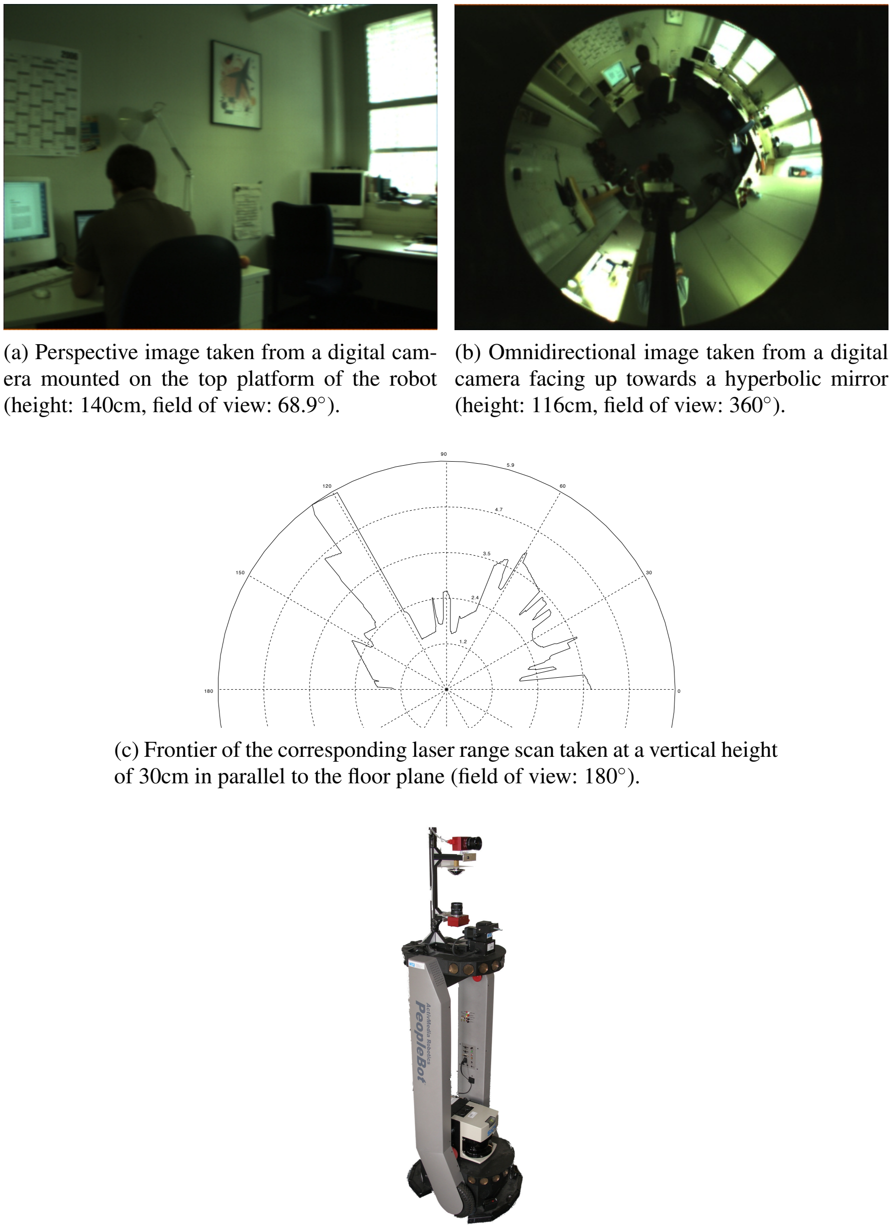
Spatial understanding comprises two aspects. For one, it concerns *categorization* of space. That is, which are the concepts that describe spatial units, and how are they determined? Secondly, it concerns *structuring* of spatial organization. That is, how are the units related that a human-oriented environment is composed of? We call spatial knowledge representations that address these issues *human-compatible representations* of space.

3 Related Work

This work builds upon and extends the author's previous research on *multi-layered conceptual spatial mapping* [97, 100] in the tradition of approaches like the *(Hybrid) Spatial Semantic Hierarchy* by Kuipers *et al.* [50, 51, 5], the *Route Graph* model by Krieg-Brückner *et al.* [92, 46], Buschka and Saffiotti's *hybrid maps* [10], as well as *multi-hierarchical semantic maps* for mobile robots by Galindo *et al.* [28, 27].

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.* [89] suggest a hierarchical probabilistic representation of space based on objects. The work by Galindo *et al.* [28, 27] presents an approach containing two parallel hierarchies, spatial and conceptual, connected through anchoring. Inference about places is based on objects found in them. This approach is based on the Multi-AH-graph model by Fernandez and Gonzalez [22]. The work by Diosi *et al.* [16] creates a metric map through a guided tour. The map is then segmented into discrete rooms according to the labels given by the instructor. Furthermore, the *Hybrid Spatial Semantic Hierarchy* (HSSH), introduced by Beeson *et al.* [5], allows a mobile robot to describe the world using different representations, each with its own ontology.

More recently, Pronobis *et al.* [75] have presented a refined approach to multi-layered mapping, in which, inter alia, the representations of the lower map layers



(a) Perspective image taken from a digital camera mounted on the top platform of the robot (height: 140cm, field of view: 68.9°).

(b) Omnidirectional image taken from a digital camera facing up towards a hyperbolic mirror (height: 116cm, field of view: 360°).

(c) Frontier of the corresponding laser range scan taken at a vertical height of 30cm in parallel to the floor plane (field of view: 180°).

(d) The mobile robot used for acquiring the sensor data. The cameras and the laser scanner can be seen on the top and bottom platforms, respectively.

Fig. 2: Office environment "seen" from the point of view of different robot sensors. Still images and sensor readings taken from the CoSy Localization Database (COLD) [72].

were re-defined, and a probabilistic inference engine is used for reasoning with the discrete symbols in the conceptual map layer.

Lemaignan *et al.* [54] present a similar approach to endowing robots with spatial representations that allow them to act in and talk about their environment. Their framework has the advantage of providing a kind of *theory of mind* that allows the robot to reason about the perspective of its interlocutor in order to disambiguate and ground natural-language instructions. While our approach addresses the specific challenges involved when engaging in dialogues about spatial environments that are larger than what can be perceived at once (cf. Section 4.2), their approach focusses on adequate reasoning techniques for shared visual scenes, like, e.g. tabletop scenarios.

With the availability of affordable 3D sensors and appropriate techniques for using them for robotic mapping purposes, a number of approaches for building layered representations of 3D space have been proposed recently. The KnowRob-Map framework [84] combines low-level metric costmaps, maps of 3D point clouds, and ontological knowledge bases into a semantic environment model of places, object locations, and afforded actions. Pangercic *et al.* [70] use natural-language task instructions from the WWW to construct a Description Logics-based knowledge base for tabletop scenarios. Tenorth *et al.* [83] present a framework that allows mobile service robots to use multiple web-based knowledge sources (including OMICS, WordNet and an internet image search engine) in order to perform everyday manipulation tasks. While these approaches are especially useful for (mobile) manipulation in human-oriented environment (e.g., kitchens [7]), our approach has a stronger focus on human-robot interaction and situated human-robot dialogues.

Viswanathan *et al.* [90, 91] propose another approach that makes use of existing commonsense knowledge resources. They use the LabelMe dataset to train an automated place classifier that relies on the presence of detected objects to infer which other objects are likely to occur nearby and which kind of place (e.g., kitchen or office) is seen in the scene.

4 Background

An important issue in cognitive science, psychology, and linguistics is the question how the mind processes sensorimotor stimuli in order to form abstract representations that are available for higher-level reasoning as well as language production and understanding. A related question is how words, being arbitrary *symbols*, get their meaning and how this meaning is *grounded* in reality, i.e., how words can refer to things and circumstances in the world.

On the lowest level of sensorimotor abstraction, the mind performs *categorization*. Categorization is a basic skill of structuring sensory input by abstraction and simplification. It is an essential capability of every neural system in humans and animals alike, or as Lakoff and Johnson put it, “every living being categorizes,” and every “living system must categorize” [52]. By categorization, it is possible to reduce the complexity of the input by relating it to previous input patterns, i.e., past experiences. With more and more experience, more and more categories are formed, and existing ones are refined. Most of category-forming and categorization is a sub-conscious process, while only a small part of it can be subject to conscious, deliberate cognitive action [52].

Concepts are higher-level cognitive representations of our mental categories. The concept system is accessible for reasoning and inference and thus part of our conscious thinking. Concepts are often formed around *prototypes* – either ideal or average representatives of their concept, or ones that possess only elementary properties. Prototypes allow to draw inferences about category members in the absence of any special contextual information [52].

4.1 Categorizing space

We are concerned with the question of how one can refer linguistically to a spatial structure – e.g., a room, a place, or an object in a specific location – in a given situation, and how one can appropriately act in such a space. Categories determine how people can interact with, and linguistically refer to entities in the world. By naming a referent, people categorize it.

Brown identifies that people in one community prefer the choice of one particular name for classes of things over the many other possible names. “The most common name is at the level of usual utility” [9]. This theory is regarded as the first approach towards the notion of *basic-level categories* further developed by Rosch [79]. 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. It has also been shown that the concept of an object evokes certain expectations about how to interact with it [8]. In a nutshell, *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.

Our work rests on the assumption that the basic-level categories of spatial entities in an environment are determined by the actions they afford. 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 receiving and entertaining guests, spending time with the family, and other recreational and leisure activities. These activities, in turn, can be afforded by certain objects, such as couches, chairs and tables, or TV sets. Living rooms are typically furnished with such objects. This means 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 that room.

We furthermore assume that the basic-level categories that people use to refer to spatial areas are located at one level lower than the more general category ‘room’. Of course, rooms can have proper names and it is common usage in office environments to label rooms systematically, e.g., by assigning unique, ordered numbers, but still it is uncommon in everyday talk that people use these proper names to refer to a spatial entity. People instead refer to rooms with their general names, which correspond to basic-level categories such as ‘kitchen,’ ‘library,’ or ‘lobby.’

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.

4.2 Structuring space

Research in cognitive psychology addresses the inherently *qualitative* nature of human spatial knowledge. It tries to answer the question how the human mind represents spatial information in a so-called *cognitive map*. Following the results of empirical studies, it is nowadays generally assumed that humans adopt a *partially hierarchical* representation of spatial organization [81, 61]. The basic units of such a qualitative spatial representation are *topological* regions [14], 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 [40]. In our approach, topological areas are the primitive units of the conceptual map that is used for human-robot interaction and dialogue, and the basic spatial relation is topological inclusion.

Recent advances in cognitive neuroscience have found evidence for brain structures that supply the topological representations of the so-called “place-cells” with a metric one encoded in the so-called “grid cells” [41]. This does not contradict the assumption that the global-scale representation of *large-scale space* in the cognitive map is a topological one. It rather provides insight into how local scenes, i.e., *small-scale space*, might be represented in the human mind and speaks in favor of a multi-layered, hybrid representation of space in the cognitive map.

Large-scale space and small-scale space

There is an important distinction to make when investigating any kind of spatially situated behavior, be it acting, planning, observing, learning, or communicating, namely if it pertains to space that constitutes the agent’s immediate surroundings (*small-scale space*), or if it pertains to larger spatial structures (*large-scale space*) [39, 38].

Kuipers defines large-scale space as “a space which cannot be perceived at once; its global structure must be derived from local observations over time,” whereas small-scale space consist of the here-and-now. For example, a drawing is a large-scale space “when viewed through a small movable hole, while a city can be small-scale when viewed from an airplane” [49]. In more common everyday situations, an office environment, one’s house, a city, or a university campus are large-scale spaces. A table-top or a particular corner of one’s office are examples of small-scale space.

Segmenting and partitioning space

The physical properties of containers and surfaces belong to the “first and most frequent spatial concepts taught” to children [25]. Since these spatial concepts are among the first to be experienced through our own embodiment, they give rise to the basic cognitive schemata for spatial and metaphorical thinking. The so-called *container schema* represents one of the most pervasive and intuitive spatial relations, namely *containment* [52]. Another schema that is acquired early on is the notion of surface-support, i.e., the *surface schema*. In natural language they are expressed by the topological locatives “in” and “on,” which are among the most frequently used prepositions [15].

As mentioned earlier, it is important that autonomous agents which are supposed to interact with humans in a human-oriented environment have a notion of spatial units that are also meaningful for humans. Topological regions are such units that are meaningful to humans. We call the units of indoor spaces *areas*. We distinguish between two basic kinds of areas. *Rooms* are spatial areas whose primary purpose is defined by the kinds of actions they afford. The other major class of indoor areas are *passages* whose primary purpose is to link rooms and provide access to other spatial areas. This very basic distinction already allows mobile robots to employ a human- and situation-aware motion behavior in the vicinity of humans [95].

The challenge for intelligent agents is to autonomously build spatial representations that are composed of such areas. The previously mentioned distinction between physical, perceptual and subjective boundaries of topological areas corresponds to a *spatial segmentation* along geometric features versus functional features. In indoor environments, walls are the physical boundaries of areas. They determine the geometric layout of the space they surround. Functional features, as mentioned before, can be determined by specific objects – but also by the spatial layout and the composition of the objects and their surroundings.[3] Similarly, the gateways that link areas can be defined geometrically or on a functional-perceptional basis.

However, the sensors of a robot are not particularly geared towards perceiving architectural structures. Neither do computer vision methods exist that allow to vi-

[3] Strictly speaking, the presence of a coffee machine alone does not turn a room into a kitchen – it could as well be a storeroom. The space in the room must afford the preparation of coffee, just as the coffee machine must be reachable and usable.

sually recognize arbitrary objects – let alone their functional affordances. Currently, the main purpose of robotic exteroceptive sensors is to discriminate free space from physical obstacles, and to provide a means for localizing the robot with respect to local landmarks. It is therefore necessary to make use of other cues to *segment* an environment into topological units.

A special kind of free space are geometrically bounded *gateways*. In a spatial representation that is based upon free space and its inter-connectivity, gateways play an important role in structuring and segmenting free space. In a map that only implicitly represents the boundaries of spatial areas, gateways divide space into regions that belong to one spatial area from regions that belong to other spatial areas. "Cognitively this allows the world to be broken up into smaller pieces" [11]. Gateways constitute an important factor for spatial cognition and navigation of autonomous agents in large-scale space [12]. Chown *et al.* [13] explain the special role of gateways for autonomous robots like this:

> "In buildings, these [gateways] are typically doorways; [. . .] Therefore, a gateway occurs where there is at least a partial visual separation between two neighboring areas and the gateway itself is a visual opening to a previously obscured area. At such a [location], one has the option of entering the new area or staying in the previous area."

Later we show how our approach makes use of information about doorways in order to maintain a representation that is composed of rooms as spatial units that correspond to how humans segment indoor environments.

Hierarchical subdivision of space

One prominent spatial relation we experience physically and abstractly every day is spatial *containment*. Egenhofer and Rodriguez [17] consider the space within a room as a small-scale space in which people experience cognitive image schemata, e.g., the *container-surface schema*. However, people routinely employ the same schemata to larger structures, for example when saying "the bench is in the garden" [52]. Similar to objects that are *inside* a room, streets are *in* a city, and several districts form a country. The space around us can thus be decomposed into smaller units, or can combine with other spatial units to larger regions. The container schema can – with a few constraints – also be applied to large-scale space – at least when considering objects of comparable size and similar observation scale [78]. Figure 3 illustrates such a spatial containment hierarchy for a large-scale space environment.

Containment of objects or spatial units is a productive schema for spatial language [15], and one of the structuring principles in the cognitive map [81, 61]. Likewise, hierarchical subdivisions of space are a basic topological relation for *geographical information systems* (GIS) [60, 88].

Topological hierarchies can be expressed as spatial-relation algebras, which, unlike usual computational geometry-based calculations, "rely on symbolic computations over small sets of relations. This method is very versatile since no detailed information about the geometry of the objects, such as coordinates of boundary points or shape parameters, is necessary to make inferences" [17]. This makes them

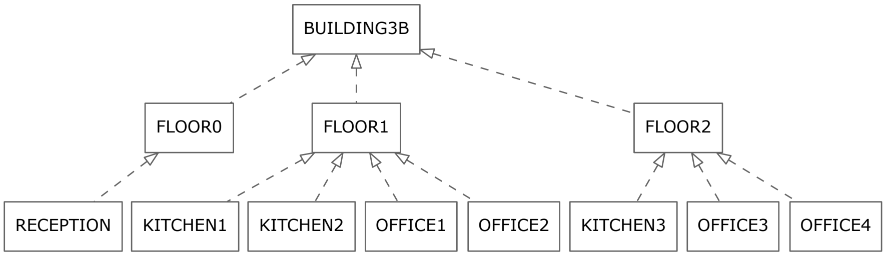


Fig. 3: Example for a hierarchical subdivision of an office environment. The arrows denote the containment relation.

a prime candidate for a basic human-compatible relation to structure and subdivide space.

Conceptually, containment does not form a strict hierarchy. One spatial region can be contained in several different spatial regions, which, in turn, might not be in a containment relation. Consider, for example, an intersection of two corridors. While the intersection itself forms a spatial region, it can also be assumed to be a part of each individual corridor. The representation of spatial abstraction hierarchies is thus rather a *partially ordered set* (poset) [42].

5 Multi-Layered Conceptual Spatial Mapping

If an autonomous robot is required to perform navigation tasks, it must have access to low-level spatial representations that are suitable for fine-grained hardware control. These are typically *quantitative* spatial representations, such as *metric coordinate systems*. Metric maps rely on accurately measurable distances and dimensions. The sensors modern robots are typically equipped with, such as time-of-flight cameras or laser range finders, provide quite exact measurements of free and occupied space in the robot's surrounding. Such sensor readings are hence often stored in metric maps of different kinds.

Humans, on the other hand, use the topological structuring of space to form a more *qualitative* sense of space. This is reflected in natural language, which is full of vague, qualitative spatial expressions. In order to be able to communicate successfully and naturally with humans, robots must be able to establish such a quantitative spatial understanding on the basis of the low-level maps they can build from their sensory input.

To this end, we present *multi-layered conceptual spatial mapping*. The approach addresses the problems of human-compatible structuring and categorization of space. It comprises spatial representations at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations, as illustrated in Figure 4.

In order to address the diverse requirements ranging from “low-level” robot control and “high-level” human-robot interaction, a multi-layered conceptual spatial map can be divided into three major strata. The low-level *metric map layer* encompasses sensor-based maps with specialist representations for robot navigation, localization, and control. The intermediate *topological map layer* provides a basic abstraction of metric space into regions. On the highest level of abstraction, the *conceptual map layer* augments spatial units with human-compatible symbols inside a representation that allows for reasoning and inferencing. In the presented approach, the available spatial information gets coarser while the conceptual knowledge increases with each abstraction step. The details of performing reasoning in the conceptual map layer will be described in Section 6.

5.1 The metric map layer

The lowest level of robot mapping (also referred to as the *sensory map layer*) typically makes use of *metric maps* that serve the principal purpose of allowing the robot to safely navigate its environment while staying localized within its representation of large-scale space.

This self-localization can be performed in an absolute frame of reference or in a relative frame of reference with respect to a local landmark. Different existing approaches to robot mapping of large-scale space hence generate metric maps of different sizes. While several approaches construct global metric maps of the whole operating environment [24, 23, 77], there is a tendency to reduce mapping complexity by representing larger environments by means of interrelated local maps [6, 43]. Many of these metric maps are constructed using the Simultaneous Localization and Mapping technique (SLAM) [55]. This has the consequence that the features of the spatial representation are typically only *meaningful* with respect to the algorithms that work on these representations. These include, for instance, occupancy grid maps (cf. Figure 5a), which address the challenge of representing which parts of an environment are likely to be free and unobstructed, and which ones contain potential obstacles [85], or line maps that represent static features of the environment that facilitate SLAM (cf. Figure 5b).

As a result, such maps are essentially metric representations of positions of free versus occupied space, rather than faithful models of the architectural structure around that free space. In contrast to this, what we need are human-like features. In order to be able to talk in and about space, the agent needs to abstract from its internal, machine-compatible representations of space to a level that is at least comparable to the way humans perceive of space.

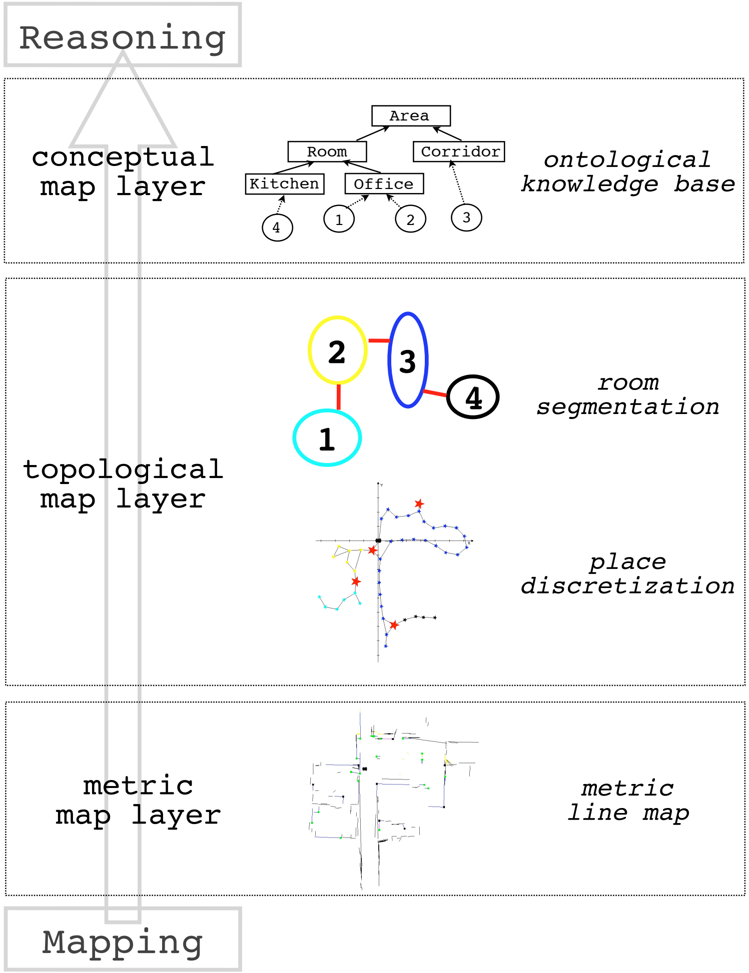
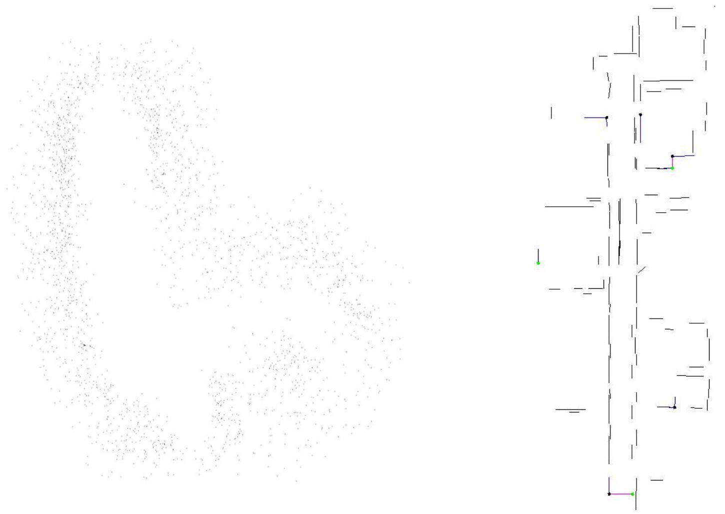


Fig. 4: The different layers of a multi-layered conceptual spatial map.

5.2 The topological map layer

In order to allow for efficient path planning it is common practice to abstract away from sensor-based metric maps. The first abstraction step is *discretization* of the continuous metric space. Examples of such a discretization are *free-space markers* [53, 69] which are used to form a *navigation graph map layer* in the implementation of the CoSy Explorer [100]. Recently [74] introduced the notion of *places* to form an intermediate map layer, which is part of the integrated robotic system Dora [34]. Such representations of the connectivity of free space provide the input to efficient graph-based path planning algorithms, e.g., A*. Often, such graph nodes (i.e.,



(a) Underwater grid map of a marina in San Pere Pescador, Spain [77, 45].

(b) Sparse line-feature map of an indoor environment [100].

Fig. 5: Examples of robotic spatial representations for SLAM.

distinct places) are augmented with semantic mapping information based on *place classification* [66] or *place categorization* [73].

This level of discretization provides a basic notion of the topological structure of an environment. However, the discrete units are not guaranteed to be meaningful to humans. It is thus necessary to aggregate the units of the intermediate layer into *human-compatible spatial units*, such as rooms. This then provides a *topological partitioning* that can be used for *human-compatible structuring* and *categorization* of space. In this view, the exact shape and boundaries of an area are irrelevant. Basic notions that are represented in such a map are *adjacency* and *connectivity*.

As discussed earlier in Section 4.2, the boundaries between such human-compatible units can be established on the basis of gateways, based on geometric, or perceptual features.

Martinez Mozos *et al.* [66] extract a topological semantic map from a metric one using appearance-based features derived from laser range and visual data. Alternatively, Friedman *et al.* [26] use Voronoi Random Fields for extracting the topological structure from a fully explored grid map. Tapus *et al.* [82] describe an approach to topological segmentation using a Bayesian door detector.

The approaches used in the CoSy Explorer [100] and in the Dora robot system as described in [93, 35] use a door detector for on-line room segmentation during the system's exploration of the environment. Based on the information about the connectivity of places and whether they constitute gateways or not, the topological layer forms rooms by clustering places that are transitively interconnected without passing a doorway. Since door detection can malfunction (e.g., a doorway

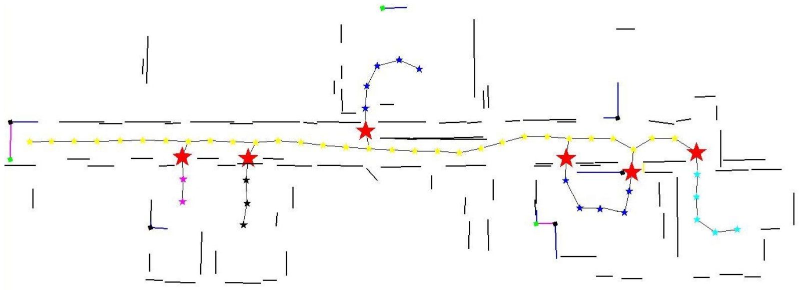


Fig. 6: The line feature map from Figure 5b overlaid with a graph of visited places. The coloring of the nodes indicates their segmentation into rooms, based on detected doorways (large red stars).

is not identified correctly and only later it is found and added to the map; or an erroneous door detection is removed later in the light of new sensor information), room formation must be a non-monotonic process in order to support the potential for knowledge revision. Such an approach can be used as an on-line process that is continuously operating during the robot's run-time, maintaining instances of places and rooms with acquired connectivity relations.

Figure 6 shows a graph of visited places within a metric line-feature map. Figure 7 illustrates the inherent non-monotonic nature of the mapping process we model. (1) shows the initial state. Blue points indicate laser range readings, gray rectangles are walls, and colored circles are (linked) nodes on a navigation graph. If nodes have the same color, they are interpreted as belonging to the same room. (2) shows a sequence of nodes formed after moving around. All nodes belong to a single room (the corridor) because the robot failed to detect the door it was passing through. In (3) the robot has passed through, and successfully detected, a doorway (red node). This triggers the creation of a new room. In (4) the robot has exited this room through another doorway, re-entering the corridor. At this point, the robot is unaware that it has returned to the same corridor as before. Only in (6) nodes become fully connected. Now, the hypothesis for a new room raised in (4) is fused with the already existing corridor hypothesis, creating a single room. In (7), the robot detects the doorway that it had not spotted earlier, i.e., in (2). This leads to a separation of already observed nodes, creating a new room (8).

Together, the intermediate place discretization layer and the topological layer presented above provide an abstraction over continuous, sensor-based metric data. The discrete units of the respective maps (e.g., places, navigation nodes, areas, objects, and landmarks) and the relations that hold between them (e.g., adjacency, inclusion, visibility) serve as the symbolic basis for the *conceptual map layer*.

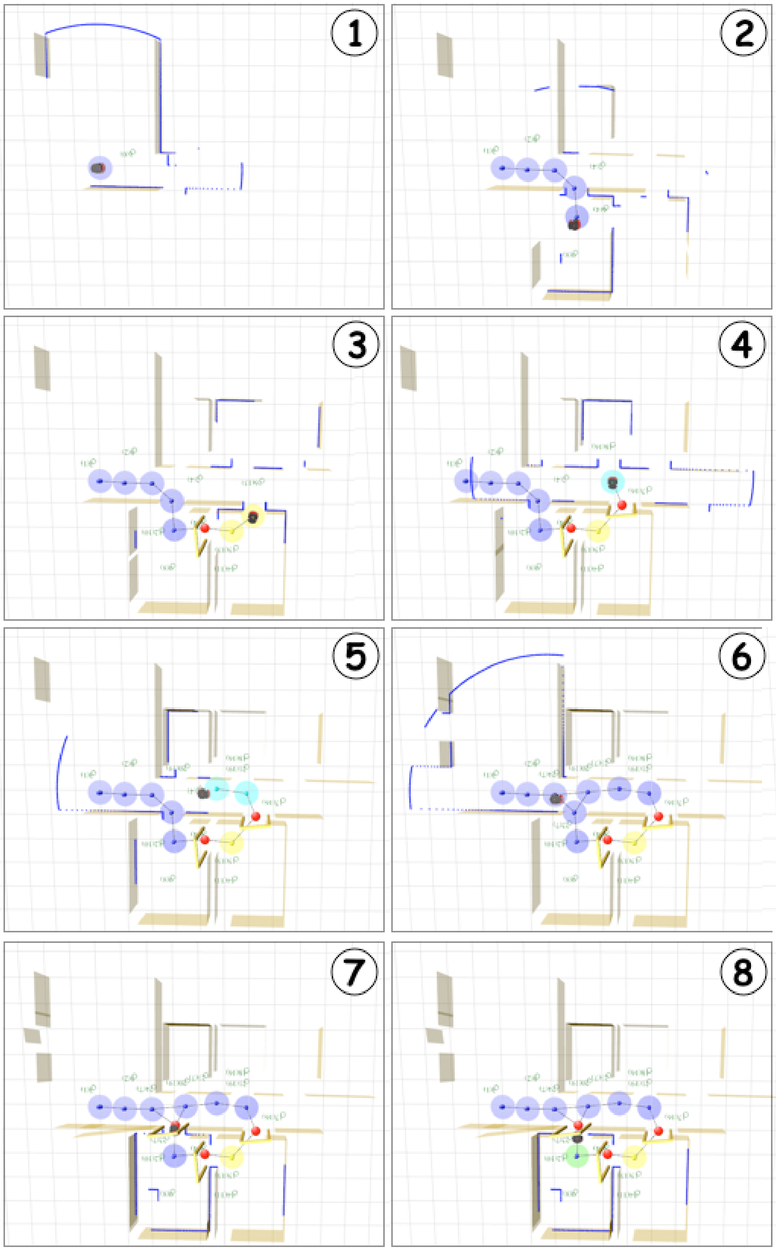


Fig. 7: Exploration sequence of the robot Dora. Red nodes are doorways, colored circles are free space nodes. Nodes having the same color are interpreted as belonging to the same room. Color changes of a node indicates a revision of a room hypothesis, e.g., fusion of nodes into a single room (5 → 6) or separation into a new room after observing a doorway (7 → 8).

Room categorization

In recent years, the task of autonomous sensor-based room categorization for mobile robots has received a lot of attention. Below we will briefly introduce room categorization techniques that have been used with the multi-layered conceptual mapping approach presented here.

A robust room categorization can be achieved by performing local semantic classifications of distinct places inside a room (cf. Section 5.2), and integrating these classifications into a coherent room categorization. The multi-layered mapping approach by Zender *et al.* [100], a predecessor to the work presented here, makes use of a semantic place classifier [66] that can classify a place belonging to either a room or a corridor. A majority vote approach is then used to determine the area category (i.e., either room or corridor).

The approach by Pronobis and Jensfelt [73] is a more recent technique for room categorization that has been used for multi-layered conceptual spatial mapping. Their approach derives properties (shape, size, appearance, topology, and occurrences of certain objects, inter alia) from low level sensory data (laser range scans and camera images). A chain graph representation that supports incrementality and non-monotonicity is then used to perform probabilistic inference of room categories (anteroom, bathroom, computerlab, conferencehall, doubleoffice, hallway, kitchen, meetingroom, professorsoffice, robotlab, and singleoffice). For each room, their approach then estimates the probability distribution of these category labels, over which then the maximum a posteriori estimate is computed in order to obtain the single best room category. More details can be found in [73].

Object detection

Besides information about rooms and their topology, the conceptual map layer needs information concerning the locations and categories of concrete objects that exist in the environment. This kind of information is provided by computer vision modules that are trained on detecting and recognizing certain household or office environments. While these approaches are not discussed in detail, the following pointers might act as starting points for further reading for the interested reader. The Scale Invariant Feature Transform (SIFT) by Lowe [58] is a widely used algorithm for object detection. The multi-layered conceptual mapping instantiation in the CoSy Explorer [100, 29] makes use of a SIFT detector, combined with receptive field cooccurrence histograms (RFCH) [18] for detecting smaller objects according to the approach put forward by Ekvall *et al.* [19]. The Dora robotic system [34] uses the vision algorithms from the BLORT toolkit [65] to detect objects in the images from the robot's cameras. Whenever an object is detected, its existence is stored in the conceptual map along with its detected type and its location (see [100] for more details).

5.3 The conceptual map layer

The basic spatial unit of the conceptual map are the rooms of an environment. The topological segmentation of the environment into rooms is provided by the lower map layers. The conceptual map layer allows to reason about the types of those rooms and about the kinds of objects that they contain. The rooms are assigned their respective class, as determined by the room categorization module (e.g., [66] or [73], see above)

In addition to information stemming from computer vision and sensor-based room categorization, the conceptual map stores and incorporates information that is given to the robot in natural language. This is typically done in situated dialogues with a human user or tutor, e.g., in a so-called *human-augmented mapping* guided tour [86, 87, 71]. This allows, for instance, to represent that the human user said that "this room is the living room." Kruijff *et al.* [48] explain natural language interpretation for human-augmented multi-layered conceptual spatial mapping in more detail.

Using a reasoner, *new knowledge* can then be inferred. The conceptual map affords different kinds of reasoning (see Section 6) in order to provide a human-compatible structuring and categorization of space that can be used for situated human-robot interaction. This reasoning comprises instance knowledge about the given environment as well as conceptual knowledge about indoor environments in general. For example, suppose the robot knows that it is in an area classified as "room" where there is a coffee machine and an oven, and suppose that it has the knowledge that a kitchen typically contains such kitchen appliances, it can then infer that this area is a kitchen. Like this, linguistic references to areas can be generated and resolved even if the robot's observations did not yield complete information. This is a typical case of partial information that occurs, for instance, in human-augmented mapping: the human user shows the robot where the coffee machine is, and later asks the robot to "go to the kitchen" [87].

More information on linking natural language expressions to the spatial representation can be found in [48, 100]. An approach to generating natural language descriptions to spatial entities in large-scale space that makes use of the spatial representations presented here can be found in [98, 99, 96]. Kruijff *et al.* [47] present a comprehensive account on interpreting situated natural language in human-robot interaction.

In the following, we describe how inference over known instances can be performed in the conceptual map, and how hypotheses about unknown parts of the environment can be generated.

6 Reasoning with Changing and Incomplete Spatial Knowledge

Our approach models conceptual knowledge in an ontological taxonomy. It is composed of a commonsense ontology for indoor environments that describes necessary

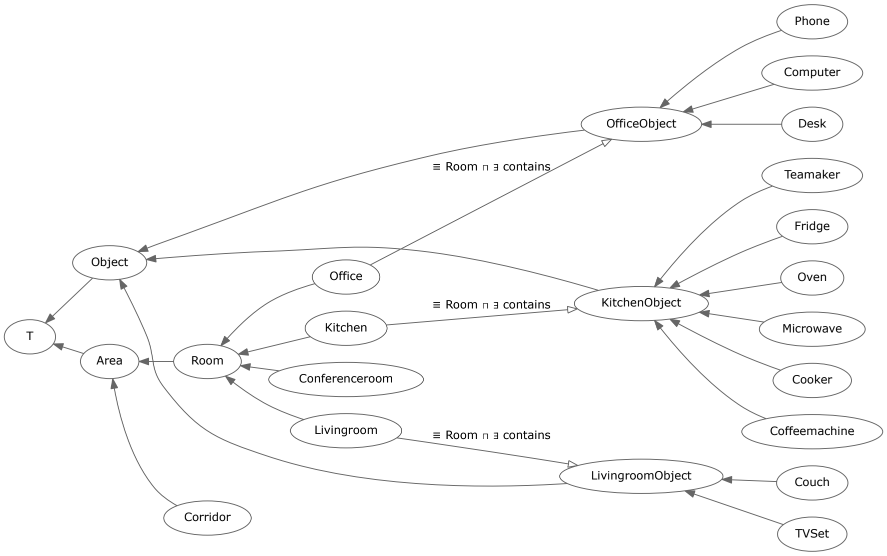


Fig. 8: A part of a handcrafted commonsense ontology of an indoor office environment. Solid arrows denote the taxonomical subclass relation. Labeled edges express that the given subclass of Room is *defined* as being a Room that contains at least one instance of the pointed-to Object subclass. ⊤ stands for the universal top-level concept.

and sufficient conditions that spatial entities must fulfill in order to qualify for belonging to a certain concept. Our definitions of the concepts in the terminological taxonomy are inspired by the way humans *categorize* spatial areas.

6.1 Ontologies

An *ontology* is a formal "explicit specification of a conceptualization" of an area of interest [33]. Ontologies describe classes of objects, their properties, and relations that can hold between them. Ontologies are used to formally define a shared terminology, and to provide a semantic interpretation. They can be used as knowledge base for automated reasoning. *Description Logics (DL)* comprise a family of logical formalisms for ontology-based reasoning. Ontologies are suitable for representing the knowledge about a given domain in a way that is understandable by humans and computers.

In order to allow the robot to draw conclusions about its domain, the conceptual map is equipped with an ontology of an indoor environment. This contains *taxonomies* (i.e., *subclass* relations) of room types, and couples room types to typical objects found therein through *contains* relations. Figure 8 shows the handcrafted commonsense ontology that underlies the previous examples and that was used in the Explorer robot [100].

6.2 Description Logics

We make use of the *Web Ontology Language OWL*[4], more specifically its sublanguage OWL-DL, as the ontology language for the present work because of the availability of different OWL reasoning software, its wide acceptance as a standard for ontology engineering, and the resulting re-usability of resources. OWL-DL is one kind of Description Logic.

Description Logics comprise a whole family of knowledge representations and associated reasoning formalisms that are based on fragments of first-order logic [2]. DL-based knowledge representations distinguish three kinds of knowledge. Firstly, a *taxonomy* of *concepts* represents the so-called terminological knowledge of the domain. This part of the knowledge base is referred to as *TBox*. Secondly, the *ABox* (for assertional knowledge) holds the knowledge about individuals in the domain. We say that an individual a is an *instance* of a concept A if a instantiates A or any of its subconcepts. Finally, DL ontologies contain a set of *roles*, sometimes referred to as *RBox*, that can hold between individuals, and which are defined over concepts. While the TBox expresses general, abstract knowledge of the domain, the ABox contains a description of a specific state of affairs of the world. The interested reader is referred to the Description Logic Handbook [2] for a more detailed account of Description Logics, especially the formal DL syntax and semantics [3].

6.3 Reasoning

Subsumption and instance checking are the standard reasoning services that DL affords [67]. The iterative process in which the different DL reasoning services infer new facts from the TBox, ABox, and RBox axioms is called *expansion*. In pure Description Logics, this is a monotonic process, i.e., the *full expansion* of a knowledge base results from repetitive applications of the DL rules, irrespective of their order. Unless the knowledge base is *inconsistent*, there is exactly one full expansion for a given knowledge base.

Robotic systems, however, are faced with a world that is dynamic and only partially observable at any point in time. Moreover, the robot's perception of the world might be incomplete or error-prone. As a consequence, its representation of the world might be initially false and only over time become more accurate. Spatial knowledge representations for autonomous mobile robots should thus be able to

[4] `http://www.w3.org/TR/owl/`

address these two challenges: *reasoning with changing information*, and *reasoning with incomplete information*.

Belief revision

Belief Revision provides mechanisms for reasoning with changing information [30, 31, 68]. This is the case, e.g., in a world that is not static, or if the agent acquires new information that invalidates older, potentially erroneous information.

The Jena reasoning framework[5] offers built-in OWL-DL reasoning and rule inference facilities. It allows for a basic form of belief revision by re-classifying the knowledge base if a fact is withdrawn from it. This leads to the retraction of previously inferred facts once the conditions that allowed to draw the inference are invalidated.

Default reasoning

The approach presented so far allows a mobile robot to reason about the known parts of an environment, including reasoning with changing information.

However, if the robot needs to reason about (partially) unknown parts of its environment, it is faced with potentially incomplete information. In order to overcome this, and still come up with hypotheses and expectations about the unknown part of its environment, it must be equipped with some form of background knowledge, e.g., about what is typically found where. This, in turn, enables a planning module to infer where an action might have its intended effect.

In one implementation of the Explorer system [37, 80], we presented an approach to deriving *default knowledge* from OWL-DL ontologies. In brief, Default Logic [76] allows to draw *risky* (i.e., potentially false or contradicting) conclusions from a set of certain, but possibly incomplete, facts using rules called *defaults* [1]. Inference from defaults differs from usual entailment in that defaults permit the derivation of their consequences based on the absence of counter-evidence for their truth.

The standard syntax of a default δ is [1]:

$$\delta = \frac{\alpha : \beta}{\gamma}$$

α, β, γ are first-order logic formulae. α is the *prerequisite* of the default rule, β is called the *justification*, and γ is its *consequent*. Informally speaking, a default δ can be interpreted like this: if α is true, and if it is consistent to assume β, then conclude γ.

A special form of default reasoning is *prototypical reasoning*, which expresses typical properties of instances of a concept. This notion is closely related to the

[5] available online at `http://jena.sourceforge.net/`

intuition behind the ontological knowledge representation we chose for our conceptual spatial knowledge base. Below we briefly sketch how generalized introspective mechanisms can be applied to derive defaults from existing OWL-DL ontologies in a principled way. The interested reader is referred to [94] for a more detailed discussion.

Let us start with the following example that expresses the commonsense knowledge that ovens are usually found in kitchens:

$$\delta_{oven} = \frac{Oven(x) \wedge Kitchen(y) : in(x,y)}{in(x,y)} \tag{1}$$

The above default contains free variables. It is a so-called *open default* that represents a set of defaults, where all variables are assigned values. Practically only those substitutions are considered for which the prerequisite is satisfiable, i.e., in our case only oven instances would be used to substitute the free variable x in the first place. The same holds for the other free variable y. Note that this explicitly rules out hypothesizing about unknown *individuals*. Nevertheless such a *closed default* would allow an autonomous robot to hypothesize about the whereabouts of certain objects in case their existence can be assumed. The robot could use this default knowledge to come up with an informed guess where to look first for an oven. This can be helpful both for the purely epistemic goal of achieving a better and more complete knowledge of the world, and for executing a task, like finding a particular object. Hawes *et al.* [37] illustrate how this kind of reasoning helps to infer facts that a symbolic planner can use for goal-directed knowledge gathering, and planning of complex actions.

Such prototypical knowledge is implicitly already represented in OWL-DL knowledge bases. In order to generate a default from a concept definition, we propose to use introspective meta-reasoning over necessary conditions. The concept definition of Kitchen in Figure 8 can be decomposed into the following two necessary conditions: $\mathsf{Kitchen} \sqsubseteq \mathsf{Room}$ and $\mathsf{Kitchen} \sqsubseteq \exists\, \mathsf{contains.KitchenObject}$. On the basis of such concept definitions, open defaults like the following one can be constructed:

$$\delta_{\mathsf{contains}} = \frac{\mathsf{Kitchen(x)} \wedge \mathsf{KitchenObject(y)} : \mathsf{contains(x,y)}}{\mathsf{contains(x,y)}} \tag{2}$$

Following up on the human-augmented mapping example given in Section 5.3, the robot knows which room contains the coffee machine, and it has already successfully inferred that that room (let us call it AREA1) is of type Kitchen. Suppose the robot is then asked to “turn off the oven.” So far the robot hasn’t known anything about the existence of an Oven instance in the environment. The human’s mention of “the oven” warrants the creation of a new symbol (let us call it OBJ5) of type Oven in the knowledge base. This process is called *presupposition accommodation*, and reflects the fact that if someone intends to make a felicitous reference to an object, then that object must exist [44]. Given all this, it makes sense as a case of *prototypical default reasoning* for the robot to assume that the given kitchen object is contained in a known kitchen.[6] The open default above can then be instantiated like this:

$$\delta_{\mathsf{contains}_1} = \frac{\mathsf{Kitchen}(\mathsf{AREA1}) \wedge \mathsf{Oven}(\mathsf{OBJ5}) : \mathsf{contains}(\mathsf{AREA1}, \mathsf{OBJ5})}{\mathsf{contains}(\mathsf{AREA1}, \mathsf{OBJ5})} \qquad (3)$$

Note that the substitution of Oven for KitchenObject follows directly from ontological class subsumption. The following tentative fact can then be concluded: contains(AREA1, OBJ5). Through ordinary DL inference ($\mathsf{in} \sqsubseteq \mathsf{contains}^{-}$, *role inversion*), the robot can infer: in(OBJ5, AREA1), allowing the planner to start its search for an oven in the kitchen. An approach to distinguishing tentative facts (derived from applying default reasoning) and crisp facts (pure DL reasoning) in the knowledge base through reification is discussed in [94].

7 Commonsense Knowledge for Situated Action and Interaction

While the ontologies mentioned so far were handcrafted, it is possible to leverage existing resources of commonsense and linguistic knowledge for bootstrapping an ontology of typical indoor places and objects.

Object and location types that are relevant for our domain are taken from the *Open Mind Indoor Common Sense*[7] (OMICS) project conducted by the Honda Research Institute USA Inc. The OMICS project offers a large collection of user-submitted common-sense facts that were collected with the express aim of making indoor mobile robots more intelligent. Its advantage is its focus on indoor household environments, which makes it valuable for our purposes.

The OMICS locations database (henceforth referred to as OMICS-L) comprises more than 5,800 user-given associations between common everyday objects (ca. 2,900 unique types) and their typical locations (ca. 500 unique types). This provides us with a rich set of objects and locations, and their typical co-occurrence, that are relevant for intelligent mobile indoor robots. Table 1 shows an excerpt from OMICS-L.

Using the thus collected object and location terms, we then construct a taxonomy that puts these terms in a subclass/superclass relation using the *WordNet* lexical database[8]. WordNet [62, 21] is an extensive lexical database of English that has found wide use for the word sense disambiguation task, e.g., [57, 4], but has also been employed in the context of robotics [59].

[6] Of course it is impossible to know for sure – without actually trying to perceptually verify its truth. That is why it is not desirable to add the consequents as crisp facts to the OWL-DL knowledge base, but instead only make it available as *tentative* fact to the planning domain [80].

[7] http://openmind.hri-us.com/

[8] http://wordnet.princeton.edu/

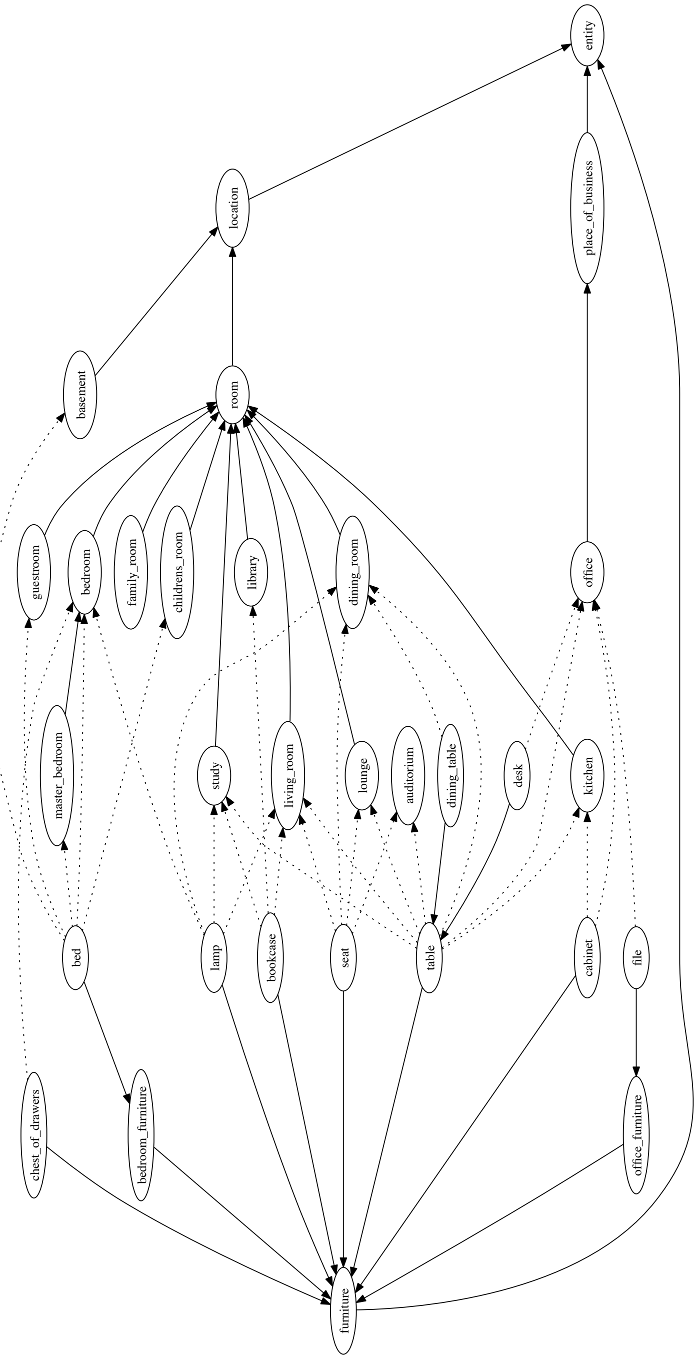


Fig. 9: Part of the WordNet taxonomy with OMICS-L-asserted co-occurrences.

Table 1: Some contributor asserted occurrences of ‘sink’.

object	*location*
sink	kitchen
sink	laundry
sink	washroom
sink	bathroom
sink	restroom
sink	garage
sink	bar
sink	laundry room

The WordNet hypernym/hyponym relation constitutes a taxonomical order over synsets. Combining the lexical taxonomy of WordNet with the commonsense knowledge from OMICS-L therefore allows us to relate object and location types in an ontology. In order to restrict the ontology to cover commonsense knowledge about the indoor environment domain, we performed a bottom-up taxonomy extraction based on the distinct synsets in OMICS-L.

Figure 9 shows a small and sparse subset of the extracted taxonomy, along with links that express OMICS-L co-occurrence statements.[9]

The resulting ontology is stored as an OWL-DL domain ontology. It can readily be used as symbolic spatial knowledge base in the conceptual map of our multi-layered conceptual spatial mapping approach, and it can be manually aligned with the smaller handcrafted ontology described earlier.

7.1 Quantifying commonsense knowledge

The approach to default reasoning described earlier relies on the presence of Description Logics concept definitions like the ones in our handcrafted ontology. Such a kind of information can not necessarily be assumed to be present in the automatically built ontologies presented above. Moreover, there are many kinds of objects that are not uniquely *defining* for the kind of room they are located in. On the other hand, knowing what kinds of rooms exist in the environment, gives indications of where certain objects are more likely to occur than elsewhere.

As an example, suppose that the robot is asked to “find a cornflakes box.” Such boxes can be in many places: in the cabinet, in the storeroom, in the kitchen, in the dining room, but it could also have been left on someone’s desk – all of which are possible, but arguably not equally likely. A planner that is able to deal with probabilities assigned to facts [32] can then come up with a plan that efficiently searches an environment for, e.g., the cornflakes box. A detailed description and discussion

[9] Synsets have been replaced with their associated word labels for ease of reading. Solid arrows denote hyponymy (WordNet), dotted arrows denote a co-occurrence assertion (OMICS-L).

of this scenario can be found in [34]. Here, we focus on obtaining quantifiable expectations about object occurrences in the conceptual mapping approach.

The aforementioned OMICS-L database provides us with a rich set of objects and locations that are relevant for intelligent mobile indoor robots. However, their respective likelihoods are not quantified. E.g., normally each bathroom has at least one sink, whereas some garages do contain a sink and some do not (cf. Table 1). In order for a robot to make judgments about whether it should start searching for a sink in the bathroom or in the garage, it must have quantitative information that allow it to assess the expected outcomes of searching either of these rooms.

Using quantitative priors for the likelihood of an observation of the respective object in a given set of rooms, a decision theoretic planner, e.g., the *switching planner* [32] used for the robot Dora [34], is able to decide about the prioritization of the tasks. To construct such a prior, we obtain *co-occurrence frequency estimates* for all unique object types o (e.g., 'milk') with all unique location types l (e.g., 'office') in OMICS-L by counting the number of hits an image search engine[10] returns when resolving 'o in the l' queries for each of the 1.5 million object-location pairs $\langle o, l \rangle$. Writing $\#q(o\&l)$ for the number of hits returned by that query, and $\#q(l)$ for the number when we query the noun term l alone, then the *co-occurrence prior* $c(o,l)$ that o is located in l is given by Equation 4.

$$
\begin{aligned}
& c(o,l) = \left(\frac{\sqrt{\#q(o\&l)}}{\sqrt{\#q(l)}} \right)^{B} \\
& \text{with } c(o,l) \text{ ranging over } [0,1], \\
& B{=}\frac{1}{2} \text{ if } (o,l) \text{ in OMICS-L, else } B{=}1
\end{aligned}
\tag{4}
$$

We avoid using the raw frequencies from the search engine results to mitigate the problems of: (1) occluded objects being underrepresented in image search queries – e.g., cups are stored in cupboards, and (2) image search queries are often biased to human interest, and omit the mundane and ordinary – e.g., ducks and baths are common, however faucets and baths are rarely mentioned together. We mitigate those problems by first applying the square root function to the counts, and then boost counts selectively using B. The resulting prior then arguably better expresses commonsense knowledge which a contributor to the OMICS project considered relevant for intelligent indoor robots. Table 2 shows some examples of the obtained co-occurrence priors.

Word sense disambiguation

The resulting co-occurrence matrix associates object *words* with location *words*. Consequently it suffers from the vagueness that penetrates natural language. As

[10] http://images.bing.com, September–October 2010

Table 2: Co-occurrence matrix $c(o,l)$ for some select objects and locations.

	kitchen	bathroom	garage	office
sink	0.394958	0.24747899	0.053361345	0.05630252
faucet	0.45874125	0.40419582	0.018181818	0.043776225
computer	0.048387095	0.028830646	0.019112904	0.111693546

mentioned earlier, the same word can have many different meanings, depending on the context. Although the domain of interest already restricts the context, most mentioned words can still denote different concepts in the indoor domain. Consider, e.g., the word 'fan', cf. Table 3. It is quite clear that 'a device for creating a current of air by movement of a surface or surfaces' is meant, rather than 'an enthusiastic devotee of sports,' or 'an ardent follower and admirer.'[11] While in that case the indoor domain provides enough context to disambiguate the different meanings of the word 'fan', consider the word 'keyboard', cf. Table 4. We as humans know that there is one kind of keyboard that is a musical instrument, and that there exists a different kind of keyboard that constitutes a computer input device and which shares only rather superficial properties with the musical instrument.

Table 3: Contributor asserted occurrences of 'fan'.

object	*location*
fan	bedroom
fan	den
fan	kids room
fan	entryway
fan	office
fan	living room
fan	attic

In order to address the ambiguities stemming from word polysemy we make use of the WordNet[12] resources. To disambiguate between different senses of the mentioned words, we linked the OMICS-L terms with WordNet synsets in the spirit of a semantic concordance [63], in which every noun occurrence is tagged with its corresponding word sense.

The tagging was done manually. For each term in OMICS-L, all possible synsets along with their WordNet definition glosses were displayed. An annotator then had to select the appropriate word sense. In order to overcome the problem of typos and spelling errors (e.g., 'jitchen' for 'kitchen') present in the OMICS-L collection, we

[11] Definitional glosses are taken from WordNet.

[12] `http:/wordnet.princeton.edu/`

Table 4: Contributor asserted occurrences of ‘keyboard’.

object	*location*
keyboard	bedroom
keyboard	office
keyboard	study
keyboard	basement
keyboard	computer room
keyboard	computer lab
keyboard	music room

employed a Levenshtein-distance based [56] comparison of word forms that could not otherwise be resolved to words in WordNet.

The version of OMICS-L that we worked with contains 6,293 *object-location* statements. Overall these contain 3,338 distinct words, out of which there are 2,906 distinct object and 509 distinct location terms.[13] For 942 words (908 words mentioned as object, 75 words as location), corresponding WordNet synsets could be found. Among these words there are polysemous words, so that in total OMICS-L was tagged with 1,372 distinct WordNet synsets – among which there are 1,264 object synsets and 157 location synsets.

They cover 3,034 of the total 6,293 *object-location* pairs. The remainder comprises both noise (i.e., nonsensical statements that found their way into the OMICS-L database) as well as concepts that have no clear counterpart in WordNet. A large portion of the latter are sub-concepts of existing senses, most of which are expressed by compound nouns, e.g., ‘ceiling fan’, ‘computer room’, or ‘printer paper’. Note that an *object-location* pair did not count as linked when at least one of the two words could not be resolved to a synset. We call the WordNet-tagged subset of the OMICS-L OMICS-L_{WN}.

As mentioned earlier, we are not just interested in aligning the user-contributed *object-location* statements from OMICS-L with WordNet synsets. We want to go one step further and disambiguate the OMICS-based matrix of co-occurrence priors. However, only a negligibly small portion of all image data on the world wide web is tagged with WordNet senses, and by far the biggest part is just indexed with natural language words. The queries that we retrieve from the image search engine thus have to be expressed with (ambiguous) words.

Since we are unable to inspect the image search results, we cannot determine the actual contribution and relevance of each possible synset for the resulting co-occurrence prior. In order to establish informative priors for the full OMICS-L_{WN} co-occurrence matrix, we need to revise Equation 4.

The object types o and location types l under consideration are, more precisely put, words. We write s_o and s_l if s_o is a synset of o and s_l is a synset of l, respectively.

[13] The discrepancy of the sum of the distinct objects and locations to the total number of distinct terms is a result of several words appearing both as objects and as locations.

We then compute co-occurrence pairs of object synsets with location synsets from OMICS-L$_{\text{WN}}$ as $c(s_o, s_l)$ according to Equation 5.

$$c(s_o, s_l) = \left(\frac{\sqrt{\#q(o\&l)}}{\sqrt{\#q(l)}} \right)^B \tag{5}$$

with $c(s_o, s_l)$ ranging over $[0,1]$,
$B = \frac{1}{2}$ if (s_o, s_l) in OMICS-L$_{\text{WN}}$, else $B=1$

By selectively boosting the intended word senses of the user contributed OMICS-L assertions, we achieve a context sensitivity of our quantitative data that differentiates on the level of word meanings, rather than word forms.

8 Conclusions

In this chapter, we have presented the principle of multi-layered conceptual spatial mapping. In this approach, spatial knowledge is represented at different levels of abstraction, ranging from low-level metric maps to symbolic conceptual representations. It addresses the diverse needs involved in representing spatial knowledge for situated action and human-robot interaction.

The presented approach comprises a symbolic abstraction layer for conceptual reasoning about the properties of locations in an indoor environment. Being inspired by human cognition, it lends itself to being used with state-of-the-art techniques in situated human-robot dialogue processing in that it can be readily used as a knowledge base for situated natural language generation and interpretation. While it neglects aspects of representing small, shared, visual scenes (tabletop scenarios, so-called small-scale space), its main strength is the abstraction over large-scale spatial environments (e.g., indoor environments consisting of multiple rooms). The approach has been implemented and integrated in a number of autonomous, intelligent, and interactive mobile robot systems.

Acknowledgments

This work has been supported by the European Community under contract number FP6-004250-CoSy in the EU FP6 IST Cognitive Systems Integrated Project "Cognitive Systems for Cognitive Assistants *CoSy*"[14] and under contract number FP7-

[14] `http://www.cognitivesystems.org`

ICT-215181-CogX in the EU FP7 ICT Cognitive Systems Large-Scale Integrating Project "*CogX* – Cognitive Systems that Self-Understand and Self-Extend" [15].

This work makes use of the Open Mind Indoor Common Sense Project data by Honda Research Institute USA Inc.

The author wishes to thank Patric Jensfelt, Óscar Martínez Mozos, Geert-Jan M. Kruijff, Andrzej Pronobis, Kristoffer Sjöö, and Alper Aydemir who have collaborated on different instantiations of the multi-layered conceptual mapping approach. Their contributions are gratefully acknowledged.

References

1. G. Antoniou. *Nonmonotonic Reasoning*. The MIT Press, Cambridge, MA, USA, 1997.
2. F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, Cambridge, UK; New York, NY, USA, 2003.
3. F. Baader and W. Nutt. Basic description logics. In F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, chapter 2. Cambridge University Press, Cambridge, UK; New York, NY, USA, 2003.
4. S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 117–171. Springer Berlin / Heidelberg, 2002.
5. P. Beeson, M. MacMahon, J. Modayil, A. Murarka, B. Kuipers, and B. Stankiewicz. Integrating multiple representations of spatial knowledge for mapping, navigation, and communication. In *Interaction Challenges for Intelligent Assistants*, Papers from the AAAI Spring Symposium, Stanford, CA, USA, 2007. AAAI.
6. P. Beeson, J. Modayil, and B. Kuipers. Factoring the mapping problem: Mobile robot mapbuilding in the Hybrid Spatial Semantic Hierarchy. *International Journal of Robotics Research*, 29(4):428–459, 2010.
7. N. Blodow, C. Goron, Z.-C. Marton, D. Pangercic, T. Rühr, M. Tenorth, and M. Beetz. Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4263–4270, San Francisco, CA, USA, September 2011.
8. A. M. Borghi. Object concepts and action. In D. Pecher and R. A. Zwaan, editors, *Grounding Cognition – The Role of Perception and Action in Memory, Language and Thinking*. Cambridge University Press, Cambridge, UK; New York, NY, USA, 2005.
9. R. Brown. How shall a thing be called? *Psychological Review*, 65(1):14–21, 1958.
10. P. Buschka and A. Saffiotti. Some notes on the use of hybrid maps for mobile robots. In *Proceedings of the 8th International Conference on Intelligent Autonomous Systems (IAS)*, Amsterdam, The Netherlands, March 2004.
11. E. L. Chown. Making predictions in an uncertain world: Environmental structure and cognitive maps. *Adaptive Behavior*, 7(1):17–33, December 1999.
12. E. L. Chown. Gateways: An approach to parsing spatial domains. In *Proceedings of the International Conference on Machine Learning Workshop on Machine Learning of Spatial Knowledge*, pages 1–6, Palo Alto, California, 2000.
13. E. L. Chown, S. Kaplan, and D. Kortenkamp. Prototypes, location, and associative networks (PLAN): Towards a unified theory of cognitive mapping. *Cognitive Science*, 19(1):1–51, 1995.

[15] `http://cogx.eu`

14. A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29, 2001.
15. K. R. Coventry and S. C. Garrod. *Saying, Seeing and Acting – The Psychological Semantics of Spatial Prepositions*. Essays in Cognitive Psychology. Psychology Press, 2004.
16. A. Diosi, G. Taylor, and L. Kleeman. Interactive SLAM using laser and advanced sonar. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, Barcelona, Spain, April 2005.
17. M. J. Egenhofer and M. A. Rodríguez. Relation algebras over containers and surfaces: An ontological study of a room space. *Spatial Cognition and Computation*, 1(2):155–180, 1999.
18. S. Ekvall and D. Kragic. Receptive field cooccurrence histograms for object detection. In *Proc. IEEE/RSJ International Conference Intelligent Robots and Systems, IROS'05*, pages 84–89, 2005.
19. S. Ekvall, D. Kragic, and 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):175–187, March/April 2007.
20. D. Estrin, D. Culler, K. Pister, and G. Sukhatme. Connecting the physical world with pervasive networks. *IEEE Pervasive Computing*, 1(1):59–69, 2002.
21. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA, 1998.
22. J.-A. Fernández and J. González. *Multi-Hierarchical Representation of Large-Scale Space – Applications to Mobile Robots*, volume 24 of *International Series on Microprocessor-Based and Intelligent Systems Engineering*. Kluwer Academic Publishers, Dordrecht / Boston / London, 2001.
23. J. Folkesson, P. Jensfelt, and H. I. Christensen. Vision SLAM in the measurement subspace. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, Barcelona, Spain, April 2005.
24. U. Frese and L. Schröder. Closing a million-landmarks loop. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 5032–5039, 2006.
25. S. M. Freundschuh and M. Sharma. Spatial image schemata, locative terms and geographic spaces in children's narrative. *Cartographica*, 32(2):36–49, 1996.
26. S. Friedman, H. Pasula, and D. Fox. Voronoi random fields: Extracting the topological structure of indoor environments via place labeling. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, Hyderabad, India, January 2007.
27. C. Galindo, J.-A. Fernández-Madrigal, and J. González. *Multiple Abstraction Hierarchies for Mobile Robot Operation in Large Environments*, volume 68 of *Studies in Computational Intelligence*. Springer Verlag, Berlin/Heidelberg, Germany, 2007.
28. C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.-A. Fernández-Madrigal, and J. González. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-05)*, pages 3492–3497, Edmonton, Canada, August 2005.
29. D. Gálvez López. Combining object recognition and metric mapping for spatial modeling with mobile robots. Master's thesis, Royal Institute of Technology, Stockholm, Sweden, July 2007.
30. P. Gärdenfors. *Knowledge in Flux – Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, MA, USA, 1988.
31. P. Gärdenfors. Belief revision: An introduction. In P. Gärdenfors, editor, *Belief Revision*. Cambridge University Press, Cambridge, UK; New York, NY, USA, 1992.
32. M. Göbelbecker, C. Gretton, and R. Dearden. A switching planner for combined task and observation planning. In *Electronic Proceedings of the Workshop on Decision Making in Partially Observable, Uncertain Worlds: Exploring Insights from Multiple Communities at the Twenty-Second International Join Conference on Artificial Intelligence (DMPOUW 2011)*, 2011.

33. T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, Deventer, The Netherlands, 1993.
34. M. Hanheide, C. Gretton, R. Dearden, N. Hawes, J. Wyatt, A. Pronobis, A. Aydemir, M. Göbelbecker, and H. Zender. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*, Barcelona, Catalonia, Spain, July 2011.
35. N. Hawes, M. Hanheide, J. Hargreaves, B. Page, H. Zender, and P. Jensfelt. Home alone: Autonomous extension and correction of spatial representations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA 2011)*, Shanghai, China, May 2011.
36. N. Hawes, M. Hanheide, K. Sjöö, A. Aydemir, P. Jensfelt, M. Göbelbecker, M. Brenner, H. Zender, P. Lison, I. Kruijff-Korbayová, G.-J. M. Kruijff, and M. Zillich. Dora the explorer: A motivated robot. In *AAMAS '10: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 1617–1618, Toronto, ON, Canada, May 2010. International Foundation for Autonomous Agents and Multiagent Systems.
37. N. Hawes, H. Zender, K. Sjöö, M. Brenner, G.-J. M. Kruijff, and P. Jensfelt. Planning and acting with an integrated sense of space. In A. Ferrein, J. Pauli, N. T. Siebel, and G. Steinbauer, editors, *HYCAS 2009: 1st International Workshop on Hybrid Control of Autonomous Systems – Integrating Learning, Deliberation and Reactive Control*, pages 25–32, Pasadena, CA, USA, July 2009.
38. N. L. Hazen, J. J. Lockman, and H. L. Pick, Jr. The development of children's representations of large-scale environments. *Child Development*, 49(3):623–636, September 1978.
39. J. F. Herman and A. W. Siegel. The development of cognitive mapping of the large-scale environment. *Journal of Experimental Child Psychology*, 26:389–406, 1978.
40. S. C. Hirtle and J. Jonides. Evidence for hierarchies in cognitive maps. *Memory and Cognition*, 13:208–217, 1985.
41. K. J. Jeffery and N. Burgess. A metric for the cognitive map: Found at last? *Trends in Cognitive Sciences*, 10(1), January 2006.
42. W. Kainz, M. J. Egenhofer, and I. Greasley. Modeling spatial relations and operations with partially ordered sets. *International Journal of Geographical Information Systems*, 7(3):215–229, 1993.
43. M. Karg, K. M. Wurm, C. Stachniss, K. Dietmayer, and W. Burgard. Consistent mapping of multistory buildings by introducing global constraints to graph-based SLAM. In *Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*, pages 5383–5388, Anchorage, AK, USA, May 2010.
44. L. Karttunen. Presupposition and Linguistic Context. *Theoretical Linguistics*, 1(1/2):182–194, 1974.
45. S. Keshavdas. Grid based SLAM using Rao-Blackwellized particle filters. Unpublished master's thesis, Heriot Watt University, Edinburgh, UK, May 2009.
46. B. Krieg-Brückner, U. Frese, K. Lüttich, C. Mandel, T. Massokowski, and R. J. Ross. Specification of an ontology for Route Graphs. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, and Interaction*, volume 3343 of *Lecture Notes in Artificial Intelligence*, pages 390–412. Springer Verlag, Heidelberg, Germany, 2005.
47. G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, and I. Kruijff-Korbayová. Situated dialogue processing for human-robot interaction. In H. I. Christensen, G.-J. M. Kruijff, and J. L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, chapter 8, pages 311–364. Springer Verlag, Berlin/Heidelberg, Germany, 2010.
48. G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138, March 2007.
49. B. Kuipers. *Representing Knowledge of Large-Scale Space*. PhD thesis, MIT-AI TR-418, Massachusetts Institute of Technology, Cambridge, MA, USA, May 1977.

50. B. Kuipers. The Spatial Semantic Hierarchy. *Artificial Intelligence*, 119:191–233, 2000.
51. B. Kuipers, J. Modayil, P. Beeson, M. MacMahon, and F. Savelli. Local metrical and global topological maps in the Hybrid Spatial Semantic Hierarchy. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA 2004)*, New Orleans, LA, USA, April 2004.
52. G. Lakoff and M. Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York, NY, USA, 1999.
53. J.-C. Latombe. *Robot Motion Planning*. Academic Publishers, Boston, MA, 1991.
54. S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2):181–199, 2012. 10.1007/s12369-011-0123-x.
55. J. J. Leonard and H. F. Durrant-Whyte. Simultaneous map building and localization for an autonomous mobile robot. In *Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems (IROS '91)*, pages 1442–1447, Osaka, Japan, November 1991.
56. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *USENIX Technical Conference*, 1966.
57. X. Li. A wordnet-based algorithm for word sense disambiguation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1368–1374, 1995.
58. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
59. M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
60. R. W. Marx. The TIGER system: Automating the geographic structure of the United States census. *Government Publications Review*, 13(2):181–201, March–April 1986.
61. T. P. McNamara. Mental representations of spatial relations. *Cognitive Psychology*, 18:87–121, 1986.
62. G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
63. G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.
64. F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J.-C. Zufferey, D. Floreano, and A. Martinoli. The e-puck, a robot designed for education in engineering. In *Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions (Robotica 2009)*, pages 59–65, May 2009.
65. T. Mörwald, J. Prankl, A. Richtsfeld, M. Zillich, and M. Vincze. BLORT - The Blocks World Robotic Vision Toolbox. In *Proceedings of the ICRA 2010 Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation*, 2010.
66. O. M. Mozos, A. Rottmann, R. Triebel, P. Jensfelt, and W. Burgard. Semantic labeling of places using information extracted from laser and vision sensor data. In *IEEE/RSJ IROS Workshop: From Sensors to Human Spatial Concepts*, Beijing, China, 2006.
67. D. Nardi and R. J. Brachman. An introduction to description logics. In F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, chapter 1. Cambridge University Press, Cambridge, UK; New York, NY, USA, 2003.
68. B. Nebel. A knowledge level analysis of belief revision. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 1st International Conference (KR'89)*, pages 301–311, Toronto, Canada, May 1989.
69. P. M. Newman, J. J. Leonard, J. D. Tardós, and 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)*, pages 1802–1809, Washington, D.C., USA, 2002.
70. D. Pangercic, R. Tavcar, M. Tenorth, and M. Beetz. Visual scene detection and interpretation using encyclopedic knowledge and formal description logic. In *Proceedings of the International Conference on Advanced Robotics (ICAR).*, Munich, Germany, June 2009.

71. J. Peltason, F. H. K. Siepmann, T. P. Spexard, B. Wrede, M. Hanheide, and E. A. Topp. Mixed-initiative in human augmented mapping. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA2009)*, Kobe, Japan, May 2009.
72. A. Pronobis and B. Caputo. COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)*, 28(5):588–594, May 2009.
73. A. Pronobis and P. Jensfelt. Hierarchical multi-modal place categorization. In *Proceedings of the 5th European Conference on Mobile Robots (ECMR'11)*, Örebro, Sweden, Sept. 2011.
74. A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, and P. Jensfelt. A framework for robust cognitive spatial mapping. In *Proceedings of the 14th International Conference on Advanced Robotics (ICAR 2009)*, Munich, Germany, June 2009.
75. A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, and P. Jensfelt. Representing spatial knowledge in mobile cognitive systems. In *11th International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada, Aug. 2010.
76. R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.
77. D. Ribas, P. Ridao, J. D. Tardós, and J. Neira. Underwater SLAM in man made structured environments. *Journal of Field Robotics*, 25(11):898–921, December 2008.
78. M. A. Rodríguez and M. J. Egenhofer. Image-schemata-based spatial inferences: The container-surface algebra. In S. C. Hirtle and A. U. Frank, editors, *Spatial Information Theory: A Theoretical Basis for GIS (COSIT '97)*, volume 1329 of *Lecture Notes in Computer Science*, pages 35–52. Springer Verlag, Berlin, Germany, 1997.
79. E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1978.
80. K. Sjöö, H. Zender, P. Jensfelt, G.-J. M. Kruijff, A. Pronobis, N. Hawes, and M. Brenner. The Explorer system. In H. I. Christensen, G.-J. M. Kruijff, and J. L. Wyatt, editors, *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*, chapter 10, pages 395–421. Springer Verlag, Berlin/Heidelberg, Germany, 2010.
81. A. Stevens and P. Coupe. Distortions in judged spatial relations. *Cognitive Psychology*, 10:422–437, 1978.
82. A. Tapus, G. Ramel, L. Dobler, and R. Siegwart. Topology learning and recognition using Bayesian programming for mobile robot navigation. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, 2004.
83. M. Tenorth, U. Klank, D. Pangercic, and M. Beetz. Web-enabled Robots – Robots that Use the Web as an Information Resource. *Robotics & Automation Magazine*, 18(2):58–68, 2011.
84. M. Tenorth, L. Kunze, D. Jain, and M. Beetz. KNOWROB-MAP – Knowledge-Linked Semantic Object Maps. In *Proceedings of the 10th IEEE-RAS International Conference on Humanoid Robots*, pages 430–435, Nashville, TN, USA, December 2010.
85. S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, Cambridge, MA, USA, 2005.
86. E. A. Topp, H. Hüttenrauch, H. I. Christensen, and K. Severinson Eklundh. Acquiring a shared environment representation. In *Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI 2006)*, pages 361–362, Salt Lake City, UT, USA, 2006.
87. E. A. Topp, H. Hüttenrauch, H. I. Christensen, and K. Severinson Eklundh. Bringing together human and robotic environment representations – a pilot study. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October 2006.
88. T. Trainor. U.S. Census Bureau geographic support: A response to changing technology and improved data. *Cartography and Geographic Information Science*, 30(2):217–223, April 2003.
89. S. Vasudevan, S. Gachter, V. Nguyen, and R. Siegwart. Cognitive maps for mobile robots – an object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, May 2007.
90. P. Viswanathan, D. Meger, T. Southey, J. J. Little, and A. K. Mackworth. Automated spatial-semantic modeling with applications to place labeling and informed search. In *CRV '09: Proceedings of the 2009 Canadian Conference on Computer and Robot Vision*, pages 284–291, Washington, DC, USA, 2009. IEEE Computer Society.

91. P. Viswanathan, T. Southey, J. J. Little, and A. K. Mackworth. Automated place classification using object detection. In *Proceedings of the Seventh Canadian Conference on Computer and Robot Vision (CRV 2010)*, Ottawa, Canada, 2010.
92. S. Werner, B. Krieg-Brückner, and T. Herrmann. Modelling navigational knowledge by Route Graphs. In C. Freksa, W. Brauer, C. Habel, and K. F. Wender, editors, *Spatial Cognition II*, volume 1849 of *Lecture Notes in Artificial Intelligence*, pages 295–316. Springer Verlag, Heidelberg, Germany, 2000.
93. J. L. Wyatt, A. Aydemir, M. Brenner, M. Hanheide, N. Hawes, P. Jensfelt, M. Kristan, G.-J. M. Kruijff, P. Lison, A. Pronobis, K. Sjöö, D. Skočaj, A. Vrečko, H. Zender, and M. Zillich. Self-understanding and self-extension: A systems and representational approach. *IEEE Transactions on Autonomous Mental Development*, 2(4):282–303, December 2010.
94. H. Zender. *Situated Production and Understanding of Verbal References to Entities in Large-Scale Space*, volume 36 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany, 2011.
95. H. Zender, P. Jensfelt, and G.-J. M. Kruijff. Human- and situtaion-aware people following. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2007)*, pages 1131–1136, Jeju Island, Korea, August 2007.
96. H. Zender, C. Koppermann, F. Greeve, and G.-J. M. Kruijff. Anchor-progression in spatially situated discourse: a production experiment. In *Proceedings of the Sixth International Natural Language Generation Conference (INLG 2010)*, pages 209–213, Trim, Co. Meath, Ireland, July 2010. Association for Computational Linguistics.
97. H. Zender and G.-J. M. Kruijff. Multi-layered conceptual spatial mapping for autonomous mobile robots. In H. Schultheis, T. Barkowsky, B. Kuipers, and B. Hommel, editors, *Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems – Papers from the AAAI Spring Symposium*, Technical Report SS-07-01, pages 62–66, Menlo Park, CA, USA, March 2007. AAAI, AAAI Press.
98. H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová. A situated context model for resolution and generation of referring expressions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 126–129, Athens, Greece, March 2009. Association for Computational Linguistics.
99. H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová. Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1604–1609, Pasadena, CA, USA, July 2009.
100. H. Zender, O. M. Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June 2008.

Semantic map segmentation using function-based energy maximization

Kristoffer Sjöö

Abstract— **This work describes the automatic segmentation of 2-dimensional indoor maps into semantic units along lines of spatial function, such as connectivity or objects used for certain tasks. Using a conceptually simple and readily extensible energy maximization framework, segmentations similar to what a human might produce are demonstrated on several real-world datasets.**

In addition, it is shown how the system can perform reference resolution by adding corresponding potentials to the energy function, yielding a segmentation that responds to the context of the spatial reference.

I. Introduction

In the field of mobile robotics, one of the main goals is the integration of robots into the daily lives of humans, aiding us by carrying out tasks for us at home, at the workplace or in outdoor environments. There are many challenges still to overcome before this vision can become reality, however. One of them is that in order to make sure the robots do the right thing, and in the right place, means of intuitive communication between man and machine are needed – in particular, communication concerning their mutual environment.

Robots will need to parse humans' statements and requests and to formulate their own questions and reports in return, using expressions that can be understood by both human and machine. The treatment of such expressions are the subject of this paper; in particular, those describing different parts of space and which an agent might use for navigation or to carry out specific tasks.

The fundamental assumption adopted herein is that *functional* properties are key to dividing up and referring to the world, see Tversky [1]. An indoor environment is constructed intentionally with different functions compartmentalized: this room for eating, this one for sleeping, this for working; and the words we use to refer to those spaces likewise pertain to those functional distinctions. Consequently, this paper attempts to use functional aspects of space to achieve a subdivision and labeling of 2-D maps that corresponds well to human intuitions.

A. Related work

There has been a great deal of work related to the subdivision of maps into discrete units, in many different contexts. One common approach to discretizing space is by using Voronoi diagrams [2]. Another is partitioning it on the basis of the navigational actions it affords, such as in Kuipers et al. [3]. Milford et al. [4] accomplish a similar structuring using neural networks. Pronobis et al. [5] discuss the general problem of partitioning the world into distinct "places" based on perceptual distinctiveness and spatial relationships.

Given a discretization, the next step is to label the units in some relevant way. Diosi et al. [6] and Milford et al. [7] impose labels externally, through a user that the robot is talking to at different locations. Mozos et al. [8] classify regions using metric features, while Vasudevan et al. [9] utilize spatial relations between objects. A graph-based approach is taken by Friedman et al. [2], by performing place classification based on potentials defined on nodes in a graph, with arity up to 4, making it similar to the framework used in this paper although with a model that is more local, and learned as opposed to specified by functional criteria.

Work that examines the functional properties of space include Kuhn [10], who discusses the problem in general terms on an abstract level; and at the other end of the spectrum Dornehege and Kleiner [11], in which parts of a map are classified according to whether they afford a robot's moving through them, though not using human or linguistic concepts. Also related is Fedrizzi et al. [12] where specific places are defined on the basis of a robot's ability to manipulate objects there. Lastly, a debt is owed to Coventry and Garrod [13] who have pioneered the investigation of functional aspects of spatial relations in language.

This work is also concerned with mapping linguistic expressions to portions of space, although in a limited way. Related work has been done e.g. by Kollar et al. [14], who also use an energy optimization method to determine referents for an expression, and Mandel et al. [15], who choose the referent from among Voronoi nodes using fuzzy functions. Both of the above deal with route descriptions, and not with labeling or segmenting maps. Zender et al. [16] also deal with determining spatial entities referred to by a speaker, by finding the lowest common context in a hierarchy. Here too, the set of potential referents is assumed to be given.

B. Contributions

In this paper, a method is presented by which separate, basic, common-sense criteria of a functional nature, such as may be found in a dictionary, can be combined in a single energy maximization and yield an intuitively reasonable subdivision and labeling of a map. Furthermore it is demonstrated how the same energy maximization can be used to find the referents of a linguistic expression, through translating it into an energy potential in a straightforward way.

The author is with the Centre for Autonomous Systems at the Royal Institute of Technology (KTH), Stockholm, Sweden. This work was supported by the SSF through its Centre for Autonomous Systems (CAS), and by the EU FP7 project CogX and the Swedish Research Council, contract 621-2006-4520

C. Structure

This paper is structured as follows: in Section II the reasoning behind using functionality as the basis for spatial segmentation is explained; Section III outlines the energy maximization framework and the solution algorithm. Experiments on various datasets are described in Section IV and Section V presents their outcomes. Section VI summarizes the paper and discusses future work.

II. Functional properties of space

The basic concept this work is based on is the idea that *function* is key to the way humans understand space, and thus also key to any successful robotic representation intended to interact with humans and human-designed environments.

As an example, consider the concept of a kitchen. For a robot to be able to follow orders from humans in a home environment, it will be necessary for it to understand what the word means. A typical approach is to have a human "tag" points in space with the fact that a region is a kitchen [7]. The tag might be attached to a single point, or a region, segmented out by some independent process – such as using laser scans to detect doorways and grouping places on each side of the doorway into different regions [8]. The tagging might be replaced by using machine learning to train models of different regions' appearance.

However, what makes a kitchen a kitchen at a fundamental level is not its appearance, nor a person calling it "kitchen", but the fact that it is used to prepare (and store and consume) food. An appearance-based model might fail if the kitchen is of a novel layout or unfamiliar design, and an algorithm that uses doorways as cues might fail for a studio apartment, where there is no such clear boundary between "kitchen" and "living room". But if a robot can be made to recognize the potential for the *function* of a kitchen, e.g. food preparation, this will improve its ability to generalize and its capacity to communicate effectively with humans.

The semantic labels humans use for space may also vary depending on context. In the case of the aforementioned studio apartment, sometimes "kitchen" will be used to refer to the part of it that houses the sink and oven, while sometimes "room" will be used of the entire room including the kitchen area. This context-sensitivity is an additional necessary feature of a robot's system for spatial understanding.

In the following section, a framework is presented that attempts to incorporate both functional segmentation and context-sensitive reference resolution.

III. Framework for functional labeling of space

The problem is the following: given a 2-dimensional map of an environment, including an over-segmentation of it into a number of small units, "places", find a combination of clusters of places and labels for these clusters such that all the labels well describe the functional features of the associated place cluster. The map that is given may contain various additional information, such as occupancy data, paths existing between places, and objects associated with places.

A. Basic definitions

The set of all places in the map is termed $\mathcal{P}$. A *region* $\mathcal{R}$ is a set of places: $\mathcal{R} = \{p \in \mathcal{P}\}$.

A *label* L is a linguistic symbol corresponding to a region's perceived functional purpose. Labels used in this paper are "room", "corridor", "entrance", "kitchen", "office".

A *relational label* is a label that additionally refers to another region by its definition. Of the above, "entrance" is relational; an entrance is always an entrance *to* something.

A *labeling* is a set of 3-tuples, each consisting of a region $\mathcal{R}_i$, a label for that region L_i, and a relational index k_i indicating which other region the label relates to if it is relational. The regions are subject to the constraint that each place in $\mathcal{P}$ is in exactly one region:

$$\mathcal{L} = \{\langle \mathcal{R}_i, L_i, k_i \rangle\}, \begin{cases} \bigcup \mathcal{R}_i = \mathcal{P} \\ \bigcap \mathcal{R}_i = \oslash \\ 1 \leq k_i \leq |\mathcal{L}| \end{cases}$$

B. Energy function

Every 3-tuple in a labeling has an associated energy, representing how well that particular label describes that particular group of places. A higher energy means a better fit.

$$E(\langle \mathcal{R}_i, L_i, k_i \rangle) = f(\mathcal{R}_i, L_i, k_i, \mathcal{L}) \in [0, |\mathcal{R}_i|] \tag{1}$$

Note that the energy depends on the entire labeling in general. (It also depends on the map; however, that is considered a constant here and left out of the notation.) Because the number and size of regions can vary arbitrarily, in order to avoid any bias for large or small regions the label energies should be proportional to the size of the region, other things being equal, and the average energy per place be within $[0, 1]$.

The energy function is the sum of the energies of each region in the labeling:

$$E(\mathcal{L}) = \sum_i E\left(\langle \mathcal{R}_i, L_i, k_i \rangle\right) \tag{2}$$

The energies assigned to a label for a given region should correspond to the degree to which that region possesses the functional features that define that label. Features are combined in a weighted sum, where the weights may be negative:

$$\begin{aligned} & E(\langle \mathcal{R}_i, L_i, k_i \rangle) = \\ = \quad & \max\left\{\sum_k w_l(L_i)\phi_l(\langle \mathcal{R}_i, L_i, k_i \rangle), 0\right\} \end{aligned} \tag{3}$$

where ϕ_l is the value of the l^{th} feature, and $w_l(L_i)$ is the weight assigned that feature for label L_i. For example, the food preparation feature has a positive weight for the kitchen label. The label energy is bounded from below to 0, and the weights and features must be such that the per-place energy is in $[0, 1]$ as mentioned previously. The weights used below are selected manually, and would be a suitable object for learning in future work.

C. Labels

Below is a list of the labels used for the experiments in this paper, followed by the formulation of the functional features used.

1) Room: The Oxford English Dictionary (OED) [17] provides this definition of a "room":

> A compartment within a building enclosed by walls or partitions, floor and ceiling, esp. (freq. with distinguishing word) one set aside for a specified purpose; (with possessive) a person's private chamber or office within a house, workplace, etc. [. . .]

The functional aspects focused on in the following are the *enclosure* of a room and the *specified purpose* associated with it (the ownership angle is beyond the scope of this paper as it entails social considerations besides purely spatial ones). Enclosure affords a room protection from outside disturbances and influences, and helps an agent form a definite boundary when speaking or thinking about a region. The room also supports some purpose or task for agents who are in it. It will typically do this through some object or set of objects located in the room, with which an agent interacts. The agent needs to perceive those objects; if it cannot the task functionality is undermined. This is encapsulated in a feature that will be referred to as *perceptual convexity*, meaning that each place in the room is visible from the others.

2) Corridor: The following is the OED's definition of "corridor":

> A main passage in a large building, upon which in its course many apartments open.

Here, the functional aspect implied is *connecting*, i.e. a corridor serves as a main route of communication between different parts of the map.

3) Kitchen:

> That room or part of a house in which food is cooked; a place fitted with the apparatus for cooking.

The focus is here on the function of *cooking*, as supported by specific objects. Having room-like features are also of relevance, although not stated as absolute requirements.

4) Office:

> A room, set of rooms, or building used as a place of business for non-manual work; a room or department for clerical or administrative work. [...]

In this case the function is that of *work*, specifically non-manual work. Again, room attributes appear as non-essential aspects of the term.

5) Entrance:

> That by which anything is entered, whether open or closed; a door, gate, avenue, passage; the mouth (of a river). Also, the point at which anything enters or is entered.

Evidently *entering* is the key aspect here.

D. Features

The above labels make use of the following set of function-related features:

1) Enclosed: The functional feature of being "enclosed" that applies to rooms is treated as follows:

$$\phi_{encl} = |\mathcal{R}| \left(1 - \frac{B_{external}(\mathcal{R})}{B_{total}(\mathcal{R})}\right) \tag{4}$$

where $B_{external}$ is the length of the boundary shared by places in this region and places in other regions, and B_{total} is the total boundary length (excluding internal boundaries between places within the region). This formulation rewards labelings where room-labeled regions are compact and largely delineated by walls. The $|\mathcal{R}|$ factor ensures the energy grows as the size of the region.

2) Perceptually convex: The measure of perceptual convexity within a region is

$$\phi_{perc} = \frac{\sum_{\{p,p'\}\in\mathcal{R}\times\mathcal{R}} Vis(p,p')}{|\mathcal{R}| - 1} \tag{5}$$

where

$$Vis(p,p') = \begin{cases} 1, & \text{if } p \text{ and } p' \text{ are visible from each other} \\ 0, & \text{otherwise} \end{cases}$$

Again, the $|\mathcal{R}| - 1$ term is in order to normalize the energy to the order of the size of the region.

3) Connecting: The connecting function of corridors is evaluated as the number of pairs of places in the map that have a *shortest path* that passes through the (prospective) corridor. If any path passes through multiple places in the corridor it counts multiple times. Thus, places that are crossed by many paths in the map contribute strongly to the connecting function of a region, while "dead ends" do not contribute at all. The feature can be expressed:

$$\phi_{conn} = \sum_{\substack{p\in\mathcal{R} \\ \{p^{from},p^{to}\}\in\mathcal{P}\times\mathcal{P}}} \frac{C(p,p^{from},p^{to})}{C_{max}} \tag{6}$$

where

$$C(p,p^{from},p^{to}) = \begin{cases} 1, & \text{if } p \neq p^{from},\ p \neq p^{to} \\ & \text{and } p \text{ is on the shortest} \\ & \text{path between } p^{from} \text{ and } p^{to} \\ 0, & \text{otherwise} \end{cases}$$

C_{max} is a normalizing constant equal to the highest value of $\sum_{\{p^{from},p^{to}\}} C(p,p^{from},p^{to})$ for any single p.

4) Entering: The entering feature is similarly defined to the connecting feature, except only paths leading to the region specified by the relational index k_i are counted, and paths starting inside the active region are similiarly discounted:

$$\phi_{ent,k_i} = \sum_{\substack{p\in\mathcal{R}_i, p^{to}\in\mathcal{R}_{k_i} \\ p^{from}\in\mathcal{P}\setminus\mathcal{R}_i}} \frac{C(p,p^{from},p^{to})}{|\mathcal{R}_i||\mathcal{R}_{k_i}|} \tag{7}$$

5) Food-preparing: The potential of food preparation is here modeled as a function of the distance to objects needed for the task. Two objects are taken as determinants: "refrigerator" and "stove", although this should only be regarded as an illustration; more study will be needed to determine exactly which objects support the function and to what degree, in humans' minds. The value falls off as a sigmoid with the navigation distance (not the straight-line distance):

$$\phi_{food} = \sum_{p \in \mathcal{R}} \left(\alpha \frac{1+C}{e^{d_1(p)/B} + C} + \beta \frac{1+C}{e^{d_2(p)/B} + C} \right) \quad (8)$$

where B and C are constants determining the shape of the sigmoid, and the d_1 is whichever distance (stove or refrigerator) is smaller, d_2 the larger. This formulation allows a non-zero value even if one object is missing entirely.

6) Working: The working feature is treated analogously to the food-preparing feature, except that there is only one object, "desk" and so only one corresponding term in Eq. 8.

E. Referring expression matching

Maximizing the energy described above serves to produce a context-less labeling of the map. In the following it is explained how a spatial referring expression, such as "the room next to the corridor", can be matched to a part of the map using the same framework.

A *description* $\mathcal{D}$ consists of a set of *attributes* and an n-tuple of regions taken from a labeling, each called an *operand*. n is called the arity of the description. Attributes are similar to labels, but may be defined on more than one region. Each attribute is associated with some subset of the descriptions' n-tuple.

Example: A description of arity 2 might have 3 attributes:

1) Region 1 should be labeled "Corridor" (unary)
2) Region 1 and region 2 should be neighbors (binary)
3) Region 2 should be a room (unary)

This description encodes: "find a room that is next to a Corridor".

Attributes each evaluate to a number $a_i \in [0, 1]$, and their geometric mean is taken as the "fit" of the description:

$$F(\mathcal{D}) = \sqrt[n]{a_1 \dots a_n} \in [0, 1] \quad (9)$$

The energy of the description is the product of its fit and the energy of the corresponding labeling:

$$E(\mathcal{D}) = \gamma F(\mathcal{D}) E(\mathcal{L}) \quad (10)$$

This energy is added to that of the labeling itself, and when this sum is maximized it will tend to assign the n-tuple to regions from the labeling which possess all the attributes – which may involve influencing the labeling such that there exists a match, e.g. by reinterpreting two otherwise separate rooms as a single large room. This effect is desirable, because the description implicitly injects information that the unbiased labeling does not have access to about e.g. how a human user conceptualizes different parts of the map. The weight constant γ determines how strongly the description influences the labeling. Its value will in general depend on the application and the linguistic context; $\gamma = 0.1$ is used in this paper.

Attributes used here are:

- Operand region A should have a specific label
- Operand region A should contain a specific place p^*
- Operand region A should be large
- Operand region A should be located toward a given direction in the map
- Operand region A should be located in a given drection relative to operand region B

IV. Experiments

This section describes experiments done using the above framework, operating on three grid maps: FR079, Intel and SDR (see Figure 1). The maps were thresholded and a morphological closure operation performed to eliminate spurious holes in walls. In order to obtain the initial oversegmentation of places $\mathcal{P}$ that the framework needs, a set of nodes and connections were added manually in the manner of an exploring robot to produce a graph similar to e.g. Mozos et al. [8]. Each free grid cell was then assigned to the closest (via free space) node, forming a place and permitting the computation of border lengths (see Sec. III-D). Objects were also assigned manually to places in two of the three maps, for illustrative purposes. The SDR map was left without objects.

A. Energy maximization

The high-level features making up the energy function make it problematic for standard graphical solving methods. For the purposes of this paper a stochastic method, simulated annealing, was found to provide adequate optimization. Simulated annealing works by taking random moves, and may move against the energy gradient in order to escape local minima, but does so at an ever-decreasing probability as time passes; see Algorithm 1.

All experiments used $T_{start} = 2$ and $T_{end} = 0.001$. The cooling-down rate, κ was set to 0.9998, leading to a step count of circa 40 000.

The perturb function changes the labeling using one of the following moves, picked at random:

1) *Transfer*: A donor region is picked at random, and a receiver region is picked from among the donor's neighbors. Places are transferred from the donor to the receiver until a random trigger stops it, or that entire connected component is transferred.
2) *Split*: A seed place is picked at random from the map, and another seed is picked from the neighbors of that place within the same region. The two seeds then grow competitively within the region, until a random trigger stops the process or that entire connected component is covered. Finally one of the grown seeds is picked at random to generate a new region with a random label.
3) *Relabel*: A random region is picked and given a random new label.
4) *Reassign index*: The relational index k_i of a relational label is set to a new random region

Algorithm 1 Energy maximization procedure

```
begin
  $T := T_{start}$;
  while $T > T_{end}$
      do
      $\mathcal{L}_{new} :=$ perturb($\mathcal{L}_{cur}$);
      if $E(\mathcal{L}_{new}) > E(\mathcal{L}_{cur})$
        then
            $p_{accept} := 1$;
        else
            $p_{accept} := e^{\frac{E(\mathcal{L}_{new})-E(\mathcal{L}_{cur})}{T}}$;
      fi;
      if rand() $< p_{accept}$
        then
            $\mathcal{L}_{cur} := \mathcal{L}_{new}$;
      fi;
      $T := T \cdot \kappa$
  od;
end
```

5) *Reassign description*: If a description is being used, change one of its operands to a new random region

Note that nothing in these rules keeps a region from becoming disconnected in the process. Maintaining a region's integrity comes out of the energy maximization.

After each perturb move above (except #5), additionally the description – if one is in use – is locally optimized by taking each of the regions that was affected by the change, and trying it in the place of each current operand in turn, to see if the description's value is improved by switching. This is done before p_{accept} is computed, and permits the description to effectively steer the labeling toward an optimum for both description and labels.

V. Results

Figure 1 shows the result of a context-less segmentation of the three maps. For the most part, the result accords with what a human might come up with. Some corridors in the upper half of the SDR map are mislabeled as rooms, probably because the many loops make for many alternative paths that "dilute" the connected property compared to the southern corridor. This might be remedied by normalizing that property more locally.

Note that this segmentation comes about purely from commonsense functional semantics, without the training of perceptual models, heuristics such as detected doorways or explicit tagging by humans.

No regions are classified as offices or kitchens even where there is functional support – this is not suprising, since they are also good representatives of *rooms*, and there is no context to decide between them until it is imposed, see below.

A. Description resolution

Below are some examples of reference resolution performed on the maps as described in Sec. III-E. They demonstrate that the functional framework can provide both flexibility and simplicity to spatial reference resolution. The labelings are shown in Figure 2 (note that some are cutouts of the full map).

1) Fig. 2(a): "The eastern corridor" (Operand A: Labeled "corridor"; operand B: Labeled "corridor", located east of A). The expression implies there is at least one other corridor that is less easterly.
2) Fig. 2(b): "A big room" (Operand A: Labeled "room", large size).
3) Fig. 2(c): "A kitchen" (Operand A: Labeled "kitchen"). What is otherwise a single room (Fig. 1(a)) is contextually reinterpreted as a kitchen and another region (because the work function crowds out the food function at the upper end of the room).
4) Fig. 2(d): "The room at place $< p^* >$" (Operand A: labeled "room", contains p^*). Although not part of a context-less labeling (Fig. 1(b)), the best fit was found through extending the room into the corridor.
5) Fig. 2(e): "Entrance to a kitchen" (Operand A: labeled "entrance", relational index must point to B; Operand B: labeled "kitchen").

An example of a failed resolution is displayed in Figure 2(f): "Entrance to a big room". Here the search got stuck in a local minimum, where any move to reduce the size of the room led to an energy decrease.

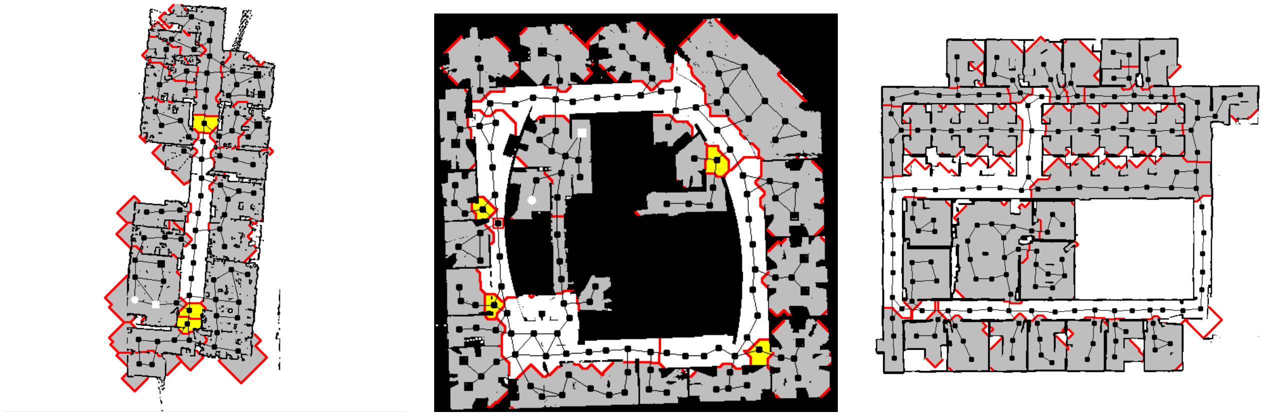
VI. Conclusions

This paper has shown how a conceptually very simple – and, consequently, flexible – energy maximization approach can be used to perform segmentation of 2D maps into units, using features taken from the functional aspects that form the core of spatial semantics. The resulting clusters correspond well to human intuitions. Additionally, it is shown how the framework can use the same mechanism to find matches for referring expressions, even adjusting the segmentation to accommodate the context implicit in those expressions.

A. Future work

The set of different labels used in this work was small. Future work must investigate how increasing the number of possible labels affects outcomes and performance. More complex contexts should also be investigated, as well as the opportunities for combining the framework with language parsing or production. In addition, the different parameters used in the energies for the different labels are good candidates for learning.

The simulated annealing method used for solving the energy in this paper leaves much to be desired in terms of efficiency. It might be worthwhile to explore other approaches; however, because of the general nature of the energies used few simplifying assumptions can be made by any algorithm.

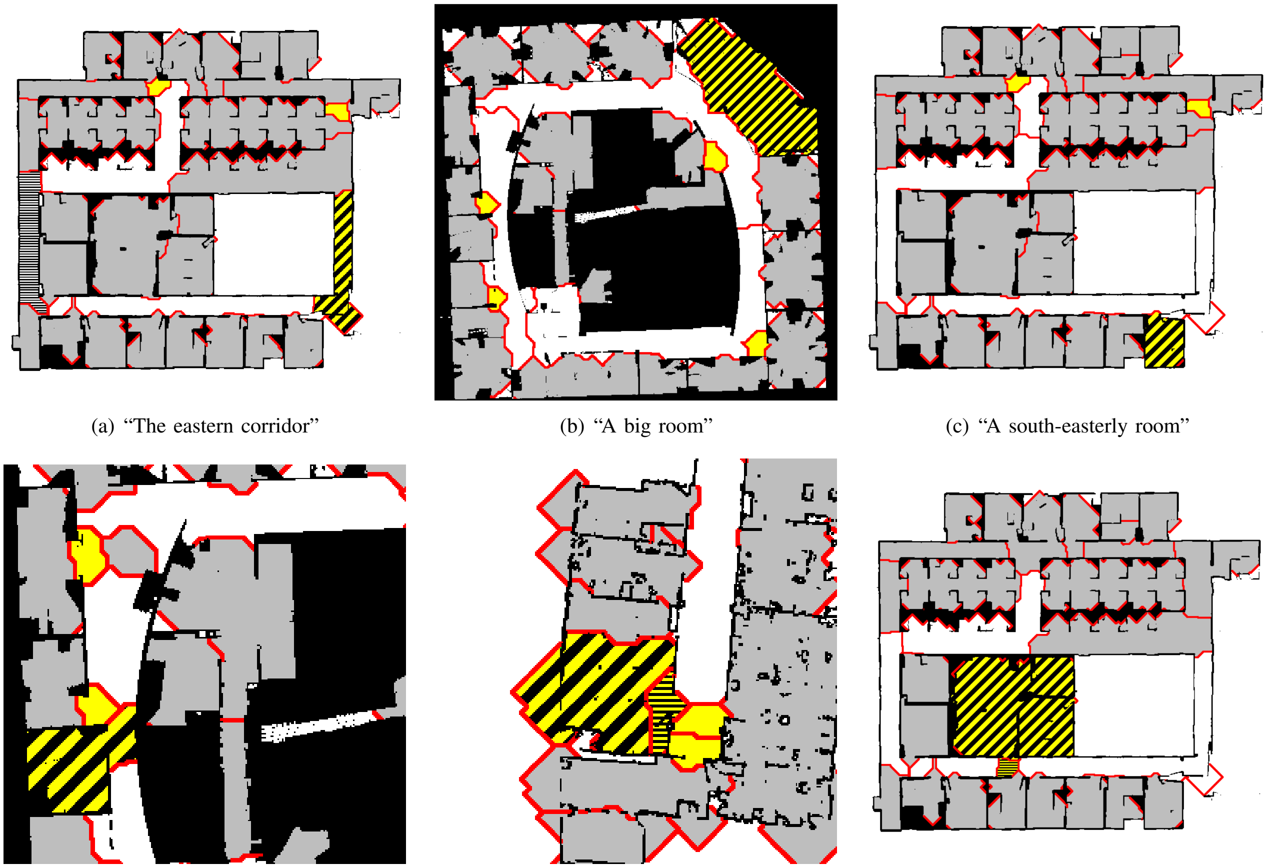


(a) FR079 map

(b) Intel map

(c) SDR map

Fig. 1. Labeling of regions. Grey signifies rooms, white corridors and yellow entrances. Red lines delimit regions. Nodes used to create the places are also shown, with connectivity. A white square represents a refrigerator object; a dot, a stove; a black square, a desk. A red box indicates the place used in description 4 in Sec. V-A



(a) "The eastern corridor"

(b) "A big room"

(c) "A south-easterly room"

(d) "The room at place $< p^* >$"

(e) "Entrance to kitchen"

(f) "Entrance to a big room"

Fig. 2. Fitting descriptions to map. Diagonal stripes indicate the primary operand of the description, horizontal ones the secondary when applicable.

B. Acknowledgements

The data sets used to create the maps used in this paper were obtained from the Robotics Data Set Repository (Radish) [18]. Thanks go to Cyril Stachniss, Andrew Howard and Dirk Hähner for providing this data. The FR07 and Intel maps were obtained from Burgard et al. [19].

References

[1] B. Tversky, "Structures of mental spaces: How people think about space," *Environment and Behavior*, vol. 35, pp. 66–80, 2003.

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

[3] B. Kuipers, J. Modayil, P. Beeson, M. MacMahon, and F. Savelli, "Local metrical and global topological maps in the hybrid spatial semantic hierarchy," in *IEEE International Conference on Robotics and Automation (ICRA 2004)*, 2004.

[4] M. Milford, G. Wyeth, and D. Prasser, "Ratslam: a hippocampal model for simultaneous localization and mapping," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 1, April-1 May 2004, pp. 403–408 Vol.1.

[5] A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, and P. Jensfelt, "A framework for robust cognitive spatial mapping," in *10th International Conference on Advanced Robotics (ICAR 2009)*, June 2009.

[6] A. Diosi, G. Taylor, and L. Kleeman, "Interactive slam using laser and advanced sonar," *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pp. 1103–1108, April 2005.

[7] M. Milford, R. Schulz, D. Prasser, G. Wyeth, and J. Wiles, "Learning spatial concepts from ratslam representations," in *Robotics and Autonomous Systems*, December 2007.

[8] O. M. Mozos, P. Jensfelt, H. Zender, G.-J. Kruijff, and W. Burgard, "From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots," in *Proc. of the Workshop "Semantic information in robotics" at the IEEE International Conference on Robotics and Automation (ICRA'07)*, 2007.

[9] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart, "Cognitive maps for mobile robots – an object based approach," *Robotics and Autonomous Systems*, 2007.

[10] W. Kuhn, *Modeling the Semantics of Geographic Categories through Conceptual Integration*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2002, vol. 2478, pp. 108–118.

[11] C. Dornehege and A. Kleiner, "Behavior maps for online planning of obstacle negotiation and climbing on rough terrain," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[12] A. Fedrizzi, L. Mösenlechner, F. Stulp, and M. Beetz, "Transformational planning for mobile manipulation based on action-related places," in *10th International Conference on Advanced Robotics (ICAR 2009)*, June 2009.

[13] K. Coventry and S. Garrod, *Saying, seeing and acting : the psychological semantics of spatial prepositions*. Hove, 2003.

[14] T. Kollar, S. Tellex, and N. Roy, "A discriminative model for understanding natural language route directions," in *Dialog with Robots: Papers from the AAAI Fall Symposium*, 2010.

[15] C. Mandel, U. Frese, and T. Rofer, "Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, Oct. 2006, pp. 205–210.

[16] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová, "Situated resolution and generation of spatial referring expressions for robotic assistants," in *Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09).*, July 2009.

[17] "The oxford english dictionary." [Online]. Available: http://www.oed.com

[18] A. Howard and N. Roy, "The robotics data set repository (radish)," 2003. [Online]. Available: http://radish.sourceforge.net/

[19] W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kummerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. Tardos, "A comparison of slam algorithms based on a graph of relations," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, oct. 2009, pp. 2089 –2095.

Towards a Cognitive System That Can Recognize Spatial Regions Based on Context

Nick Hawes
Intelligent Robotics Lab
University of Birmingham, UK
n.a.hawes@cs.bham.ac.uk

Matthew Klenk
Palo Alto Research Center
Palo Alto, CA
matthew.klenk@parc.com

Kate Lockwood
ITCD Department
California State University - Monterey Bay
klockwood@csumb.edu

Graham S. Horn
Intelligent Robotics Lab
University of Birmingham, UK
gsh148@cs.bham.ac.uk

John D. Kelleher
Applied Intelligence Research Centre
Dublin Institute of Technology
john.d.kelleher@dit.ie

Abstract

In order to collaborate with people in the real world, cognitive systems must be able to represent and reason about spatial regions in human environments. Consider the command *"go to the front of the classroom"*. The spatial region mentioned (the front of the classroom) is not perceivable using geometry alone. Instead it is defined by its functional use, implied by nearby objects and their configuration. In this paper, we define such areas as *context-dependent spatial regions* and present a cognitive system able to learn them by combining qualitative spatial representations, semantic labels, and analogy. The system is capable of generating a collection of qualitative spatial representations describing the configuration of the entities it perceives in the world. It can then be taught context-dependent spatial regions using *anchor points* defined on these representations. From this we then demonstrate how an existing computational model of analogy can be used to detect context-dependent spatial regions in previously unseen rooms. To evaluate this process we compare detected regions to annotations made on maps of real rooms by human volunteers.

1 Introduction

Consider a janitorial robot cleaning a classroom. While performing this task, it encounters a teacher working with a student. The teacher tells the robot to "start at the front of the classroom", expecting it to go to the front of the classroom and begin cleaning that area. This response requires that the robot is able to *determine the spatial region in the environment that satisfies this concept*.

The ability to understand and reason about *spatial regions* is essential for cognitive systems performing tasks for humans in everyday environments. Some regions, such as whole rooms and corridors, are defined by clearly perceivable boundaries (e.g. walls and doors). However, many regions to which humans routinely refer are not so easily defined. Consider, for example, the aforementioned region *the front of the classroom*. This region is not perceivable using just the geometry of the environment. Instead, it is defined by the objects present in the room (chairs, a desk, a whiteboard), their role in this context (seats for students to watch a teacher who writes on the whiteboard) and their configuration in space (the seats point toward the whiteboard). We refer to such regions as *context-dependent spatial regions* (CDSRs).

Current cognitive systems are not capable of representing and reasoning about CDSRs, yet it is an important ability. If cognitive systems are to collaborate with humans in everyday environments then they must be able to understand and refer to the same spatial regions humans do. Many regions are best defined in a context-dependent manner, for example, a kitchen in a studio apartment, an aisle in a church or store, behind enemy lines in a military engagement, etc. In order to represent and reason about such regions, cognitive systems must integrate different types of information, including geometric, semantic, and functional knowledge. Creating systems able to integrate such a range of information is a key challenge in the cognitive systems paradigm (Langley in press).

This paper presents an artificial cognitive system (specifically a mobile robot) able to represent and reason about CDSRs. Our approach is founded on two assumptions. The first assumption is that CDSRs can be defined using *qualitative spatial representations* (QSRs) corresponding to sensor data of the system (Cohn and Hazarika 2001). The second assumption is that semantically and geometrically similar areas (e.g. two different classrooms) will feature similar CDSRs, and that these similarities can be recognised through *analogy*. The rest of the paper is structured following these assumptions. Section 2 describes how we generate QSRs from sensor data taken from an existing, state-of-the-art, cognitive system and use these to define CDSRs. Section 3 then describes how we use the structure-mapping model of analogy (Gentner 1983) to transfer a CDSR from a labelled example to a new situation. Section 4 presents a worked example of the entire process, and Section 5 evaluates our system in comparison to data from human subjects performing the same task.

2 Metric to Qualitative Representations

The context which defines a CDSR is a combination of the functional and geometric properties of a room, i.e. what can be done there and where. In this work we implicitly represent context using the types of objects present in a room and their location relative to each other. The following sections describe how we construct symbolic representations of these ingredients of context from robot sensor data.

2.1 The Dora System

We base our work on Dora, a mobile cognitive robot with a pre-existing multi-layered spatial model (Hawes et al. 2011). In this paper, we draw on the metric map from this model. For more information on Dora's other competences, see recent papers, e.g. (Hawes et al. 2011; Hanheide et al. 2011).

Dora's metric map is a collection of lines in a 2D global coordinate frame. Two example maps are pictured in Figure 4. Map lines are generated by a process which uses input from the robot's odometry and laser scanner to perform simultaneous localization and mapping (SLAM). Lines in this SLAM map represent features extracted from laser scans wherever a straight line is present for long enough to be considered permanent. In practice, lines are generated at the positions of walls and any other objects that are flat at the height of the laser (e.g. bins, closed doors etc.). The robot's location in the metric layer is represented as a 2D position plus an orientation.

Dora is capable of using vision to recognize pre-trained 3D object models. Recognition can either be triggered through autonomous visual search or at a user's command. When an object is detected it is represented in the metric map by placing a copy of the model at the detected pose. The recognizer associates each object with a semantic type that was provided during a training phase.

To enable us to generate a range of different evaluation situations in a reasonable length of time, we have generated data from Dora in both real rooms and in simulation. Simulation is performed using the Player/Stage hardware abstraction layer (Gerkey, Vaughan, and Howard 2003) allowing us to run the system mostly unchanged in a predefined environment. Also, to enable us to detect a wider range of objects than is usually possible (from armchairs to whiteboards), we used a simulated object recogniser in all runs. The recogniser was configured with types and positions of objects in the environment and was guaranteed to detect them when the robot was orientated towards them. This eliminated any errors from the recognition process, but was still influenced by errors in robot localisation.

2.2 Qualitative Spatial Representation Extraction

For each object that Dora detects we compute the strengths of 8 spatial relations between that object and each of the objects adjacent to it; adjacency is determined using a voronoi diagram, as is standard in geometric reasoning (Forbus, Usher, and Chapman 2003). The strength of a computed relation for a given pair of objects represents the applicability of that relation to the pair. Strength ranges from 0 to 1, with 0 being unsuitable. The model used to compute these relations was inspired by the literature on modeling the semantics of spatial terms (Kelleher and Costello 2009; Kelleher and van Genabith 2006; Regier and Carlson 2001; Gapp 1994). The model accommodates both direction and distance as factors in the relative position of objects.

The relations we compute between each given *landmark* object and its adjacent neighbours are analogous to the cardinal and intermediate points on the compass when the compass is centered on the object. The canonical directions of these relations are defined using the following vectors: $\langle 0,1\rangle$, $\langle 1,1\rangle$, $\langle 1,0\rangle$, $\langle 1,-1\rangle$, $\langle 0,-1\rangle$, $\langle -1,-1\rangle$, $\langle -1,0\rangle$, $\langle -1,1\rangle$. The predicates used to denote these relations are named accordingly, e.g. *xZeroYPlus*, *xPlusYPlus*, *xPlusYZero*, *xPlusYMinus*, etc.

We generate the strengths of these spatial relations as follows. First we compute the maximum distance d_{max} between any two points in the room, this value is used to normalize the distances between objects. Next, taking each object in turn to be the landmark, we translate the origin of the room to the landmark's centroid. This results in the coordinates of the all the other objects in the room being translated into a frame of reference whose origin is the centroid of the landmark. We then compute the strength of each of the 8 spatial relations between the landmark and each of the objects adjacent to it by calculating: (a) the distance d between the landmark's centroid and the adjacent object's location, and (b) the inner angle θ between the direction vector of the relation and the vector from the origin (the landmark's centroid) to the neighbour's location. These two spatial components are integrated to compute the strength s of a given relationship using Equation 1. Figure 1 provides a visualization of a spatial relationship across a region.

$$s = \begin{cases} \left(1 - \frac{\theta}{90}\right) * \left(1 - \frac{d}{d_{max}}\right) & \text{if} \quad \theta \leq 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

These spatial relationships between adjacent objects provide the structure necessary for analogical processing. Generating the relationships in this way (as opposed to, for example, simple coordinate-based thresholding) has the advantage that the presence and absence of relationships is represented on a continuous scale. This provides our representations with the flexibility necessary to manage the variation in perceptual information (i.e. the position of walls and objects) inevitable in human environments and robot perception.

In addition to spatial relations, we also create *grouping entities* from the robot sensor data. Grouping entities collect together sets of adjacent objects of the same type. For example, a classroom would likely have a group entity created in which all of the students' desks were members.

2.3 Representing CDSRs

We use *anchor points* (Klenk et al. 2005) to define the boundaries of CDSRs. Anchor points are symbolic descriptions which link a conceptual entity to a perceived entity. The perceived entities we use are the objects recognised by Dora, and the room itself. The room representation is created by putting a convex hull around the lines in Dora's SLAM map. Anchor points are created from perceived entities using unary functions, e.g. `(XMaxYMostFn Desk1)`

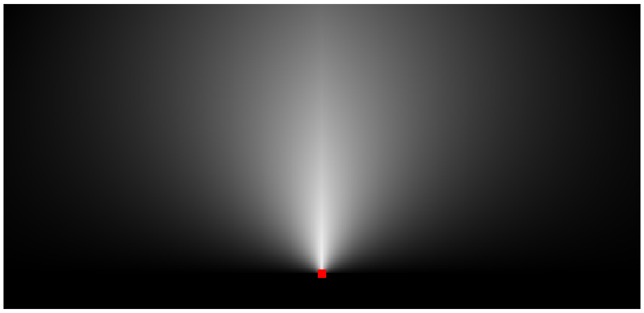


Figure 1: A visualisation of a the strength of a spatial relation across a region. The landmark is the red square and the direction vector used was $\langle 0,1\rangle$ (i.e. above of the landmark). The lighter the pixel the stronger the spatial relation is deemed to be at that point.

represents the point on the Desk1 with the largest x coordinate taken from the set of points with a y coordinate within 5% of the maximum y coordinate. Anchor points are linked to particular CDSRs using a boundarySegment ternary relation. After we have defined the boundary of the region, we assign it a semantic label using the regionType relation. Therefore, each CDSR has one type and a variable number of boundary segments.

```
(regionType CDSR9 FrontRegion)
(boundarySegment CDSR9
    (YMaxXFewestFn Room3)
    (YMinXFewestFn Room3))
(boundarySegment CDSR9
    (YMinXFewestFn Room3)
    (YMinXFewestFn Group1))
```

Figure 2: Three of the five expressions representing the front of the classroom context-dependent region CDSR9

Figure 2 contains three of the five expressions defining the front of classroom Room3 which is pictured in the top of Figure 4. The boundary segments (shown in orange in Figure 4) define the extent of the region. (YMaxXFewestFn Room3) and (YMinXFewestFn Room3) are the points with the highest and lowest y coordinate out of the set of points within 5% of the minimum x coordinate of Room3. The next segment connects the lower left coordinate in the figure to the (YMinXFewestFn Group1), where Group1 includes the eight desks. There are two more boundary segments completing a polygon for this region. The semantic label FrontRegion ties this polygon to a conceptual region, "the front of the room". This definition for the front of the room is specific to Room3 and its entities. It is clearly context-dependent because its extent is dependent on the arrangement of the anchor points used to define its boundary. If the desks were in a different position then the region would cover a different extent (e.g. if they were further to the left then the region would be smaller).

3 Analogical Transfer of Spatial Regions

We assume that a cognitive system will have a way of initially acquiring examples of CDSRs, e.g., by being taught through dialogue, sketching, or hand-coding. To avoid burdening potential users with the task of teaching the system every CDSR individually, it is desirable for a cognitive system to be able to automatically recognize similar regions after initial training. For example, after a janitorial robot has been taught where the front of one classroom is, it should be able to identify the fronts of other classrooms in the building. Our system uses *analogy* to solve this problem. We chose this approach because analogy has been previously used to successfully combine semantic and geometric information in spatial reasoning tasks (Lockwood, Lovett, and Forbus 2008).

Analogy is an essential cognitive process. In humans, analogical processing has been observed in language comprehension, problem-solving, and generalization (Gentner 2003). The structure-mapping theory of analogy and similarity postulates this process as an alignment between two structured representations, a *base* and a *target* (Gentner 1983). We use the Structure-Mapping Engine (SME) (Falkenhainer, Forbus, and Gentner 1989) to perform analogical matching in our system. Given base and target representations as input, SME produces one or more mappings. Each mapping is represented by a set of *correspondences* between entities and expressions in the base and target structures. Mappings are defined by expressions with an identical relation and corresponding arguments. When provided with expression strengths, such as, our spatial relationships, SME prefers mappings with closely aligned fact strengths. SME can be given *pragmatic constraints* that require certain entities in the base to be included in the mapping. Mappings also include *candidate inferences* which are conjectures about the target using expressions from the base which, while unmapped in their entirety, have subcomponents that participate in the mapping's correspondences. SME operates in polynomial time, using a greedy algorithm (Forbus, Ferguson, and Gentner 1994).

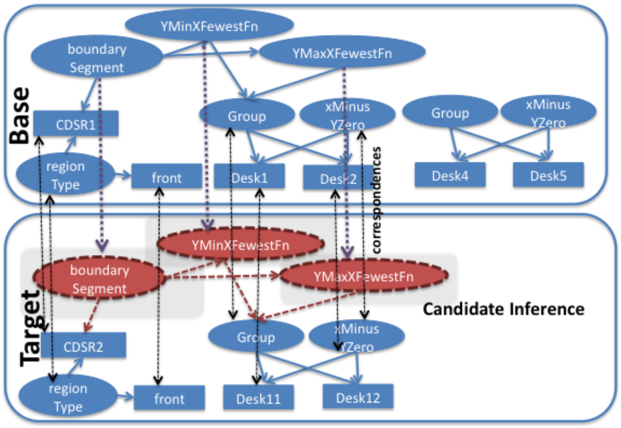


Figure 3: Analogical mapping between six base expressions and three target expressions.

Figure 3 illustrates a sample mapping between six base expressions and three target ones. Each oval represents a predicate, and the entity arguments are represented by squares. SME generates a mapping between the base expressions (group Desk1 Desk2) and (xMinusYZero Desk1 Desk2), and the target expressions (group Desk11 Desk12) and (xMinusYZero Desk11 Desk12) as well as between the regionType expressions in each case in the following manner. First, the predicates of these expressions are placed in correspondence, as identical predicates are preferred by SME. Then SME aligns the arguments of the aligned predicates, Desk1 with Desk11, Desk2 with Desk12, and CDSR1 with CDSR2. While there is another XMinusYZero statement in the base about two desks, it cannot correspond to either of the target expressions in the same mapping due to the one-to-one constraint in SME which allows each element in the target to map to at most one element in the base and vice versa. In Figure 3, the correspondences are highlighted by the hashed bi-directional arrows. Next, SME creates a candidate inference for the boundary segment expression, because both the mapped Group and regionType predicates participate in the mapping. The candidate inference is shown in red in the figure. Note that inference is selective, with no candidate inferences generated for the entirely unmapped expressions.

In our system, the base and target representations consist of the entities Dora has perceived in two different rooms, the QSRs between them and any groups that have been identified. The base also contains a labeled CDSR of the type sought in target. The result of running SME on these representations is a set of correspondences between the base and target, and a set of candidate inferences about the target. We use these to transfer the CDSR from base to target (i.e. recognizing the CDSR in the target) as follows. First, we identify the CDSR of the sought type in the base and use SME's pragmatic constraints to ensure that the entities referred to its anchor points participate in the mapping. To transfer the CDSR to the target, we collect the candidate inferences that result from boundarySegment statements mentioning the base CDSR. The second and third arguments of these candidate inferences are anchor points in the target environment. We use these to define the boundary of the CDSR in the target.

4 Example System Run

To elucidate the workings of our system, we now present an example of how it can transfer a CDSR describing the front of a known classroom (the base) to a new classroom (the target).

We first create the base and target representations by running Dora in the two different classrooms. In each case, Dora is manually driven around the room to allow it to create a metric map. Once the map is created, Dora is then positioned such that the objects are observable and the visual recognition system is run. The map and object data that result from this are then passed on to the QSR generator. The base and target maps are pictured in the top and bottom of Figure 4 respectively. In the base case, Dora perceives 8

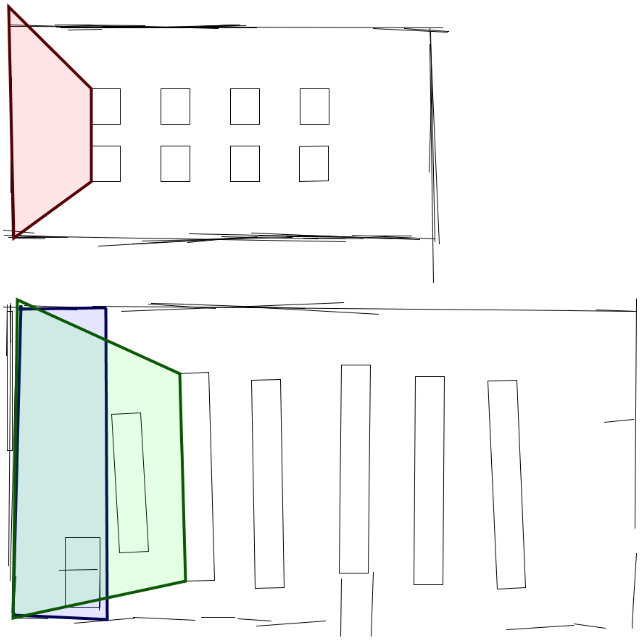


Figure 4: Maps of 2 real classrooms generated by our system. The lines around the perimeter are walls, the unfilled polygons are the outlines of objects and the filled polygons are CDSRs. The maps show an expert-annotated CDSR (red, top image), a subject-annotated CDSR (blue, bottom image) and a CDSR transferred by analogy (green bottom image). The classroom used to generate the bottom classroom is pictured in Figure 5.

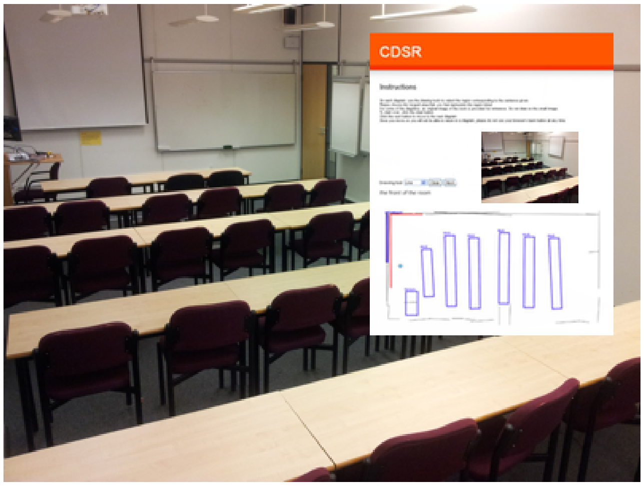


Figure 5: One of the classrooms used in our evaluation. This image was presented to subjects who were asked to annotate a copy of the image in the bottom half of Figure 4. The inset shows a screenshot from the data collection webpage.

individual desks, a group entity containing these desks and the room area. To this we add the CDSR representing the front of the room. The case includes a total of 50 expression relating the 20 entities. Six of these expressions are used to define the boundary segments and CDSR representing the front of the room. The target case includes 26 expressions and 11 entities.

SME generates an analogy between the base and target cases enabling the transfer of the symbolic description of the front of the room to the new situation requiring `Room3` and `Group1` participate in the mapping as they are referenced by the anchor points in the base. The resulting analogy includes 26 correspondences between the entities and expressions and 32 candidate inferences. Four of these candidate inferences define the CDSR in the target with anchor points defined on the room and the group of desks in the target. The green region in the lower image of Figure 4 illustrates the transferred CDSR.

5 Evaluation

To evaluate our progress toward building a cognitive system capable of reasoning about CDSRs, we conducted the an experiment focusing on the following questions:

- Are anchor points able to encode context-dependent spatial regions?
- When provided with a base representation containing a labelled CDSR, how well does our approach identify the CDSR in a given target?

5.1 Materials

We evaluated our approach on six classrooms (two simulated and four real) and two simulated studio apartments. The simulated rooms were based on real-life counterparts. For each room we manually encoded appropriate CDSRs that could be represented by our approach. For the classrooms these were the front and back, and the front and back rows of desks. For the studios these were the kitchen, office and living areas. These manually encoded regions were used as the base CDSRs for analogical transfers, and can be considered the training data for our evaluation.

To determine how people define CDSRs, we asked three naïve users to draw polygons for each region type for each room. This task was performed using a webpage on which each user was presented with an image of the real room plus an image of the map data generated by the robot onto which the drawing could be done. The webpage[1] is shown in the inset in Figure 5. The user-defined polygons define the *target regions* against which we evaluate our transfers.

We consider a *problem instance* to be a room and a sought CDSR type. For each room containing a manually encoded CDSR of the sought type, we generate a *transferred region* using analogical transfer. To assess the quality of the transfer, we calculate precision (p, the proportion of the transferred region that overlaps with the target region) and recall (r, the proportion of the target region that overlaps with the transferred region) as follows:

$$p = \frac{area(transferred\ region \cap target\ region)}{area(transferred\ region)} \tag{2}$$

$$r = \frac{area(transferred\ region \cap target\ region)}{area(transferred\ region)} \tag{3}$$

Using this approach we generate results showing the matches between each of the following pairs of regions: the transferred region and the appropriate target region; the CDSR we manually encoded for the target room and target region; and the region for the whole room and the target region. Results comparing transferred and target regions measure how well our system is able apply a single example to new situations. The comparisons between the manual annotations to the target regions measure how well the anchor points we chose capture the users' regions (who were not constrained to anchor points). Results from the whole room regions provide a baseline performance for comparison.

5.2 Results

To assess overall performance, Table 1 summarizes the results across all problem instances against user-defined target regions from three different users. The transferred regions achieved a precision of .47 (σ=.37) and a recall of .46 (σ=.38). Comparing the manually encoded regions against each target region results in a mean precision of .71 (σ=.30) and recall of .67 (σ=.25). The region defined by the room corresponds to the target region with a precision of .17 (σ=.11) and recall of .98 (σ=.05).

To identify how our approach fairs under different conditions, Table 2 separates the results by CDSR type. The mean precision for the transferred regions ranged from .76 for the front rows of classrooms to 0 for the office in studio apartments. Comparing manually encoded against target regions resulted in a minimum mean precision of .60. This occurred for the front of the classroom. The whole room precision, which is directly proportionally to the size of the target region, varied from .08 for the office to .35 for the living area.

5.3 Discussion

These results support the hypothesis that anchor points can provide a symbolic representation on top of sensor data for context-dependent spatial regions, and, when combined with qualitative spatial relations, they facilitate learning from a single example through analogical transfer. Collaboration with human users requires a high precision and recall, because cognitive systems must be able to understand as well as refer to these regions in human environments. Consequently, the high manually encoded precisions and recalls indicate that the defined anchor points are a reasonable starting point for a symbolic representation. Our future work seeks to further evaluate the utility of this representation by embedding the cognitive system within tasks with human users.

The transferred regions were considerably more precise (.47) when compared to the room as whole (.17), and their recalls (.46) indicate that they captured almost half of the area indicated by the human user. As we create CDSRs

[1] http://home.csumb.edu/k/katherinelockwood/world/

Transferred	Manually Encoded	Entire Room
$\bar{p}$=.47 σ=.37, $\bar{r}$=.46 σ=.38	$\bar{p}$=.71 σ=.30, $\bar{r}$=.67 σ=.25	$\bar{p}$=.17 σ=.11, $\bar{r}$=.98 σ=.05

Table 1: Overall Performance Compared Against Target Regions Defined by Three Users

Region	Transferred	Manually Encoded	Entire Room
Front	$\bar{p}$=.32 σ=.33, $\bar{r}$=.49 σ=.41	$\bar{p}$=.60 σ=.29, $\bar{r}$=.83 σ=.19	$\bar{p}$=.16 σ=.10, $\bar{r}$=1 σ=0
Back	$\bar{p}$=.44 σ=.37, $\bar{r}$=.56 σ=.41	$\bar{p}$=.66 σ=.25, $\bar{r}$=.84 σ=.17	$\bar{p}$=.11 σ=.06, $\bar{r}$=.99 σ=.03
Front Rows	$\bar{p}$=.76 σ=.27, $\bar{r}$=.28 σ=.21	$\bar{p}$=.83 σ=.31, $\bar{r}$=.50 σ=.11	$\bar{p}$=.22 σ=.08, $\bar{r}$=1 σ=0
Back Rows	$\bar{p}$=.72 σ=.30, $\bar{r}$=.42 σ=.26	$\bar{p}$=.80 σ=.29, $\bar{r}$=.43 σ=.26	$\bar{p}$=.19 σ=.06, $\bar{r}$=1 σ=0
Kitchen	$\bar{p}$=.60 σ=.05, $\bar{r}$=.59 σ=.34	$\bar{p}$=.78 σ=.20, $\bar{r}$=.71 σ=.13	$\bar{p}$=.16 σ=.02, $\bar{r}$=.92 σ=.13
Office	$\bar{p}$=.00 σ=.00, $\bar{r}$=.00 σ=.00	$\bar{p}$=.78 σ=.29, $\bar{r}$=.55 σ=.20	$\bar{p}$=.08 σ=.03, $\bar{r}$=.94 σ=.06
Living Room	$\bar{p}$=.40 σ=.39, $\bar{r}$=.01 σ=.01	$\bar{p}$=.63 σ=.34, $\bar{r}$=.54 σ=.13	$\bar{p}$=.35 σ=.22, $\bar{r}$=.96 σ=.06

Table 2: Performance by Region Type

using anchor points defined on perceived entities, our approach performs best when the boundary of the target CDSR is closely tied to such entities. This is the case in the front rows of the classroom, with p of .76 and .82 for the inferred and the manually encoded regions respectively. The system performs worst when the extent of the CDSR is defined as an unbounded area near or around particular objects. The office of a studio apartment is loosely defined as the region around the desk. This motivates one direction of future work: expanding the vocabulary of anchor points to better capture these notions of space.

6 Related Work

Typical approaches to spatial representation for mobile robots tend to focus on localization, and thus mostly represent the world uniformly without subdivision into meaningful (semantic) units (Thrun 2003). When a more structured representation is required, many turn to Kuipers' Spatial Semantic Hierarchy (Kuipers 2000). This paper follows in this tradition, adding CDSRs to his qualitative topological representations. Whilst mobile robots exist which can determine the type of a room from the objects in it (Hanheide et al. 2010; Galindo et al. 2005), they only concern themselves with types of whole rooms, and cannot represent regions within rooms. This is also true for those systems which use some elements of QSR (Aydemir et al. 2011). The need for an autonomous system to ground references to human-generated descriptions of space has been recognized in domains where a robot must be instructed to perform a particular task, however existing systems are restricted to purely geometrically-defined regions (Tellex et al. 2011; Dzifcak et al. 2009; Brenner et al. 2007).

There is mounting evidence that analogy, operating over structured qualitative representations, can be used to simulate a number of spatial reasoning tasks. Forbus *et al.* showed that analogy between course of action diagrams could be used to identify potential ambush locations in new situations by focusing on only the relevant aspects of sketched battle plans (Forbus, Usher, and Chapman 2003). A core contribution of their work was the definition of a *shared similarity constraint* between a spatial reasoning system and its user; where users and spatial reasoning systems agree on the similarities between situations. This has close parallels to what we are trying to accomplish, where a cognitive system is able to reason about context-dependent spatial regions by identifying the same salient features as its human user. The anchor points in our work were originally used in teaching a system how to solve problems from the Bennett Mechanical Comprehension Test that require spatial and conceptual reasoning. For example, identifying which wheelbarrow will be more difficult to lift based on the relative locations of its loads as depicted in a sketch (Klenk et al. 2005). In that work, the anchor points defined the endpoints of lines. We go beyond that result to use anchor points to specify 2D regions.

7 Conclusion

In this paper we presented an integrated cognitive system capable of representing and reasoning about context-dependent spatial regions. The system identifies CDSRs in previously unseen environments through analogy with a single example. This is a difficult cognitive systems task requiring integration of semantic and geometric knowledge to identify regions as small as 8% of the room. Our system demonstrates a successful integration of a range of technologies including vision, SLAM, qualitative spatial reasoning and analogy to achieve this task. In order to make this rich collection of components work together, our work takes a number of short-cuts that we plan to address with future work. These include a reliance on the initial orientation of a room in a global coordinate frame, the lack of a mechanism to retrieve relevant rooms from memory (e.g. MAC/FAC (Forbus, Gentner, and Law 1995)), and a lack of transfer post-processing (e.g. comparing the QSRs present in both base and transferred regions) to improve results. In addition, we must complement our system development work with more comprehensive human studies assessing how people define and use these regions as well as how well anchor points capture them. Despite the preliminary nature of this work, our evaluation demonstrates that the system is able to transfer CDSRs that overlap with user-defined regions for 6 out of 7 region types.

8 Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No. 215181, CogX. The authors would like to thank Jeffery Usher for providing the code to compute voronoi diagrams, Scott Friedman for assistance on probabilistic SME, and Danny Bobrow comments on an earlier draft of this paper.

References

Aydemir, A.; Sjöö, K.; Folkesson, J.; Pronobis, A.; and Jensfelt, P. 2011. Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*.

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, 2072–2077.

Cohn, A. G., and Hazarika, S. M. 2001. Qualitative spatial representation and reasoning: an overview. *Fundam. Inf.* 46(1-2):1–29.

Dzifcak, J.; Scheutz, M.; Baral, C.; and Schermerhorn, P. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA'09)*.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41(1):1 – 63.

Forbus, K. D.; Ferguson, R. W.; and Gentner, D. 1994. Incremental structure-mapping. In Ram, A., and Eiselt, K., eds., *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 313–318.

Forbus, K.; Gentner, D.; and Law, K. 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19(2):141 – 205.

Forbus, K.; Usher, J.; and Chapman, V. 2003. Qualitative spatial reasoning about sketch map. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*.

Galindo, C.; Saffiotti, A.; Coradeschi, S.; Buschka, P.; Fernandez-Madrigal, J. A.; and Gonzalez, J. 2005. Multi-hierarchical semantic maps for mobile robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2278 – 2283.

Gapp, K. P. 1994. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, 1393–1398. AAAI Press.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155 – 170.

Gentner, D. 2003. Why we're so smart. In Gentner, D., and Goldin-Meadow, S., eds., *Language in mind: Advances in the study of language and thought*. MIT Press. 195–235.

Gerkey, B. P.; Vaughan, R. T.; and Howard, A. 2003. The Player/Stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the International Conference on Advanced Robotics (ICAR'03)*, 317–323.

Hanheide, M.; Hawes, N.; Wyatt, J.; Göbelbecker, M.; Brenner, M.; Sjöö, K.; Aydemir, A.; Jensfelt, P.; Zender, H.; and Kruijff, G.-J. M. 2010. A framework for goal generation and management. In *Proceedings of the AAAI'10 Workshop on Goal-Directed Autonomy*.

Hanheide, M.; Gretton, C.; Dearden, R.; Hawes, N.; Wyatt, J.; Pronobis, A.; Aydemir, A.; Göbelbecker, M.; and Zender, H. 2011. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11)*, 2442–2449.

Hawes, N.; Hanheide, M.; Hargreaves, J.; Page, B.; Zender, H.; and Jensfelt, P. 2011. Home Alone : Autonomous Extension and Correction of Spatial Representations. In *Proc. Int. Conf. on Robotics and Automation*.

Kelleher, J. D., and Costello, F. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics* 35(2):271–306.

Kelleher, J., and van Genabith, J. 2006. A computational model of the referential semantics of projective prepositions. In Saint-Dizier, P., ed., *Syntax and Semantics of Prepositions*, Speech and Language Processing. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Klenk, M.; Forbus, K.; Tomai, E.; Kim, H.; and Kyckelhahn, B. 2005. Solving everyday physical reasoning problems by analogy using sketches. In *Proceedings of the 20th national conference on Artificial intelligence (AAAI'05)*.

Kuipers, B. 2000. The spatial semantic hierarchy. *Artificial Intelligence* 119:191–233.

Langley, P. in press. Advances in cognitive systems. *AI Magazine*.

Lockwood, K.; Lovett, A.; and Forbus, K. 2008. Automatic classification of containment and support spatial relations in english and dutch. In *Proceedings of the international conference on Spatial Cognition VI: Learning, Reasoning, and Talking about Space*, 283–294. Springer-Verlag.

Regier, T., and Carlson, L. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General* 130(2):273–298.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M.; Banerjee, A.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI'11)*.

Thrun, S. 2003. Robotic mapping: A survey. In Lakemeyer, G., and Nebel, B., eds., *Exploring Artificial Intelligence in the New Millennium*. Morgan Kaufmann Publishers Inc. 1–35.

Web Mining Driven Semantic Scene Understanding and Object Localization

Kai Zhou, Karthik Mahesh Varadarajan, Michael Zillich, Markus Vincze

***Abstract*— Knowledge acquisition from the Internet for robotic applications has received widespread attention recently. It has turned out to be an important supplementary or even a complete replacement to conventional robotic perception. In this paper, we investigate state-of-the-art online knowledge acquisition systems for robotic vision applications and present a framework for further fusion and tighter integration. Bootstrapped by an interconnected process wherein modules for object detection and supporting structure detection co-operate to extract cross-correlated information, a web text mining technique using sequential pattern retrieval is introduced for linking the search of objects with their potential localities. Experiments using an indoor mobile robot for an Active Visual Search (AVS) task demonstrate the benefits of our coherent framework for visual representation and knowledge acquisition from the Internet.**

I. Introduction

In order to observe, detect, recognize, grasp or manipulate objects, diverse sensors have been mounted on versatile robots and various perception techniques have been designed for searching potential interest areas. As visual information is the most important sensory source for humans, visual perception algorithms play the most important role of all the robotic sensory knowledge acquisition methods and have received widespread attention in the last decades. Robotic researchers have applied numerous computer vision algorithms for detecting/recognizing potential objects in environments, and most recently they provide clear evidences of success in situating isolated object detector/recognizer in holistic scene understanding frameworks. These approaches [1][2][3][4][5] focus on the relationship between object information and environment, thereby facilitating more accurate detection/recognition of potential objects. However, the knowledge about the semantic link between the object of interest and its potential surrounding environment is still missing in current holistic scene understanding methods. This paper addressed this knowledge gap.

A practical instance of visual perceptual analysis in an indoor mobile robot scenario will be first described here to depict our intuition and development of robot visual perception system. Given a mobile robot with the task of searching a mug in the apartment, 1) The robot is driven around based on pre-defined or exploratory waypoints and **isolated** mug detector processes the image streams. However, abundant wrong and redundant detections are caused due to the presence of clutter (wrong detections), illusory or noisy contour (redundant detections) and degrade the robot's performance greatly. 2) Alternatively, the **holistic** scene understanding methods consider the potential spatial layout of the surroundings of a mug (e.g., a supporting plane) and coherently perceive the mug and the surrounding scene. This approach improves the efficiency and accuracy significantly (e.g., the isolated detector will process a poster with mug inside as candidate for region analysis, but the holistic method will not). Although the paradigm of holistic understanding of entire scenes improves the efficiency of existing robotic vision systems, considerable effort is still necessary to build robots that can perceive and interact with the environment in a fashion similar to that of humans. People focus visual attention on tables rather than on the floor given the task of locating a mug, and vice versa for locating a trash bin. This intelligence is based on an existing knowledge stored in our mind – in the normal case, the likelihood that a mug stands on the table is much higher than the possibility that mug is on the floor. We term this kind of knowledge as ”Common Sense about Object Locality (CSOL)”. 3) Web mining driven semantic scene understanding and object localization with **situated** CSOL is proposed in this paper for intelligent robot visual perception system. In the aforementioned example, using either surface web mining (e.g., a direct search from Google) or deep database mining (e.g., querying the online databases such as Open Mind Indoor Common Sense database (OMICS)), robots can be programmed to obtain the CSOL predicate that mugs are usually located on the top of tables or desks.

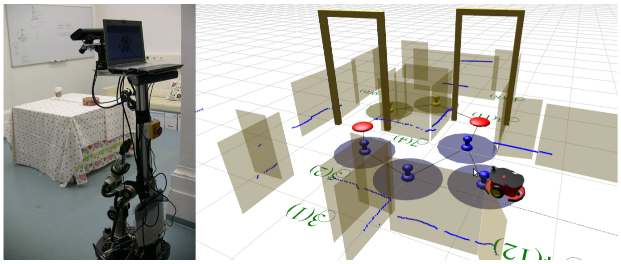


Fig. 1: Scenario and object search task at a glance, left: test scene with the robot at initial point, right: simulation/visualization of visual search task.

The paper is organized as follows. In §II we introduce the background and review state-of-the-art robot visual perception approaches. §III describes a holistic understanding approach using coherent stereo line detection and plane estimation for reasoning about the scene. We then detail how

The work was supported by EU FP7 Programme [FP7/2007-2013] under grant agreement No.215181, CogX.

All authors are with Automation and Control Institute, Vienna University of Technology, Gußhausstraße 27-29, A-1040, Vienna, Austria {zhou,kv,zillich,vincze}@acin.tuwien.ac.at

to generate CSOL predicates using web content mining in §IV. Subsequent sections present experimental results with synthetic scenes, and real robotic applications. A conclusion is given at the end of the paper and the future work is shortly discussed as well.

II. Related Work

The ultimate goal of robot visual perception is the generation of detailed 3D representations for salient objects to perform further robot manipulation. Researchers have developed many algorithms towards this goal; here we summarize the developments in three phases:

1) **Isolated** visual operators, such as specific object detector [6], sign recognizer [7] and preattentive feature based detector [8] are utilized to process the visual image captured from camera on the robot. However, isolated methods work on the entire search space thereby consuming excess computational power which is a scarce resource on a robot.

2) The **holistic** scene understanding techniques [1][2][3][4][5] consider visual operators and spatial layouts in a integrated manner for archiving accurate visual perceptive analysis of scene elements. However, these methods only use pure computer vision algorithms for robot perception and still work on a single robot agent without any prior knowledge or memories.

3) The **situated** perception methods allow the robot to make use of knowledge databases, short/long memories of the robot, learning beliefs and/or knowledge from networked robots, thereby obtaining more comprehensive information about the environment for perceiving the world it is situated in. A detailed overview of situated robotics can be found in [9] and of embodiment in [10], where it is argued to be crucial for a close coupling between brain, body and environment. Knowledge acquisition from the web or sharing databases have been adopted to supply a large corpus of training data [11] for visual recognition, to build 3D models for robot manipulation [12], to complete qualia structures describing an object [13], to guide robot planning for specific tasks such as table setting for a meal [14], and even more ambitiously to fill knowledge gaps when an indoor robot is executing sophisticated tasks [15]. However, to our knowledge, there is no robot vision system that obtains information extracted from the web for revealing the relationships of various objects and their most-likely locations.

Note that our robotic vision system as well as the entire robot platform are built atop the CoSy Architecture Schema (CAS) – a distributed asynchronous architecture [16], which facilitates integration of many relevant components that could bring additional functionality to the system in a coherent and systematic way.

III. Holistic Scene Understanding

A unified probabilistic framework, which combines stereo line detection with planar surface estimation is described in this section. Data association between planar surfaces and specific objects is addressed next. We also recommend readers [2][3] for the details.

The stereo line extraction is a bottom-up approach, First, edges are detected from image pairs with an adaptive canny edge detector before we fit lines into the extracted edge chains using the method of Rosin and West [17]. Then we match the lines of the stereo image pair using the mean-standard deviation line descriptor (MSLD) [18] together with the constraint of epipolar lines is utilized in the calibrated stereo camera setup. A confidence value $Con(f)$ for stereo matched line is then calculated based on the angle between the stereo match and the epipolar line. Note that the resulting value $Con(f)$, although in the range of $[0, 1]$, is not a probability. Rather, this value denotes the quality and correctness of the reconstructed lines.

We adopt CC-RANSAC [19] as the underlying plane estimator and assign confidence values $Con(S)$ to the estimated planes by calculating the average normal vector of connected points. This confidence value is used for the joint probability maximization and will be addressed in detail in §**??**. It is reported in [2][3] that plane refinement within a unified probabilistic framework facilitates more reliable estimation than using CC-RANSAC only.

Again the confidence $Con(S)$ does not explicitly represent a probability. However, we can use these confidence values to approximate a probability distribution by generating samples around the estimated plane and weighting these samples with confidences. Given the plane S returned by CC-RANSAC, and $\tilde{S}$ a generated sample near S, we formulate the probability distribution in the following way,

$$\begin{aligned} p(\tilde{S}|Con(\tilde{S})) &= \frac{p(Con(\tilde{S})|\tilde{S})p(\tilde{S})}{p(Con(\tilde{S}))} \\ &= \frac{[(Con(\tilde{S}) > t)]p(\tilde{S})}{p(Con(\tilde{S}))} \end{aligned} \tag{1}$$

Here t is a threshold and $[\,]$ denotes the Iverson bracket:

$$[X] = \begin{cases} 1, & \text{if } X \text{ is TRUE} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

With the Iverson bracket, the probability $p(\tilde{S}|Con(\tilde{S}))$ is proportional to the prior for the sampled plane $\tilde{S}$ whenever $Con(\tilde{S}) > t$, and 0 elsewhere. In other words, $p(Con(\tilde{S})|\tilde{S})$ facilitates thresholding of plane samples with low confidence. We draw samples randomly from the neighboring area of S to generate $\tilde{S}$, and $\tilde{S} \sim \mathcal{N}(\mu_n, \sigma_n)\mathcal{N}(\mu_h, \sigma_h)$, where n and h are the normal vector of plane S, and the distance of plane S to the origin. Hence, $p(\tilde{S})$ is a Gaussian distribution and assigns higher probabilities to the samples near to the estimated plane.

The joint probabilistic model consists of three parts, (1) the probability that the estimated plane is at $\tilde{S}$, (2) the likelihood of positive stereo line detection with the underlying plane estimation, (3) the confidence value of detected lines returned by the stereo line detection algorithm, and can be written as

$$p(S, W, E) \propto \prod_{i=1}^{K} p(\tilde{S}_i|Con(\tilde{S}_i)) \prod_{j=1}^{M} p(t_j|f_j, S)p(f_j, t_j|e_j) \tag{3}$$

The first and last probabilities are given using Eq. 1 and stereo match confidence respectively. The second probability is determined by the distance and angle between detected stereo lines and planes.

To maximize the joint probability, we present the optimization problem as $\arg\max_{s_i,t_j}(\ln p(S,W,E))$, the logarithmic formulation can be rewritten as,

$$\begin{aligned}\ln p(S,W,E) = &\sum_{i=1}^{K} \ln p(S_i|Con(S_i)) \\ &+ \sum_{j=1}^{M} [\ln p(t_j|f_j,S) + \ln p(f_j,t_j|e_j)]\end{aligned} \tag{4}$$

where S_i, t_j are the parameters to be estimated. We select the plane which has the highest confidence value of all the plane estimation results, and only consider this plane as the scene geometry for the joint probabilistic model optimization. Then the first part of Eq. 4 is a constant and the second part can be calculated independently through M 3D matched lines comparisons of $\ln p(t_j = 0|f_j, S) + \ln p(f_j, t_j = 0|e_j)$ with $\ln p(t_j = 1|f_j, S) + \ln p(f_j, t_j = 1|e_j)$. After labeling all the stereo lines, the pose of the plane with the highest confidence is refined by searching the nearby planes $\tilde{S}$. This refined pose should satisfy the criterion of maximizing the number of stereo lines parallel or orthogonal to it.

Again, we refer the authors to the previous publication [2][3] for the deduction of aforementioned formulae. A noteworthy remark of this joint probabilistic approach is that it considers all the relative elements (planes, stereo lines as objects) of the current scene in a integrated manner to obtain the optimized scene understanding, but it doesn't know whether the objects and planes in the current scene should be linked properly or not under the situated consideration. Obviously, if visual perceptive analysis is implemented only when the proper link of objects and supporting surface is detected, the object search task in the large scale environment can be executed more accurately and efficiently. The solution to break the improper link or vice versa to reveal the most appropriate link between the given objects and detected supporting planes, will be addressed in the next section.

IV. Locality Discovery With Web Mining

Locality of objects plays an important role in robotic top-down perception processes, such as active visual search. The spatial concepts reflected by the locality of objects are of great importance to robots, especially mobile ones [2][3][4].

As mentioned earlier, knowledge acquisition from the web for robots has received widespread attention in the last years [11][12][13][14][15], given that the World Wide Web is a huge, dynamic, diverse and interactive medium to gain open and free information. While these papers focus on obtaining various knowledge, they do not cater to obtaining semantic positional saliency from the Internet, which forms the core of this paper. We make use of text mining from web to generate Common Sense about Object Locality (CSOL) for efficient guiding of robot visual search.

A. Noun Of Locality: ON

The functional interpretations of the spatial language term ”on” not only act as an indicator for cognitively plausible and practical abstractions of localization knowledge in the field of mobile robotics, but have also received widespread research attention from psychology, neurobiology and linguistics. The use of web content mining technology to extract CSOL enables the exploration of large resources of information to improve efficiency of robot visual search.

1) The term ”on” is the functional abstraction of mechanical support, which is strongly relevant to the planar supporting surfaces – a dominative structure in artificial indoor environments.

2) The spatial concept implied in the noun of locality ”on”, which allows humans to analyse, generalize and internalize spatial experiences, plays a prominent role in human cognition.

3) When verbally representing scenes with mechanical support, contact or suspension, ”on” is also a keyword which can demonstrate and derive other related vocabulary. Hence researchers in the field of Natural Language Processing (NLP) have developed several algorithms around the study of the spatial language term ”on”.

4) As the 14th most common English word, ”on” serves as an exemplar of knowledge discovery or information retrieval from diverse resources. This diversification ensures the stability of the web mining results.

The spatial language term ”on” thus serves as an efficient text mining pattern for semantic knowledge representation and hence is used in this paper for discovery of CSOL for visual perception in indoor mobile robotics.

B. Basic Definition

As a fertile area for data mining, the Wide World Web has been viewed as the biggest information resource today, while the huge amount of available information also raises issues of scalability, transiency, diversification and redundancy. Web content mining, as one of the most important research directions in web mining, has reached considerable maturity in recent years (see [20][21] for good overviews). Among all web content mining techniques, Pattern Taxonomy Mining (PTM) remains a popular technique. Though inefficient in the context of information extraction from web documents [22], its specific characterizations – indirect phrase representation and absolute definitions fit perfectly to our requirements.

The definition of sequential pattern used in the paper is described as follows. Let $T = \langle t_1, t_2, t_3 \ldots, t_n \rangle$ be the representation of a sequential text pattern. The semantic representation (both singular and plural) of the object O is obtained for both user-driven mode (i.e., the user requests the robot for something) and non-situated inference mode, e.g., in [14], wherein the robot learns how to set the table for a meal through retrieval of web information, in the form of annotations of objects required. The first term of the sequential pattern, t_1 will be set to the collection of O, i.e., $t_1 = \{O_1, O_2, \ldots, O_k\}$, where k is the number of queried objects. The second term t_2 is the lemma ”be” which includes occurrences of ”was”, ”is”, ”were” and ”are”. The third term t_3 is a set of nouns of locality, including ”on”. The last term in the pattern $t_n = \{S_1, S_2, \ldots, S_h\}$ is a collection of potential supporting surfaces S in the robot exploration environment. The information of these surfaces can be provided by user predefined contexts or furniture detection algorithms.

Definition IV.1. *(Sub- and Super-sequence) Given two sequences* $\alpha = \langle a_1, a_2, \ldots, a_m \rangle$*,* $\beta = \langle b_1, b_2, \ldots, b_\ell \rangle$*, we define* α *is a sub-sequence of* β *if and only if there exist integers* $1 \leq i_1 < i_2 < \ldots < i_m < \ell$*, such that* $a_1 = b_{i1}, a_2 = b_{i2}, \ldots, a_m = b_{im}$*.*

For instance, sequence $I = \langle t_1, t_3, t_{n-1} \rangle$ is a sub-sequence of $T = \langle t_1, t_2, t_3 \ldots, t_n \rangle$. Furthermore, if sequence G is a sub-sequence of T, we call $T = \langle t_1, t_2, t_3 \ldots, t_n \rangle$ a super-sequence of G.

Definition IV.2. *(Absolute and Relative Support) Given a database* $\mathcal{D}$ *(can either be the World Wide Web or a specific robotics knowledge database, e.g., OMICS) and a sequential pattern* $\mathcal{T}$*, the* ***absolute support*** *of* $\mathcal{T}$ *in* $\mathcal{D}$*, denoted as* $supp_a(\mathcal{T}; \mathcal{D}) = ||\{\mathcal{T} | \mathcal{T} \in \mathcal{D}\}||$*, is the number of occurrences of* $\mathcal{T}$ *in* $\mathcal{D}$*. The* ***relative support*** *of* $\mathcal{T}$ *is the fraction of sentences that contain* $\mathcal{T}$ *in the entire database* $\mathcal{D}$*, denoted as* $supp_r(\mathcal{T}; \mathcal{D}) = supp_a(\mathcal{T}; \mathcal{D})/||\mathcal{D}||$*. The* ***support collection*** *is defined as a set of paragraphs, and each of the paragraphs contains the same sequential pattern* $\mathcal{T}$*, i.e.,* $\{supp(\mathcal{T}; \mathcal{D})\} = \{\mathcal{T} | \mathcal{T} \in \mathcal{D}\}$*.*

Definition IV.3. *(Frequent Sequential Pattern) A sequential pattern* $\mathcal{T}$ *is considered as a* ***frequent sequential pattern (fsp)*** *if and only if* $supp_a(\mathcal{T}; \mathcal{D}) \geq \zeta$*, where* ζ *is the minimum support (min_sup) threshold.*

The reason for using *min_sup* in our approach is to evaluate and qualify the support collections discovered in specific-scaled databases (e.g., professional robotic knowledge database), thereby enabling the selection of support collections with higher relative support for further processing, while objects with lower relative support trigger the robot to change its mining database to a lager one (e.g., Internet). Since the size of the professional robotic database is far smaller than the size of generic on-line database, this piecewise process is capable of decreasing the system burden and/or time for cognitive processing or reflection for the robot. The utilization of generic on-line database is also inevitable because the professional database delivers higher performance only in a limited scope (You may not get reasonable number of retrieval items when searching uncommon objects in professional database).

Definition IV.4. *(Object pattern, Locality pattern and Full pattern) A* ***object pattern*** $\mathcal{T}^o$ *is composed of in-sequence object representations O, lemma ”be” and a noun of locality, a* ***full pattern*** $\mathcal{T}^f$ *consists of a object pattern, a potential supporting surface at the end, and an arbitrary number of terms between. A* ***locality pattern*** $\mathcal{T}^l$ *is the full pattern without the first object term.*

Both object pattern $\mathcal{T}^o$ and locality pattern $\mathcal{T}^l$ are sub-patterns of full pattern $\mathcal{T}^f$, and full pattern $\mathcal{T}^f$ is the super-pattern of object pattern $\mathcal{T}^o$ and locality pattern $\mathcal{T}^l$.

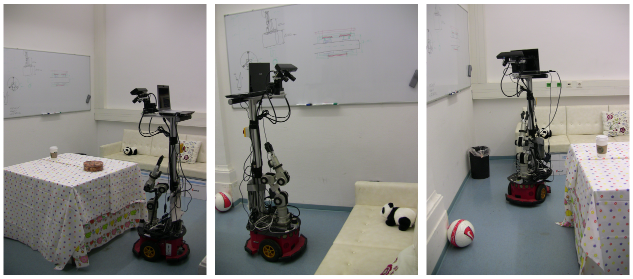
C. Pattern Retrieval

Based on the pattern representation of text documents, we present a new two-stage pattern retrieval approach for discovering locality knowledge CSOL. As we demonstrate in Algorithm 1, using pattern retrieval for robotic visual search is designed as a closely integrated two stage mining process. The mining databases are set to the specific robotic knowledge library (e.g., OMICS) or a more generic large-scale information source (e.g., Internet). The pattern retrieval algorithm operating on the specific robotic database that is of a reasonable size, can satisfy the timeliness of active visual search task while providing reasonable results for retrieving items of daily use. However, most of the robotic knowledge libraries (e.g., OMICS) are incomplete and updated periodically, and the retrieved results to queries are limited in scope. The generic large-scale information source (e.g., Internet) can be considered as an important supplementary source when the retrieval of the robotic database fails. Utilization of it increases the system burden and time consumption, not only because of the database size changing but also caused by the pruning as a preprocessing step to filter out the unrelated items. However, the robust retrieval results can facilitate more effective visual search.

Algorithm 1 Pattern retrieval of visual object search

```
1:  Set operating database $\mathcal{D}$ to robotic database $\mathcal{D}_r$
2:  if $\exists$ fsp $T^o$ , $i.e.\ supp_a(T^o; D) \geq \zeta$ then
3:     Calculate support collection $C = \{supp(T^o; \mathcal{D})\}$
4:     for $t_n = S_1 \to S_h$ do
5:        Compose $T_i^l$ with $t_n = S_i$ as the last term
6:        Compute relative support $supp_r(T_i^l; \mathcal{C})$ w.r.t. $C$
7:        Sort $\{supp_r(T_i^l; \mathcal{C}) | i = 1, \ldots, h\}$
8:     end for
9:  else
10:    if $\mathcal{D} = \mathcal{D}_r$ then
11:       Set $\mathcal{D}$ to the generic Internet database $\mathcal{D}_I$, back to
          line 2
12:    else
13:       Return failure
14:    end if
15: end if
16: Return the sorted results
```

The relative supports with respect to various elements in the support collection are sorted, thereby providing a priority table for linking the first term in sequential pattern (object) with the last term in the pattern (locality). We compute the relative support of T^l in the support collection $C = \{supp(T^o; \mathcal{D})\}$ for normalization of relative supports of various objects, since there might be a significant difference in the number of retrieved items between commonly found and uncommon objects.



(a) table scene (b) sofa scene (c) floor scene

Fig. 2: The indoor robot test scenario setting, from left to right, the robot is looking towards the table, sofa and floor for visual perception.

In our experiments, the minimum support (*min_sup*) threshold is set to 20 for the OMICS empirically, although this is a relative small number with regard to the 1184144 statements[1] in OMICS. Furthermore, *min_sup* is set to 1000 for the Internet data and we will show that this setting produces robust retrieval results in the next section.

V. Experiments

The evaluation of pattern retrieval is performed by demonstrating the validity of linkage between several common objects with their most likely locations. An indoor robot that applies this knowledge discovery methods is tested in a structured environment (Fig. 1 and Fig. 2) to depict how the online knowledge discovery facilitates effective and accurate active visual search.

A. Evaluation of Pattern Retrieval

To assess the quality of our pattern retrieval approach, several objects are used as the target term in the pattern and the two different databases (the specific database through OMICS and the large-scale generic one through Google advanced search) are applied as data mining sources. Fig. 3 displays the text mining results for three common objects in OMICS. Note that the noun of locality used for mining may be tailed by contextually unrelated nouns - not just places which do not exist in the current room/apartment context, but also some phrases or idioms. For instance, we notice that the object term ”book” has a relative high likelihood 57% for other ”location” misnomers in comparison with the locations the robot could possibly find in a room, such as table, shelf and floor. However, since most misnomers are widely used phrases, such as ”on sale”, these can be easily pruned away.

When there are not enough ($> \zeta$) retrieval results from OMICS, we use Google advanced search to retrieve results from the Internet. Fig 4 shows three pattern retrieval results. In this figure, we find that the object ”cushion” and ”trash can” are tightly bound with the locations ”sofa” and ”floor” respectively. The retrieval result of pattern $\{''football''+''be''+''on''\ldots+''location''\}$ returns two dominant locations which have similar probabilities of occurence. Although the location ”table” is dominantly picked up, the actual meaning of this word refers to ”diagram with columns of information” in the context of ”football” rather than what we need for robotic task – ”furniture upon which to work, eat”.

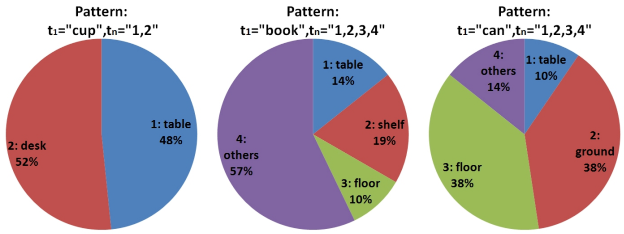


Fig. 3: The pattern retrieval results of three common objects – ”cup”, ”book” and ”can”, the source being the indoor-robot knowledge database OMICS, - only localities that exist in an office room are shown in the figure. Patterns containing ”cup”, ”book” and ”can” have absolute support values 31, 21 and 21 respectively.

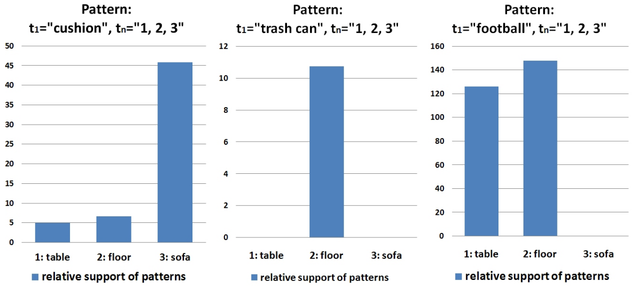


Fig. 4: The pattern retrieval results of three objects – ”cushion”, ”trash can” and ”football”, the source is the general Internet data accessed with Google advanced search, and only the localities that exist in an office room are searched for - these are displayed in the figure. Note that here we use the bar figure instead of pie figure, because comparing with the localities that are not depicted here (“others” part in Fig. 3), the number of displayed items are significantly smaller.

B. Robot Active Visual Search

We test our web content mining approach within a real indoor robotic scenario. The robot explores a room with a table in the center and a sofa next to the wall. Several objects (listed in Fig. 5) are placed on the table, floor or couch. The autonomous navigation of the robot is implemented as [23]. The visual search strategy is straightforward – at every spot, the robot will pan (±90°) and tilt (−60°) the camera to perform visual perceptive analysis. In contrast, the pattern retrieval based web content mining will prune the search when the dominant plane in the current scene does not match the object’s most likely location. Fig. 6 depicts the way points of the robot and also shows the relative positions of furnitures in the room. The greater efficiency of applying this approach for the task of object visual search is apparent in Fig. 5.

[1] According to the database statistic of OMICS project at http://openmind.hri-us.com.

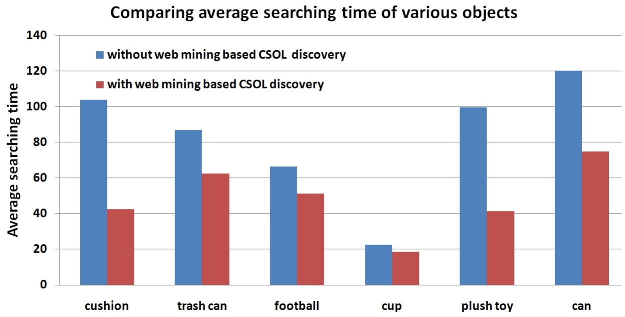


Fig. 5: Comparison of average visual search time for brute force search and the web content mining method proposed in this paper. The visual search of each object is repeated 10 times and the average processing time is recorded.

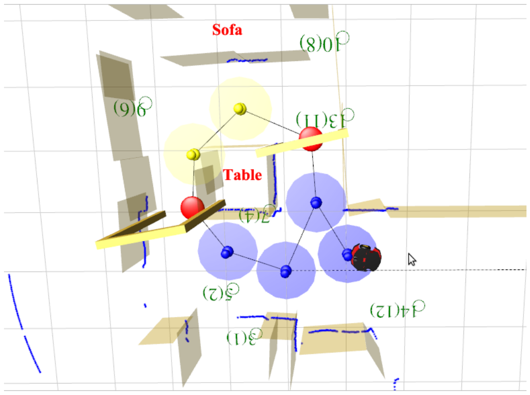


Fig. 6: Simulation/Visualization world from the top view.

VI. Conclusion and Future Work

In this paper, we present a robotic vision system which is based on the fusion of holistic visual perception and web content mining. We generate spatial information in the scene by considering plane estimation and stereo line detection coherently within a unified probabilistic framework, and show how the resulting scene information can be efficiently searched using pattern based data mining from web. Experiments demonstrate that our system can sort possible spatial locations according to their relationships with various objects, thereby providing an effective and plausible robotic visual search strategy.

Two main dimensions of using web content mining for discovering CSOL knowledge form the focus of our future work. Firstly, the assumption that the sentence containing the object and its most likely existing location has the dominant role in the online database, although intuitively correct, requires further investigation. Secondly, the selection of the objective term influences significantly the quality of retrieval results. The application of objects' synonyms or surface variants can help solve this problem.

References

[1] K. Zhou, A. Richtsfeld, M. Zillich, and M. Vincze, "Coherent spatial abstraction and stereo line detection for robotic visual attention," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011.

[2] K. Zhou, A. Richtsfeld, M. Zillich, M. Vincze, A. Vrečko, and D. Skočaj, "Visual information abstraction for interactive robot learning," in *The 15th International Conference on Advanced Robotics (ICAR 2011)*, Tallinn, Estonia, June 2011.

[3] K. Zhou, A. Richtsfeld, K. M. Varadarajan, M. Zillich, and M. Vincze, "Combining plane estimation with shape detection for holistic scene understanding," in *Advanced Concepts for Intelligent Vision Systems 2011 (ACIVS2011)*, Het Pand, Ghent, Belgium, Aug 2011.

[4] K. Sjöö, A. Aydemir, T. Mörwald, K. Zhou, and P. Jensfelt, "Mechanical support as a spatial abstraction for mobile robots," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2010.

[5] A. Vrečko, D. Skočaj, N. Hawes, and A. Leonardis, "A computer vision integration model for a multi-modal cognitive system," in *The 2009 IEEE/RSJ International Conference on Intelligent RObots and Systems*, October 2009, pp. 3140–3147.

[6] A. Nüchter, H. Surmann, and J. Hertzberg, "Automatic classification of objects in 3d laser range scans," in *Proc. 8th Conf. on Intelligent Autonomous Systems*, 2004, pp. 963–970.

[7] T. Xu, N. Chenkov, K. Kühnlenz, and M. Buss, "Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots," in *Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems*, 2009, pp. 4009–4014.

[8] C.-K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation & localization using gist and saliency," in *Intelligent Robots and Systems, 2010. IROS 2010. IEEE/RSJ International Conference on*, Oct 2010.

[9] M. J. Mataric, "Situated robotics," in *Encyclopedia of Cognitive Science*. Nature Publishing Group, Macmillan Reference Ltd, 2002.

[10] R. Pfeifer, M. Lungarella, and F. Iida, "Self-organization, embodiment, and biologically inspired robotics," *Science*, vol. 318, pp. 1088–1093, 2007.

[11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*, vol. 2, Oct. 2005, pp. 1816–1823.

[12] U. Klank, M. Z. Zia, and M. Beetz, "3d model selection from an internet database for robotic vision," in *IEEE International Conference on Robotics and Automation*, May 2009, pp. 2406 –2411.

[13] P. Cimiano and J. Wenderoth, "Automatically learning qualia structures from the web," in *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, ser. DeepLA '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 28–37.

[14] D. Pangercic, R. Tavcar, M. Tenorth, and M. Beetz, "Visual scene detection and interpretation using encyclopedic knowledge and formal description logic," in *Proceedings of the International Conference on Advanced Robotics (ICAR).*, Munich, Germany, June 22 - 26 2009.

[15] M. Waibel, M. Beetz, R. D'Andrea, R. Janssen, M. Tenorth, J. Civera, J. Elfring, D. Gálvez-López, K. Häussermann, J. Montiel, A. Perzylo, B. Schießle, O. Zweigle, and R. van de Molengraft, "RoboEarth - A World Wide Web for Robots," *Robotics & Automation Magazine*, vol. 18, no. 2, 2011.

[16] N. Hawes and J. Wyatt, "Engineering intelligent information-processing systems with CAST," *Adv. Eng. Inform.*, vol. 24, no. 1, pp. 27–39, 2010.

[17] P. Rosin and G. West, "Nonparametric segmentation of curves into various representations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 12, pp. 1140 –1153, Dec. 1995.

[18] Z. Wang, F. Wu, and Z. Hu, "Msld: A robust descriptor for line matching." *Pattern Recognition*, vol. Vol. 42, pp. 941–953, 2009.

[19] O. Gallo, R. Manduchi, and A. Rafii, "CC-RANSAC: Fitting planes in the presence of multiple surfaces in range data," *Pattern Recogn. Lett.*, vol. 32, pp. 403–410, February 2011.

[20] R. Kosala and H. Blockeel, "Web mining research: a survey," *Sigkdd Explorations*, vol. 2, pp. 1–15, 2000.

[21] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," *Knowl. Eng. Rev.*, vol. 21, pp. 1–24, March 2006.

[22] Y. Li, S.-T. Wu, and X. Tao, "Effective pattern taxonomy mining in text documents," in *CIKM*, 2008, pp. 1509–1510.

[23] N. Hawes, M. Hanheide, K. Sjöö, A. Aydemir, P. Jensfelt, M. Göbelbecker, M. Brenner, H. Zender, P. Lison, I. Kruijff-Korbayov, G.-J. M. Kruijff, and M. Zillich, "Dora the explorer: A motivated robot," in *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, May 2010.

Web Mining Driven Object Locality Knowledge Acquisition for Efficient Robot Behavior

Kai Zhou, Michael Zillich, Hendrik Zender and Markus Vincze

***Abstract*— As an important information resource, visual perception has been widely employed for various indoor mobile robots. The common-sense knowledge about object locality (CSOL), e.g. a cup is usually located on the table top rather than on the floor and vice versa for a trash bin, is a very helpful context information for a robotic visual search task. In this paper, we propose an online knowledge acquisition mechanism for discovering CSOL, thereby facilitating a more efficient and robust robotic visual search. The proposed mechanism is able to create conceptual knowledge with the information acquired from the largest and the most diverse medium – the Internet. Experiments using an indoor mobile robot demonstrate the efficiency of our approach as well as reliability of goal-directed robot behaviour.**

I. INTRODUCTION

To perform object search tasks efficiently and reliably, common-sense conceptual knowledge about the structure of the world has been introduced to guide planning for the robot [1][2][3][4][5][6][7][8]. This common-sense conceptual knowledge, which describes the relational structures between objects and their surrounding environment, probabilistically represents the confidence value of the statement "object $\mathcal{O}$ is on/in location $\mathcal{L}$". This probabilistic representation is capable of modelling the uncertainty in robotic perception, thus enhancing the plausibility and reliability of the robot's behaviour [1]. Although using common-sense conceptual knowledge about the relations between object and environment to benefit robotic visual search dates back to 1970's [9], recently it becomes popular to obtain this knowledge by automatically analyzing large-scale knowledge repositories rather than inputting manually [1][8]. The limitation with respect to the scale of professional information resources and the lack of robust knowledge extraction approaches are the main obstacles for applying the online knowledge acquisition in robotics. Certainly the trade-offs between the size/professionalisation of the information resources as well as the efficiency/reliability of the knowledge extraction approaches also affect the progresses made in this field. Thus, though automatic information acquisition by downloading the repository of knowledge has been a dream of the AI community for several decades and has appeared in many fictional movies, the robotic community is still working towards obtaining common-sense conceptual knowledge automatically.

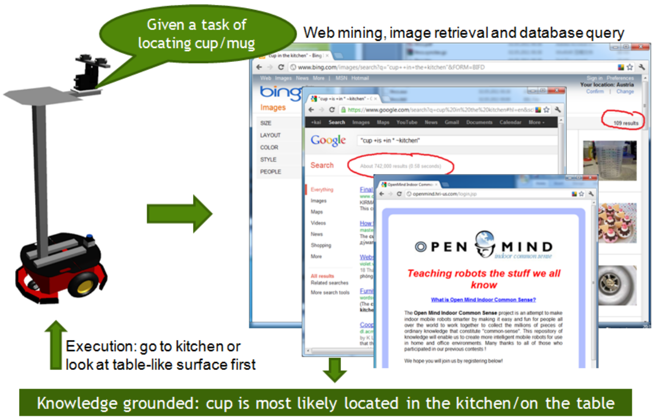

Given a task of locating cup/mug
Web mining, image retrieval and database query
bing
Google
OPEN MIND
Teaching robots the stuff we all know
Execution: go to kitchen or look at table-like surface first
Knowledge grounded: cup is most likely located in the kitchen/on the table

Fig. 1: Example scenario and object search task at a glance, note that the web search of text/image displayed here is only used to show the process, the embedded online knowledge acquisition method will be described in section IV.

The broad availability and open accessibility of the corpora on the World Wide Web (WWW) provide robots with opportunities for novel knowledge acquisition techniques and strategies. Using the WWW as the information resource for robotic applications has received widespread attentions in recent years. Knowledge acquisition from the web or sharing databases have been adopted to supply a large corpus of training data [10] for visual recognition, to build 3D models for robot manipulation [11], to complete qualia structures describing an object [12], to guide robot planning for specific tasks such as table setting for a meal [13], and even more ambitiously to fill knowledge gaps when an indoor robot is executing sophisticated tasks [14]. However, for mobile robot research, discovering common-sense conceptual knowledge about the relations of object and environment from the web is still in an early stage [3][4][1][2][8], and many progresses and improvements could be made in terms of efficiency and robustness. This paper will address this cutting-edge field in mobile robotic research.

The main contributions of this paper are: 1) *Accurate probabilistic conceptual knowledge* that represents the relations of objects and their situated environments, is extracted by fusing search engine query data and a professional database. 2) *For the first time provide a large body of experimental results* of probabilistic knowledge (hundreds of objects), to demonstrate the validity of the idea that object locality

The research leading to these results was supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No.215181, CogX.

Kai Zhou, Michael Zillich and Markus Vincze are with Automation and Control Institute (ACIN), Vienna University of Technology, Gußhausstraße 27-29, A-1040, Vienna, Austria. {zhou,zillich,vincze}@acin.tuwien.ac.at

Hendrik Zender is with Language Technology Lab, German Research Center for Artificial Intelligence (DFKI), Campus D3.2, Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany. zender@dfki.de

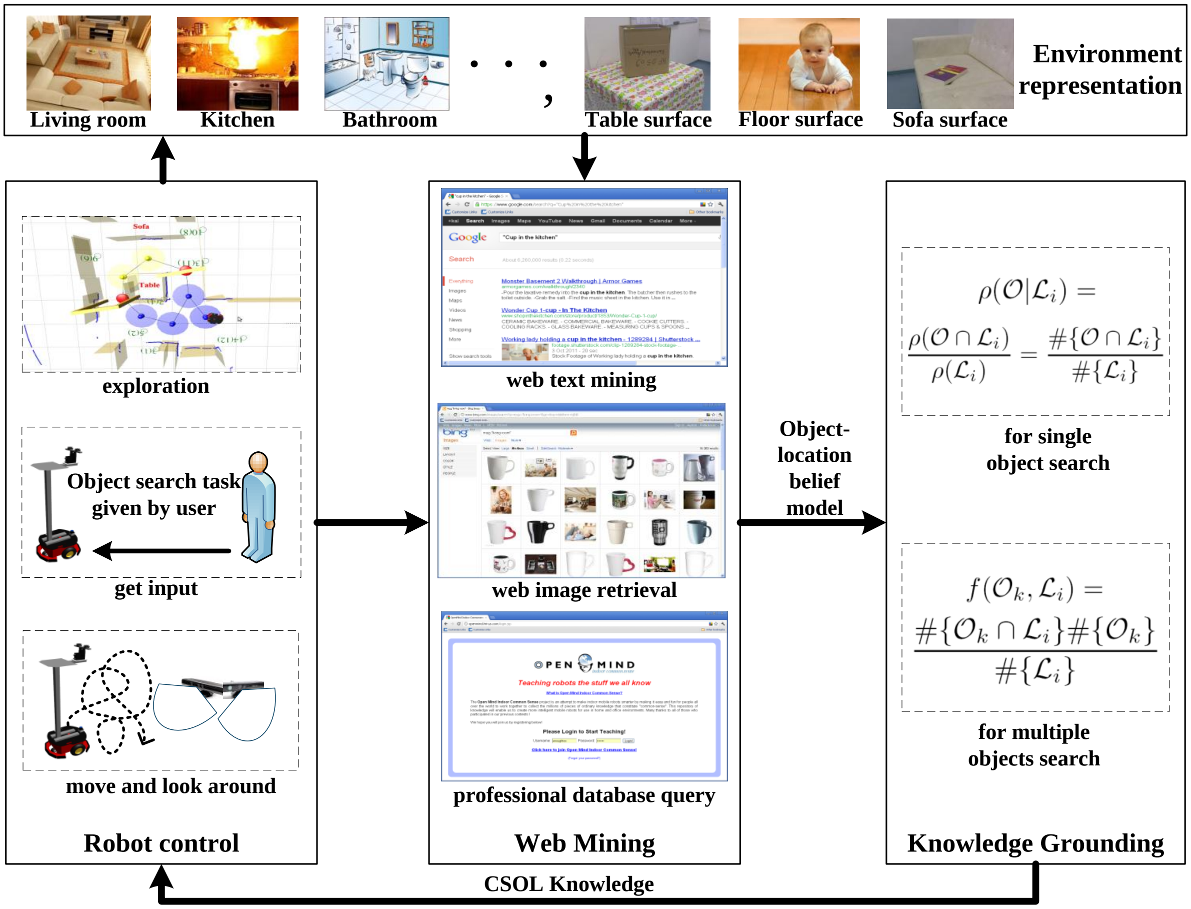

Living room
Kitchen
Bathroom
Table surface
Floor surface
Sofa surface
Environment representation
Sofa
Table
exploration
Object search task given by user
get input
move and look around
Robot control
web text mining
web image retrieval
professional database query
Web Mining
Object-location belief model
$\rho(\mathcal{O}|\mathcal{L}_i) = \frac{\rho(\mathcal{O}\cap\mathcal{L}_i)}{\rho(\mathcal{L}_i)} = \frac{\#\{\mathcal{O}\cap\mathcal{L}_i\}}{\#\{\mathcal{L}_i\}}$
for single object search
$f(\mathcal{O}_k,\mathcal{L}_i) = \frac{\#\{\mathcal{O}_k\cap\mathcal{L}_i\}\#\{\mathcal{O}_k\}}{\#\{\mathcal{L}_i\}}$
for multiple objects search
Knowledge Grounding
CSOL Knowledge

Fig. 2: The overall data flow of using the CSOL knowledge that derives from web mining results to perform robotic visual search task.

knowledge can be discovered through the analysis of Internet queries and shared database.

The remainder of this paper starts with the introduction of the related work of robotic visual search and reviews the state-of-the-art robotic applications using information acquisition from the web (Section II). Then we detail the preliminary definitions of mathematical theories in Section III. Section IV describes the online knowledge extraction approach which combines web text mining, image retrieval and database query, as well as how this approach is utilized to extract CSOL from the Internet. Subsequent sections explain the test scenarios with various experimental setups, evaluations and analyses of results. A conclusion is presented at the end and the future work is also shortly discussed.

II. Related Work

In this section, we first give an overview of the robotic visual search task, then we will briefly describe the common-sense object locality (CSOL) knowledge, and introduce recent studies about how this knowledge is applied for visual search tasks performed by various indoor mobile robots.

For indoor mobile robots the intelligence for performing complex tasks in real environments is an interconnected process wherein low-level raw data obtained from various sensors and high-level knowledge need to co-operate in order to extract cross-correlated information. Active visual search, which is a typical task required to be performed by various robots, is a popular study case which incorporates low-level data from bottom-up visual attention and high-level semantic information from users' expectations/knowledge repository. The pioneer work of robotic active search in [15] has shown that the task of optimizing the sequential locations for observing objects, given a probability distribution, is an NP-hard problem. However, much research on improving the robustness and efficiency of the approximations and simplifications of this problem has been launched in the recent years [16][17][3][4][1][8].

Recent research demonstrates that the common-sense object locality (CSOL) knowledge has played an important role in mobile robots' visual search tasks [1][2][8][3][4][18]. In [3], the CSOL knowledge is termed as spatial relations which are represented probabilistically to cast the object search problem as a fully-observable Markov decision process (MDP). [18] integrates an attentive process into the visual object search planning of a mobile robot, i.e., optimizing the probability of finding the object target using information generated from the analysis of visual attention over time. Galindo et al. solve the task planning problem of mobile robot using a semantic map [19] or the AH-graph model based abstraction of the world [20]. Both their semantic maps and world abstraction contain numerous object locality knowledge. However, although the aforementioned literature have applied CSOL knowledge to facilitate more efficient robot behavior, all of them use the conventional way to generate the CSOL knowledge, which is manual-input, predefined and restricted to searching a single object.

Rapid development of World Wide Web techniques provides researchers with opportunities for obtaining huge, dynamic, diverse and interactive information. The robotic community also noticed this trend, and various robotic tasks have benefited by using knowledge acquisition from the web or sharing databases [10][11][12][13][14][21]. [3] presented an efficient MDP-based active visual search (AVS) procedure exploiting object relational knowledge such as "book ON table1 IN room1". They could show that AVS informed by such knowledge provided in probabilistic form could significantly improve search times and success rates, and that indeed the accuracy of the provided knowledge plays an important role. Probabilities in their system however were still hand-coded. Extending this work [1] and [2] generate robot common-sense knowledge by querying cooccurrence of objects and locations from an image search engine and a robotic database, then integrate the obtained probabilistic relation into a switching planner (continual fast downward plus decision theoretic planner) for efficient robot behaviour. They demonstrate the effectiveness of the approach in searching for a given object among 19 known objects, and the applicability of probabilistic knowledge obtained form web resources. Obtaining that knowledge however still required user intervention to collect the numbers of search hits for calculating cooccurrence probabilities. Follow-up work in [4] showed extended results, searching for different types of objects. Following the above work, [8] proposed a web text mining driven CSOL knowledge extraction and combined with their robotic holistic scene understanding visual system for performing object search tasks. Their selection method of the objective term, which influences significantly the quality of retrieval results, also requires to be elaborated in advance thus limiting the degree of flexibility and expandability.

Note that our mobile robot system shares the same underlying architecture (the CoSy Architecture Schema (CAS) – a distributed asynchronous architecture [22]) with [1][2][3][4][8], thereby maintaining the functionality of the previous system and meanwhile providing increased performance of the object search task (particular multi-object search) through applying the online CSOL knowledge acquisition mechanism.

III. Preliminary Definitions

The representation and generation of knowledge for robotics is highly related to several mathematical theories, which will be firstly discussed in this section.

A. Mathematical Logic

Mathematical logic is the general approach to representing and reasoning knowledge for robotics due to its significantly important role in artificial intelligence (AI) research [23]. The conventional and state-of-the-art mechanisms use Description Logics (DL) to describe and reason about the robotic knowledge ontologically [24]. Description Logics, which consist of a family of formal knowledge representation languages, are of significant importance in providing the ontological representation of knowledge. It integrates the expressive way of Propositional Logic (PL) and efficient decision of First-order Logic (FoL). We use a practical robotic knowledge example to introduce the development of applying these mathematical logics in robotics.

PL interprets the true or false statements formally with formulas. For instance, the typical spatial knowledge in robotics – "Red cup is on the table", which is a true proposition, can be interpreted as *OnTable(Redcup)*, where *OnTable()* denotes the propositional function and *Redcup* is a variable parameter. While propositional logic covers simple declarative propositions, first-order logic additionally extends with predicates and quantification, i.e. "All the cups are on the table" is interpreted by FoL as *OnTable(X)*, $X = \{Allcups\}$, where curly brackets $\{\}$ delimit the set of variable collections. However, once the information resources involve uncertainty quantification or the reasoning process yields uncertain results, DL with the integration of PL and FoL cannot provide solutions within reasonable calculational effort to enable uncertainty-savvy logical reasoning. This is also the reason that the prior attempts of applying CSOL knowledge cannot create a holistic approach to the robotic search task. For instance, the CSOL knowledge "The possibility of locating a cup on the table is 65%" and "The possibility of locating a cup on the floor is 35%" cannot satisfy the quantification condition of DL. Thus previous literature [1][2][3][4][8] handle these information externally by taking the higher potential of object location as the dominant/unary one for the further object search task. However, this external operation works only because in both their test scenarios single-object searches in a known environment (CSOL knowledge about single object at particular locations is calculated off-line) are performed.

B. Pattern Retrieval for Text Mining

Following the definition of CSOL knowledge in [8], the structure of pattern used for web text mining is also represented using the Pattern Taxonomy Model (PTM) in this paper. An **object pattern** $\mathcal{T}^o$ is composed of in-sequence object representations O, lemma "be" and a noun of locality (NoL). A **locality pattern** $\mathcal{T}^l$ is composed of in-sequence lemma "be", a noun of locality and locality representation L. A **full pattern** $\mathcal{T}^f$ consists of an object pattern, a potential supporting surface at the end, and an arbitrary number of terms between. Table I illustrates the representations and examples of various patterns for web text mining. The tilde operator "$\sim$" takes the word immediately following it and searches both for that specific word and for the word's synonyms. The plus operator "+" highlights the keywords that had to be included in the search results exactly as we typed them.

TABLE I: The illustration of various PTM example for text retrieval

PTM	Representation	Examples	Searched in Google	Searched in Bing/Yahoo
object pattern	object+"be"+"NoL"	"sofa was in"	"sofa +was +in"	+"sofa was in"
locality pattern	"be"+"NoL"+locality	"is on the table"	"+is +on the ∼table"	+"is on the table"
full pattern	object+"be"+"NoL"+"*"+locality	"cereal is in the kitchen"	"cereal +is +in * ∼kitchen"	+"cereal is in * kitchen"

C. *Cooccurrence Prior Query*

For web image retrieval and professional database query, we adopt the same object-location cooccurrence in [1][2][3][4] as the terms for search online. The **object query** $\mathcal{Q}^o$ is the number of hits returned by the query of noun term o. The **locality query** $\mathcal{Q}^l$ is for the number when we query the noun term l. The **full query** $\mathcal{Q}^f$ is calculated by counting the number of hits that the search engine returns when resolving "o in the l" query. The pattern taxonomy model $\mathcal{T}^f$ and the cooccurrence prior query $\mathcal{Q}^f$ will be referred to as *object-location coupling representation* in the rest of the paper.

IV. Common-sense Object Locality Knowledge

The CSOL knowledge acquired from the web can be categorized into three varieties according to the different information sources, image retrieval results, web text mining results and professional database query results. The obtained knowledge from these various sources have been successfully adopted for generating spatial concepts to perform object search tasks in indoor mobile scenarios [1][3][4][8]. The combination of these three types of CSOL knowledge will be discussed in this paper and the experimental results shown in section V will demonstrate the superior performance of this integration.

A. *Assumption about CSOL Knowledge*

A basic assumption about CSOL knowledge as presented in in [1][2][3][4][8] is that the probability of the robot locating an object at the specific place is in direct proportion to the probability of finding object-location coupling representations in all the documents that contain locality representations. The semantics involved in this assumption is, roughly, that the ratio of the hits returned by searching "object in/on the location" compared to the hits returned by searching "location" only, reflects the popularity of this object at this location, and can thus be used as the likelihood of finding this object at this location when the robot is performing the search task. Following the framework laid out in [1][2][3][4][8] the mathematical representation of this assumption is formalized as follows,

$$\begin{aligned}&\rho(find\ object\ \mathcal{O}\ at\ location\ \mathcal{L}_i) \propto \rho(\mathcal{O}|\mathcal{L}_i)\\ &\rho(\mathcal{O}|\mathcal{L}_i) = \frac{\rho(\mathcal{O}\cap\mathcal{L}_i)}{\rho(\mathcal{L}_i)} = \frac{\#\{\mathcal{O}\cap\mathcal{L}_i\}}{\#\{\mathcal{L}_i\}}\end{aligned} \qquad (1)$$

where $\rho(\mathcal{O}\cap\mathcal{L}_i)$ and $\rho(\mathcal{L}_i)$ denote the probabilities of discovering documents/images that contain searched items of "object $\mathcal{O}$ + location $\mathcal{L}_i$" or just locations $\mathcal{L}_i$ in the documents/images repository. Symbol $\#\{\cdot\}$ represents the number of hits returned by the search engine when resolving task of various queries.

B. *Object-Location Belief Model*

The aforementioned way to calculate the probability of finding a specific object at various locations satisfies the fundamental requirement of the robotic search task by evaluating the popularities of various object-location couplings. However, once there are multiple objects (either object o_1 *AND* o_2 or object o_1 *OR* o_2) requiring to be searched, comparison among probabilities of multiple objects at various locations becomes necessary for planning the most efficient motion/path. Thus the popularity of an object itself should be taken into account since in general the more commonly used object would have more description/illustration in the Internet. Therefore, we propose an *Object-Location Belief Model* (OLBM) to describe the popularities of the object itself as well as object-location coupling simultaneously. It is a belief model since it implicates how strong the robot believes that the object can be located at the location. This model can be formulated as follows,

$$\begin{aligned}&OLB = \rho(find\ \mathcal{O}_k\ at\ \mathcal{L}_i) := \frac{f(\mathcal{O}_k,\mathcal{L}_i)}{\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{m} f(\mathcal{O}_k,\mathcal{L}_i)}\\ &f(\mathcal{O}_k,\mathcal{L}_i) = \frac{\#\{\mathcal{O}_k\cap\mathcal{L}_i\}\#\{\mathcal{O}_k\}}{\#\{\mathcal{L}_i\}}\end{aligned} \qquad (2)$$

The implications of $f(\mathcal{O}_k,\mathcal{L}_i)$ can be summarized as, 1) the objects' popularities (i.e. how commonly it will be found in a general indoor environment) will be considered as the most important factor for estimating the probabilities of locating various objects at diverse indoor locations. Since typically the statement "$\#\{\mathcal{O}_k\cap\mathcal{L}_i\} \ll \#\{\mathcal{L}_i\}$" is true, when the object is popular in the indoor environment (i.e. $\#\{\mathcal{O}_k\} \sim \#\{\mathcal{L}_i\}$ or even $\#\{\mathcal{O}_k\} > \#\{\mathcal{L}_i\}$ in our test configuration), $f(\mathcal{O}_k,\mathcal{L}_i)$ can be significantly large, even much more than 1. Therefore $f(\mathcal{O}_k,\mathcal{L}_i)$ is not a probabilistic function but rather the belief which depicts the expectation made by the robot about objects' locations. 2) For uncommon object, the popularities of object itself and object-location coupling are of same importance to calculate the $f(\mathcal{O}_k,\mathcal{L}_i)$.

To apply OLB for a multiple objects search task, we discuss two different cases which require the robot to search objects with logical conjunction and disjunction relations.

1) Multiple-object under Logical Conjunction: One of the most common multiple-object search cases is the attempt to find multiple objects simultaneously, i.e. both object o_1 *AND* o_2 are required to be located in one search task. For instance, a service robot might be asked for locating and grasping a fork and a knife when the user wants to eat

Algorithm 1 Search multiple-object under logical conjunction relation

```
 1: Calculate $\forall\{\mathcal{O}_k, \mathcal{L}_j\}_{k=1,\dots,n,j=1,\dots,m}$ to generate object-location set $\{\mathcal{O}, \mathcal{L}\}$,
 2: Set $\mathcal{L}_c$ to robot's current location,
 3: if $\{\mathcal{O}, \mathcal{L}\}$ is empty then
 4:     return saved object-location pairs.
 5: end if
 6: $\forall\{\mathcal{O}_k, \mathcal{L}_j\}$ In $\{\mathcal{O}, \mathcal{L}\}$, find a pair of $\{\mathcal{O}_{max}, \mathcal{L}_{max}\}$ whih has $\max(OLB(\mathcal{O}_k, \mathcal{L}_j)/d(\mathcal{L}_c, \mathcal{L}_j))$,
 7: Move robot to $\mathcal{L}_{max}$, attempt to locate $\mathcal{O}_{max}$
 8: if NOT succeed then
 9:     decrease $OLB(\mathcal{O}_{max}, \mathcal{L}_{max})$, go to step 2,
10: else
11:     Save $\{\mathcal{O}_{max}, \mathcal{L}_{max}\}$ (or break, perform other task),
12:     delete $\forall\{\mathcal{O}_{max}, \mathcal{L}_j\}_{j=1,\dots,m}$ in $\{\mathcal{O}, \mathcal{L}\}$, go to step 2,
13: end if
```

Algorithm 2 Search multiple-object under logical disjunction relation

```
1: Calculate $\forall\{\mathcal{O}_k, \mathcal{L}_j\}_{k=1,\dots,n,j=1,\dots,m}$ to generate object-location set $\{\mathcal{O}, \mathcal{L}\}$,
2: Set $\mathcal{L}_c$ to robot's current location,
3: $\forall\{\mathcal{O}_k, \mathcal{L}_j\}$ In $\{\mathcal{O}, \mathcal{L}\}$, find a pair of $\{\mathcal{O}_{max}, \mathcal{L}_{max}\}$ has $\max(OLB(\mathcal{O}_k, \mathcal{L}_j)/d(\mathcal{L}_c, \mathcal{L}_j))$,
4: Move robot to $\mathcal{L}_{max}$, attempt to locate $\mathcal{O}_{max}$
5: if NOT succeed then
6:     decrease $OLB(\mathcal{O}_{max}, \mathcal{L}_{max})$, go to step 2,
7: else
8:     return $\{\mathcal{O}_{max}, \mathcal{L}_{max}\}$,
9: end if
```

a pizza. In this case the two required objects can usually be located at the same place, thus robot is still able to perform an efficient search by considering the predominant location of each object sequentially. However, the robot could also be asked to search for two non-related objects in one task, e.g. a magazine which is predominantly located in the living room and a cup which is predominantly located in the kitchen. With a task that searches for several non-related objects, the probabilities of multiple objects at various locations should obviously be considered in planning the most efficient trajectory and movement for the mobile robot. Algorithm 1 lists the scheme of searching multiple-object under logical conjunction relations. $d(\mathcal{L}_c, \mathcal{L}_j)$ in algorithm 1 is a cost function which measures the cost of moving the robot from the current location $\mathcal{L}_c$ to an arbitrary location $\mathcal{L}_j$. Decreasing the object-location belief after an unsuccessful search can provide the possibility of handling detection failures caused by vision algorithms, since the robot will re-visit this place when it failed to find the object at all the locations.

2) Multiple-object under Logical Disjunction: Another common multiple-object search case is the attempt to find a unique object in a set of objects, i.e. alternatively object o_1 *OR* o_2 is required to be located in one searching task. For instance, a service robot might be asked for locating and grasping a cup or a mug when the user wants to drink water. In this case, alternative plans of getting a cup or a mug might be executed by the robot, thus the robot is required to compare the likelihoods that various objects can be located in all the places. Algorithm 2 lists the scheme of searching multiple objects under the logical conjunction relation.

C. CSOL Knowledge Acquisition

The CSOL knowledge can be the semantic abstraction of $OLB(\mathcal{O}_k, \mathcal{L}_j)$ information. However, a single information source (e.g. web text mining in [8] or web image retrieval in [1][2]) is not stable enough and thus often returns incorrect or incomplete results. To improve the stability of extracted CSOL knowledge, we fuse the web mining results from various sources to generate the CSOL knowledge.

Given the probability of locating object $\mathcal{O}$ at position $\mathcal{L}$ is represented as $OLB(\mathcal{O}, \mathcal{L})_t$, which is computed using pattern retrieval for text mining in Google search engine. And using cooccurrence prior query from Bing image retrieval calculates the probability $OLB(\mathcal{O}, \mathcal{L})_i$. We utilize the same boost factor as in [1][2] to take into account the influence of the professional data in the Open Mind Indoor Common Sense (OMICS) database[1]. Then the fusion probability of finding object $\mathcal{O}$ at position $\mathcal{L}$ can be formulated as follows,

$$OLB(\mathcal{O}, \mathcal{L})_{fusion} = \left(\frac{OLB(\mathcal{O}, \mathcal{L})_t + OLB(\mathcal{O}, \mathcal{L})_i}{2} \right)^B \tag{3}$$

where $B = \frac{1}{2}$ if there are hits returned when resolving the cooccurrence search of object $\mathcal{O}$ and position $\mathcal{L}$ within the OMICS database, and $B = 1$ if the query result is empty.

V. EXPERIMENTS

In order to utilize the discovered CSOL knowledge to facilitate more efficient robotic visual search, we first create the ground truth of CSOL knowledge for accuracy evaluation. Then experiments using an indoor mobile robot scenario demonstrate the superior performance of using the discovered knowledge.

A. CSOL Knowledge Ground Truth

In order to obtain the ground truth of the CSOL knowledge, five persons (two males with good experience in robotics, two females and one male without any robotic background) were asked to label the two most predominant locations of 134 household objects (both room and supporting surface levels) and 22 types of furniture (only room level). Only 37 household objects and 15 types of furniture satisfy the condition that the two predominant locations are the same in all five assignments, if the orders of two locations are taken into consideration (case *a*). 48 household objects and the same number of furniture types can be used if only considering the correctness of the most predominant location

[1] http://openmind.hri-us.com, Honda Research Institute USA

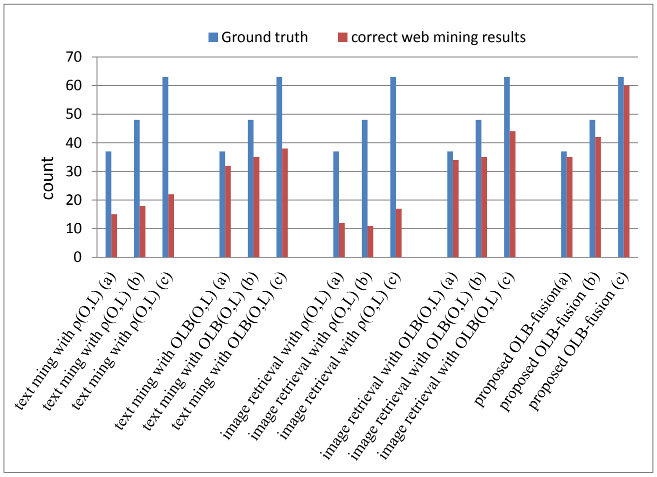


Fig. 3: Comparison of ground truth and query results of various web mining methods for discovering the CSOL knowledge of household objects.

TABLE II: The likelihood of locating a single object on various supporting surfaces

Object	Table surface $f(O, L_t)/\rho$	Floor surface $f(O, L_f)/\rho$	Sofa surface $f(O, L_s)/\rho$
book	32100000/**60.46%**	6000000/**29.71%**	505000/**9.83%**
cushion	44900/**16.32%**	38800/**37.08%**	12400/**46.60%**
blanket	51400/**8.68%**	81100/**35.99%**	31700/**55.33%**
laptop	2790000/**55.93%**	388000/**20.45%**	114000/**23.62%**
shoe	646000/**38.78%**	388000/**61.22%**	0/**0.00%**
puppy	22100/**1.56%**	419000/**77.95%**	28000/**20.49%**
kitty	661000/**5.60%**	6820000/**15.18%**	905000/**79.22%**
dog	4840000/**43.99%**	1330000/**31.77%**	258000/**24.24%**
cat	2430000/**30.89%**	1270000/**42.44%**	203000/**26.67%**

(case *b*). And 63 household objects and the same number of furniture types are accepted as ground truth when not considering the order of the two most predominant objects (case *c*). We filter out those objects where different persons disagree about the predominant locations for omitting the influence of diverse personalities and habits in the questionnaire.

We use various web mining methods to discover the CSOL knowledge, i.e. two most predominant objects, then compare the results to the ground truth and count the number of correct mining. Fig. 3 displays the counting numbers and illustrates the superior performance of the proposed CSOL knowledge discovery mechanism.

B. Object-location Beliefs Test

Table II depicts the likelihoods of locating an object on various supporting surfaces. Note that the percentages displayed in the table are calculated using Eq. 1, which means these probabilities represent the likelihoods of locating various objects in the single-object search task. Colored cells in the table highlight the predominant supporting surfaces of several examples of our experiments.

Table III and IV show the likelihoods of locating various types of furniture and household objects in the indoor environment (room category level). The cyan/orange colored cells highlight the most/second predominant locations determined using the proposed mechanism. We even test several objects/persons where no common sense about their locations can be determined, e.g. book, box, ipad, baby and kid, and illustrate these results also in Table IV. These results, to some extent, still make sense and are interesting, e.g. when a baby grows to be a kid, his/her predominant positions vary from the “bedroom+living room” to the “living room+kitchen”.

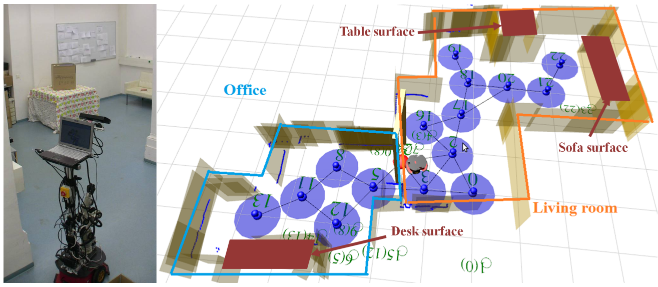


Fig. 4: Scenario and object search task at a glance, left: test scene with the robot, right: simulation/visualization of visual search task.

In case of locating two objects with logic disjunction relation, i.e. once one of the objects in the list is reached by the robot, the searching task will terminate, table V demonstrates the likelihoods of finding “book” or “box” at various locations using various methods. Updating the target object according to the current beliefs about locating various objects at all the locations, provides the flexibility and efficiency for the robotic task that requires to search alternative objects.

The full experimental results, including the likelihoods of locating 137 objects and furniture at various locations using web text mining or web image retrieval, can be downloaded from our web page [2]. Also a Python-based program for archiving these results is available there.

C. Pragmatic Test With Robotic Search Task

To evaluate the implementation of the proposed mechanism, we analyze our mobile robot system performing the multi-object search task. For the conventional single object search task, such as described in [1], [4] and [8], our mechanism is just a replacement of their off-line knowledge discovery methods, therefore similar bahaviors of the robot can be expected if their manually given probabilites of locating objects at various places just quantitatively (the predominant places for locating object are the same) differ from our online discovered OLBM. The visualization of our test environment is depicted in Fig. 4. Our experiment compares the system using the knowledge acquisition method described in this paper, to two baseline systems that discover CSOL knowledge from OMICS + image retrieval [1][2]/text mining [8].

In all the tests, a book and a box (the objects to search for) were placed in the environment, for instance the “table

[2] http://users.acin.tuwien.ac.at/kzhou/files/WMRK.zip

TABLE III: The likelihood of locating various types of furniture in indoor environment

Furniture	Living room	Kitchen	bedroom	bathroom	Dining room	Office	Corridor
armchair	28,18%	1,29%	16,02%	0,00%	0,92%	0,55%	3,04%
bed	24,00%	10,34%	49,61%	5,63%	9,06%	0,70%	0,66%
bench	7,99%	54,63%	11,61%	2,42%	7,75%	1,01%	14,59%
couch	89,67%	7,83%	0,78%	0,29%	0,59%	0,43%	0,42%
ottoman	32,76%	0,17%	15,11%	1,73%	0,00%	0,23%	0,00%
sofa	92,68%	5,29%	0,83%	0,15%	0,50%	0,28%	0,28%
television	34,79%	1,03%	58,84%	4,00%	1,25%	0,09%	0,00%
closet	21,19%	9,78%	39,05%	15,57%	1,22%	0,82%	12,36%
cabinet	13,09%	9,87%	2,79%	23,91%	46,73%	1,02%	2,59%
table	14,48%	5,66%	26,94%	9,33%	41,03%	0,63%	1,93%
desk	20,81%	3,51%	21,37%	0,61%	14,31%	29,65%	9,74%
tub	23,83%	6,79%	25,68%	43,67%	0,02%	0,00%	0,00%
piano	69,18%	4,82%	1,85%	0,47%	23,29%	0,18%	0,21%
chair	35,27%	4,83%	13,86%	15,05%	17,97%	6,71%	6,31%
dresser	26,45%	1,76%	60,03%	2,68%	8,28%	0,80%	0,00%
shelf	8,85%	22,06%	39,12%	8,89%	17,80%	2,99%	0,28%

TABLE IV: The likelihood of locating various household objects (including several non-ordinary "objects" which are animals or even nouns that refer to persons) in indoor environment

Object	Living room	Kitchen	bedroom	bathroom	Dining room	Office	Corridor
suitcase	34,41 %	0,24 %	23,59 %	1,43 %	0,48 %	0,60 %	39,26 %
soap	0,00 %	2,95 %	0,33 %	96,48 %	0,00 %	0,23 %	0,00 %
snack	0,98 %	18,72 %	0,85 %	2,02 %	18,44 %	9,00 %	0,00 %
radio	39,00 %	18,34 %	24,06 %	9,97 %	5,62 %	3,01 %	0,00 %
lamp	24,42 %	2,01 %	50,33 %	5,81 %	8,97 %	2,63 %	5,82 %
cushion	25,05 %	3,50 %	11,49 %	0,00 %	1,75 %	8,21 %	0,00 %
jacket	0,92%	3,56%	10,51%	14,80%	51,36%	18,85%	0,00%
cereal	1,22 %	32,93 %	0,00 %	4,88 %	9,76 %	1,22 %	0,00 %
candle	6,13 %	2,56 %	14,48 %	24,34 %	1,85 %	0,65 %	0,00 %
pillow	16,79 %	0,78 %	19,22 %	1,88 %	9,35 %	1,98 %	0,00 %
handbag	0,00 %	0,20 %	1,36 %	0,00 %	0,00 %	7,13 %	41,30 %
magazine	11,60 %	8,17 %	6,04 %	33,20 %	0,00 %	41,00 %	0,00 %
dish	3,83%	67,53%	2,94%	12,28%	12,83%	0,59%	0,00%
bra	4,54%	0,45%	18,68%	19,71%	0,00%	6,61%	0,00%
keyboard	15,15%	1,86%	1,14%	6,46%	3,26%	22,12%	0,00%
pot	5,93%	54,93%	2,90%	25,83%	1,88%	1,46%	7,07%
printer	21,12%	12,11%	11,21%	1,00%	0,43%	54,13%	0,00%
toy	63,32%	2,64%	21,28%	4,18%	0,00%	8,58%	0,00%
underwear	2,97%	3,40%	15,16%	66,52%	6,44%	5,51%	0,00%
guitar	16,75%	3,90%	10,62%	7,20%	4,48%	3,54%	3,51%
laptop	17,89%	7,53%	50,23%	4,44%	2,95%	16,97%	0,00%
wine	11,25%	62,74%	0,93%	2,50%	20,50%	2,08%	0,00%
briefcase	0,16%	0,16%	0,00%	1,74%	0,00%	3,48%	44,46%
bag	31,59%	20,62%	12,36%	12,43%	1,56%	8,43%	13,01%
alarm clock	1,57%	50,21%	45,20%	2,51%	0,00%	0,52%	0,00%
cherry	15,52%	16,26%	1,24%	13,38%	3,60%	0,00%	0,00%
cockroach	0,91%	2,73%	0,91%	38,18%	1,82%	0,00%	5,45%
card	4,98%	6,62%	6,39%	33,50%	2,07%	42,11%	4,33%
book	22,36%	9,62%	47,58%	5,93%	4,57%	9,95%	0,00%
ipad	8,67%	27,71%	49,94%	3,74%	0,63%	9,31%	0,00%
box	10,34%	4,80%	8,91%	6,22%	48,34%	7,15%	14,24%
kid	29,54%	23,72%	6,99%	15,39%	1,66%	19,23%	3,47%
baby	22,22%	10,07%	41,13%	17,30%	0,21%	5,04%	4,02%

TABLE V: The likelihood of finding two objects in the experimental environment

Object	Living Room with OLBM	Office with OLBM	Living Room Image retrieval	Office Image retrieval	Living Room Text mining[8]	Office Text mining[8]
book	61,77%	38,23%	7,50%	4,71%	8,15%	7,85%
box	74,93%	25,07%	2,95%	2,78%	1,72%	0,30%

surface" and "sofa surface" images in Fig. 2. A FERNS [25] object detector is running to report that the objects are successfully located by the robot. The information about the environment representation has been obtained using the same exploration process as in [1][4] (for room categories) and [8] (for supporting surfaces) beforehand. Using the likelihoods in Table V, we can make hypotheses about the behaviors of the robot using the different discovered/pre-

TABLE VI: Run time (in seconds) for various cases tested: online CSOL knowledge discovery with object-location beliefs model (on. OLBM), off-line knowledge discovery with Bing image retrieval (off. img.) and off-line knowledge discovery with Google text mining (off. txt.), both with all/partial objects in non-canonical positions. Numbers in the brackets are times of transit from one room to another. The successful executions of all the cases are recorded 5 times to calculate the average time. Offline modes have two different configurations: search the book before/after the box (left/right columns in the cells).

conf.	book loc.	box loc.	avg. time "AND"		avg. time "OR"	
on. OLBM	LR	O	412.7 (1)		188.2 (0)	
off. img			405.5 (1)	528.2 (2)	143.3 (0)	334.2 (1)
off. txt			397.4 (1)	513.6 (2)	122.6 (0)	307.8 (1)
on. OLBM	O	LR	387.3 (1)		94.9 (0)	
off. img			502.3 (2)	412.4 (1)	287.3 (1)	110.8 (0)
off. txt			487.9 (2)	422.7 (1)	313.6 (1)	99.7 (0)
on. OLBM	O	O	503.1 (1)		377.0 (1)	
off. img			745.4 (3)	735.6 (3)	322.1 (1)	331.9 (1)
off. txt			728.9 (3)	753.2 (3)	331.5 (1)	301.2 (1)

calculated beliefs/probabilities. 1) With the object-location beliefs, $G(LR) \Rightarrow S(box) \Rightarrow S(Book) \Rightarrow G(O) \Rightarrow S(book) \Rightarrow S(box)$ will be executed. 2) With the pre-calculated probabilities either from Google text mining or Bing image retrieval, behaviors $G(LR) \Rightarrow S(book) \Rightarrow G(O) \Rightarrow S(book) \Rightarrow G(LR) \Rightarrow S(box) \Rightarrow G(O) \Rightarrow S(box)$ can be predicted. $G(.)$ and $S(.)$ refer to the robot behaviors "Go to the location" and "Search for the object", respectively. Table VI demonstrates the runtime of the three configurations, it shows the superior efficiency of the proposed CSOL knowledge acquisition mechanism and also evaluates our hypotheses. Although we notice that even the run time of the object search task using the proposed method is not the shortest all the time – due to the "occasionally lucky" configuration of object search order which makes robot search at non-canonical locations in first, the proposed method is on average faster and also always provides the least transit times between various rooms.

VI. CONCLUSION AND FUTURE WORK

A common-sense object locality (COSL) knowledge acquisition mechanism by incorporating information from multiple resources has been presented in this paper. The proposed mechanism has been shown to provide plausible and reliable CSOL knowledge which depends on the proposed object-location belief model. The belief generation is achieved by considering the online popularity of the object itself and object-location coupling simultaneously. Experimental results using large numbers of household objects and furniture have demonstrated the validity of our method. The object search scenario performed by an indoor mobile robot has shown the improvement of efficiency when the acquired knowledge has been taken into consideration.

Future work will extend the fields of robotic knowledge that we could discover from the Internet, such as object affordances, or ontological representations of objects. Also utilization and fusion of more specific information gained from the Internet (e.g. time needed by the query) will be investigated in the future.

REFERENCES

[1] M. Hanheide, C. Gretton, R. W. Dearden, N. A. Hawes, J. L. Wyatt, A. Pronobis, A. Aydemir, M. Göbelbecker, and H. Zender, "Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, July 2011.

[2] H. Zender, "Multi-layered conceptual spatial mapping – representing spatial knowledge for situated action and human-robot interaction," in *Bridges Between the Methodological and Practical Work of the Robotics and Cognitive Systems Communities – From Sensors to Concepts*, ser. Intelligent Systems Reference Library, Y. Amirat, A. Chibani, and G. P. Zarri, Eds. Berlin/Heidelberg, Germany: Springer Verlag, to appear 2012.

[3] A. Aydemir, K. Sjöö, J. Folkesson, A. Pronobis, and P. Jensfelt, "Search in the real world: Active visual object search based on spatial relations," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, May 2011.

[4] A. Aydemir, M. Göbelbecker, A. Pronobis, K. Sjöö, and P. Jensfelt, "Plan-based object search and exploration using semantic spatial knowledge in the real world," in *Proc. of the European Conference on Mobile Robotics (ECMR'11)*, Örebro, Sweden, Sept. 2011.

[5] K. Zhou, A. Richtsfeld, M. Zillich, M. Vincze, A. Vrečko, and D. Skočaj, "Visual information abstraction for interactive robot learning," in *The 15th International Conference on Advanced Robotics (ICAR 2011)*, Tallinn, Estonia, June 2011.

[6] K. Zhou, A. Richtsfeld, K. M. Varadarajan, M. Zillich, and M. Vincze, "Combining plane estimation with shape detection for holistic scene understanding," in *Advanced Concepts for Intelligent Vision Systems 2011 (ACIVS2011)*, Het Pand, Ghent, Belgium, Aug 2011.

[7] K. Sjöö, A. Aydemir, T. Mörwald, K. Zhou, and P. Jensfelt, "Mechanical support as a spatial abstraction for mobile robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 18-22 2010.

[8] K. Zhou, K. M. Varadarajan, M. Zillich, and M. Vincze, "Web mining driven semantic scene understanding and object localization," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Phuket, Thailand, Dec 2011.

[9] T. Garvey and S. R. Institute, *Perpetual strategies for purposive vision*, ser. Technical note. SRI International, 1976.

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*, vol. 2, Oct. 2005, pp. 1816–1823.

[11] U. Klank, M. Z. Zia, and M. Beetz, "3d model selection from an internet database for robotic vision," in *IEEE International Conference on Robotics and Automation*, May 2009, pp. 2406 –2411.

[12] P. Cimiano and J. Wenderoth, "Automatically learning qualia structures from the web," in *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, ser. DeepLA '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 28–37.

[13] D. Pangercic, R. Tavcar, M. Tenorth, and M. Beetz, "Visual scene detection and interpretation using encyclopedic knowledge and formal description logic," in *Proceedings of the International Conference on Advanced Robotics (ICAR).*, Munich, Germany, June 22 - 26 2009.

[14] M. Waibel, M. Beetz, R. D'Andrea, R. Janssen, M. Tenorth, J. Civera, J. Elfring, D. Gálvez-López, K. Häussermann, J. Montiel, A. Perzylo, B. Schieśle, O. Zweigle, and R. van de Molengraft, "RoboEarth - A World Wide Web for Robots," *Robotics & Automation Magazine*, vol. 18, no. 2, 2011.

[15] J. K. Tsotsos, "On the relative complexity of active vs. passive visual search," *International Journal of Computer Vision*, vol. 7, no. 2, pp. 127–141, 1992.

[16] S. Ekvall and D. Kragic, "Receptive field cooccurrence histograms for object detection," in *Proc. of the IEEE/RSJ International Conference on Robotics and Automation (IROS'05)*, 2005.

[17] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. of the IEEE Int'l Conference on Image Processing (ICIP '03)*, 2003, pp. 253–256.

[18] K. Shubina and J. K. Tsotsos, “Visual search for an object in a 3d environment using a mobile robot,” *Comput. Vis. Image Underst.*, vol. 114, pp. 535–547, May 2010.

[19] C. Galindo, J.-A. Fernndez-Madrigal, J. Gonzlez, and A. Saffiotti, “Robot task planning using semantic maps,” *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 955 – 966, 2008.

[20] C. Galindo, J.-A. Fernandez-Madrigal, and J. Gonzalez, “Improving efficiency in mobile robot task planning through world abstraction,” *Robotics, IEEE Transactions on*, vol. 20, no. 4, pp. 677 – 690, Aug. 2004.

[21] D. Pangercic, M. Tenorth, D. Jain, and M. Beetz, “Combining Perception and Knowledge Processing for Everyday Manipulation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 18-22 2010, pp. 1065–1071.

[22] N. Hawes and J. Wyatt, “Engineering intelligent information-processing systems with CAST,” *Adv. Eng. Inform.*, vol. 24, no. 1, pp. 27–39, 2010.

[23] N. J. Nilsson, “Logic and artificial intelligence,” *Artif. Intell.*, vol. 47, no. 1-3, pp. 31–56, 1991.

[24] I. H. Suh, G. H. Lim, W. Hwang, H. Suh, J.-H. Choi, and Y.-T. Park, “Ontology-based multi-layered robot knowledge framework (omrkf) for robot intelligence,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 429–436.

[25] M. Özuysal, P. Fua, and V. Lepetit, “Fast keypoint recognition in ten lines of code,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, 2007.

Exploiting and modeling local 3D structure for predicting object locations

Alper Aydemir Patric Jensfelt

Abstract— **In this paper, we argue that there is a strong correlation between local 3D structure and object placement in everyday scenes. We call this the *3D context* of the object. In previous work, this is typically hand-coded and limited to flat horizontal surfaces. In contrast, we propose to use a more general model for 3D context and learn the relationship between 3D context and different object classes. This way, we can capture more complex 3D contexts without implementing specialized routines. We present extensive experiments with both qualitative and quantitative evaluations of our method for different object classes. We show that our method can be used in conjunction with an object detection algorithm to reduce the rate of false positives. Our results support that the 3D structure surrounding objects in everyday scenes is a strong indicator of their placement and that it can give significant improvements in the performance of, for example, an object detection system. For evaluation, we have collected a large dataset of Microsoft Kinect frames from five different locations, which we also make publicly available.**

I. Introduction

Locating everyday objects in indoor environments is the prerequisite of many robotics tasks such as mobile manipulation and semantic mapping. Objects are not scattered randomly in the environment. Instead man-made environments are structurally organized.

In this paper, we argue that the placement of everyday objects is highly correlated to the 3D structure of an environment – as opposed to being correlated to the appearance of the environment. As an example, cups are not typically found on tables because the former is white and the latter is brown, but the table offers physical support and easy reach. First, objects are placed in places where it is easy to interact with them; trashcans are typically on the floor and not on a high shelf. Second and more importantly, objects are placed to be physically stable at rest. As an example a cup almost never occur on a wall as there is no support for it there. We refer to the association between structure and location as *3D context*. Systems striving to efficiently locate an object should exploit *both* the shape of the object *and* the structure of the environment. One obvious benefit of this in the context of localizing objects is that although the object itself may be small or not even visible, the supporting 3D shape might be bigger and detectable at a larger distance.

Previous research has shown that exploiting structure is crucial for efficiently locating objects in complex indoor environments [1], [2]. In [3], [4] the authors exploit the notion that the performance of object detection tasks can be boosted by extracting and tracking planar surfaces since a large number of everyday objects are located on tables and shelves. Even though this approach is robust in controlled environments, not all objects rest on flat planar surfaces of a certain size and developing tailor-made methods for each situation depending on the object class is not scalable.

In another line of research, models of visual attention mechanisms aim to locate salient parts of an image. The assumption is that the sought object stands out in the image, thus creating highly salient regions which attract visual attention where the visual processes are directed to [5], [6]. This mechanism can be used to prune the search space, where an object detector is only run on salient regions of the image. This has two main advantages. First, computationally expensive algorithms can focus a subset of the image, thus lowering the overall processing time. Second, false positives that may occur in non-salient parts of the images can be eliminated. The downside with this approach is that not all target objects are visually salient, as in the case of textureless or small objects, unless this is a requirement on the target objects [7].

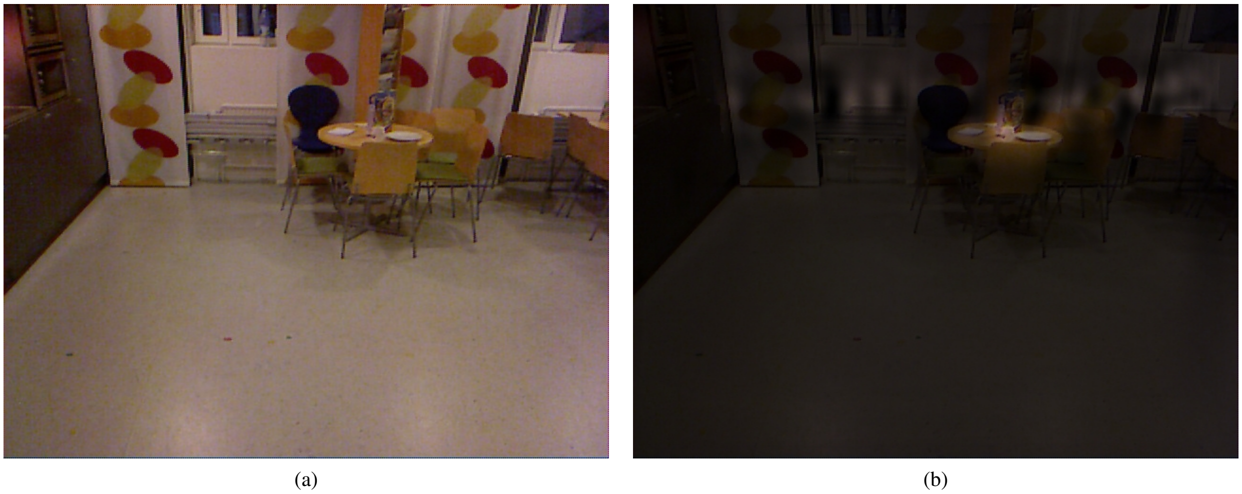
Björkman et al. demonstrate an active vision system that attempts to find and fixate on target objects in scenes [8]. The authors have utilized 3D information about the scene extracted from stereo images and detect 3D blobs that roughly correspond with the known object size. These regions are then considered more likely to contain an object. In this manner, the system will only focus on parts of the scene that are likely to contain objects..

Similarly Frintrop et al. present a saliency-based object recognition method that works with depth images constructed from a tilting laser scanner [9]. In this work, salient regions from depth data and laser intensity are combined in order to provide initial location candidates for the target object.

The work by Torralba [10], demonstrates how low-level features are extracted from the whole image for context driven attention and object detection. The appearance features are computed to capture characteristics of scenes such as spatial extent, perspective and opennes among others. This way, the implicit scene structure provides cues as to where a given object class is located in scenes.

The goal of this paper is, by leveraging the cheaply-available good quality 3D data brought by the recent advent of RGB-D sensors, to show that the 3D context of objects is a strong indicator of object placement in everyday scenes. Figure 1 gives an example output, where the system has picked out a small region in the image around the cup, corresponding mainly to the table as the most likely region

The author is with the Centre for Autonomous Systems at the Royal Institute of Technology (KTH), Stockholm, SE-100 44, Sweden. [aydemir,patric]@csc.kth.se



(a)

(b)

Fig. 1: Best viewed in color. (a) A cup on a table within a larger room. The cup occupies a very small part of the sensory data. (b) The output from our method where bright areas correspond to high probability of object presence. Most object recognizers would scan the whole image and fail to find the object in this scene. However we can rule out large portions of the image by exploiting the 3D context in which objects appear.

for a cup. No object recognition system would be able to recognize the cup itself at this distance in this image resolution. Additionally, plane fitting methods as employed in [3] might fail to detect the table plane as it occupies a very small part of the image. However, a system that exploits the local 3D structure in which objects typically occur can use this to reliably identify promising regions in the image for object presence.

A. *Contributions*

The contributions of this work are three-fold:

- We propose the use of local 3D shape around objects in everyday scenes as a strong indicator of the placement of these objects. We call this the *3D context* of an object.
- We propose a conceptually simple and effective method to capture this information. We evaluate our approach on a large RGB-D dataset across different object classes. We also quantitatively show the effect of using 3D context in an object detection task. Our results show that 3D contextual information is a strong indicator of object placement in everyday scenes.
- In order to pursue a thorough evaluation, we have collected a RGB-D data set from five different office environments in five different countries in Europe. We make the dataset and annotations publicly available to the community.

B. *Outline*

The outline of this paper is as follows. In Section II we outline important aspects of the 3D context idea. We address these aspects in Sections III and IV where we explain our method to model and exploit 3D context. In Section V we present the data collection process (V-A) and then give both qualitative and quantitative evaluations for five classes of everyday objects having very distinct 3D contexts.

II. Exploiting Local 3D Context

Most approaches to object detection look for cues/features from the object itself. In this section we describe in general terms the idea for how to make use of the 3D context to find likely object positions. The 3D context is extracted from the surrounding of objects rather than the object itself. It therefore provides information of the type ”this is a likely place for the object” rather than ”this is a likely object”. An alternative use of the 3D context could therefore also be to suggest good places to put down an object that the robot is carrying. The 3D context can nicely compliment the traditional object centred cues/features. However, as we will see in the evaluation even on its own it is quite selective as was already illustrated in Figure 1.

Figure 2 shows a scene where the target object is water tap. The presence of a kitchen sink strongly indicates that the water tap is in its image neighborhood. Parts of the scene that belong to the object's surrounding might be irrelevant to the 3D context and should not be included in the contextual model. We do not know the extent of the neighborhood that is helpful to find the water tap neither the offset between it and the object. We also need a way to represent the 3D structure appropriately. In this paper we have opted for a learning based approach where such information can be extracted from data, as opposed to manually connecting certain objects with for example planar surfaces.

In other words, certain regions of the image may predict with high accuracy the presence of an object in another region. In the previous example, the shape of a kitchen sink indicates the presence of a water tap above it. Therefore a model that captures the 3D context of an object needs to explore the object's neigborhood in the scene and find the relevant structures that consistently occur at a certain offset location (if any) with regards to the object. Also, the same object class may appear in different contexts depending on

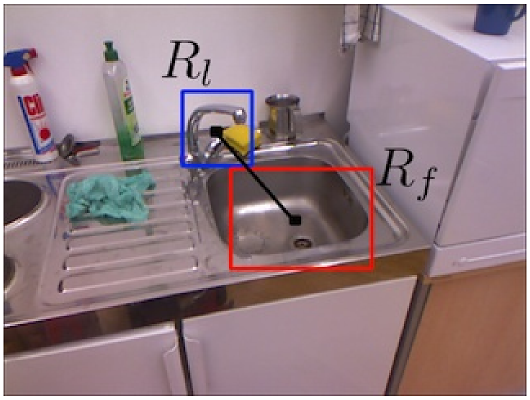


Fig. 2: The illustration of S_i regions, namely, R_f and R_l and the offset between them.

the scene. As an example, a whiteboard marker might appear to be on a vertical surface when it is attached to a whiteboard, however it may also be placed on a table that appear as a horizontal plane. Therefore, we need a model powerful enough to capture multiple contexes that an object can appear in. We argue that an approach to learn 3D context for certain object classes must capture both the location with respect to the object and the structure of the 3D context. Finally, we find that, while the 3D context captures physically plausible object placements, it can be helped by also including the height. The rationale behind using the height is that for most objects the height at which it appears is quite informative. For example, both trashcans and mugs are often found in regions of space where the surface normals are vertical (the 3D context is a horizontal surface) but the trashcan is on the floor whereas the mug is typically on a table or shelf. This difference is captured by the height.

III. Method

In this section, we will present one instantiation of the general ideas presented in Section II. We formulate the problem we address as follows. Given a 3D scene V and an object class O we want find the function

$$f_O(x, y, z, V) = P(\mathrm{X}|V, O) \tag{1}$$

This function models how likely it is for an object of class O to be at the location $\mathrm{X} = (x, y, z)$ in the scene. In our approach, we have modeled the function f_O as a binary classifier C_O, where coordinates can be labeled as being a part of the object O or not.

As discussed in Section II, C_O needs to respond to the correct region with respect to the object and needs to be able to handle multi-modal distributions. To account for this, we have chosen to compose C_O from a set of weighted weak classifiers $\mathcal{S} = \{S_1, ..., S_i, ...S_n\}$. Each weak classifier models the probability of finding the target object at (x, y, z) given a feature response, F, at an offset location $(x+o_{x,i}, y+o_{y,i}, z + o_{z,i})$. Therefore, each weak classifier explores a specific part of the object's neighborhood in the image, in line with our previous analysis. We have modeled each S_i with two rectangular image regions R_f and R_l as shown in Figure 2. Here R_f is the region from where a feature response is computed (i.e. the neighborhood hypothezised to be correlated with the object) and R_l is the object region. The weak classifiers are parametrized by the width and height of both rectangular regions and the offset between them. We model the 3D context learning as a regression problem, where the goal is, for each S_i to learn the mapping from the feature response of region R_f, to objectness in some other region R_l.

During training we provide the system with a set of annotated RGB-D images. For each S_i we calculate the feature response in R_f and the objectness for the corresponding region R_l. The objectiveness is calculated as the fraction of R_l that overlaps with the target object's annotation bounding box. We do this for each object class and every S_i, i.e. combinations of region sizes and offsets. Given this, we then move on to learn which set of the weak classifiers should be utilized[1] to get the resulting classifier C_O. That is, we learn for which S_i there is a strong correlation between a certain feature response in R_f and the object being in R_l. This is formulated as learnin the weights for each classifiers which effectively results in only a subset of the classifiers being used. Using a subset of all the weak classifiers reduces the model complexity and results in a faster processing time. During testing we slide the R_f regions corresponding to the appropriate S_i over the image and get the object responses in the corresponding R_l regions. The response from all active S_i are weighted together according to the learned weights.

IV. Implementation

In the first step of the training procedure, we slide each S_i across training images. For each, S_i we train a Support Vector Machine (SVM)[2], H_i, with a generalized radial basis function (RBF) kernel to learn aforementioned mapping from a surface patch R_f to the object response in R_l at a certain offset. This results in the set of regressors $\mathcal{H} = \{H_1...H_m\}$. We have chosen to use a discriminative approach since it has been shown that when the amount of labeled training data is large and samples accurately the problem space, discriminative methods tend to work better than their generative counterparts in terms of predictive performance [12], [13], [14]. Furthermore, SVMs being a well understood discriminative method, generally offer lower computational complexity, which is desirable in a robotics context. RBF kernels are shown to provide good results with histogram features [15], [16], which is the feature type used in this paper as we will explain later on.

The problem of combining a set of weak regressors to obtain a strong one is well researched in the field. We have chosen the widely used greedy gradient boosting algorithm described in [17] to calculate the vector β which gives the

[1] In the water tap example from before, one S_i might encode the correlation between a sink and a tap behind it and another the correlation between a vertical wall and a tap beneath it.

[2] We have used a modified version of SVM implemented in [11], with $\sigma = 9.7$ and $\epsilon = 0.001$

TABLE I: The locations from which the dataset is collected and room types from each location.

Site	Room types
University of Birmingham	kitchen, classroom, corridor, office, meeting room, computer lab
DFKI - Saarbrücken	corridor, office, meeting room, computer lab
Technical University of Vienna	kitchen, office
University of Ljubljana	kitchen, corridor, office
Royal Institute of Technology	kitchen, corridor, office, meeting room, computer lab

weights with which the output from $\mathcal{H}$ will be scaled. The elements of the weight vector β represents how much each H_i should contribute to the end result. The resulting overall regressor is then given by:

$$\mathcal{C}(F(R_f)) = \sum_{i=1}^{n} \beta_i \cdot H_i(F(R_f)) \qquad (2)$$

where $F(R_f)$ is the feature response for region R_f and $H_i(F(R_f))$ is the i^{th} weak classifier's output. The weighthing process results in a large subset of $\mathcal{H}$ having zero or near zero weights. This corresponds to regressors that has low correlation with object presence. The remaining regressors assigned with high weights allows the method to form multi-modal context models as discussed in Section II.

A. Features

We want to pick a simple, fast to compute yet expressive feature to represent R_f . We refrain from using features that build detailed models of specific 3D shapes, rather we prefer a rough descriptor of a surface patch to capture overall contextual information. For this reason, the feature we have chosen in this paper is the histogram of surface normals which has been shown to efficiently describe local geometry in 3D point clouds [18]. For a point $\mathbf{p}$ in R_f, a surface normal is computed by fitting a plane to the set of points which are inside a sphere whose center is $\mathbf{p}$ and radius is r. We perform this operation for each pair of points in R_f and obtain a set of vectors. Then these vectors are binned in a 3-dimensional histogram with each dimension containing 8 bins, resulting in 512 bins in total. The height of the center of the object in the scene according to the annotation is concatenated to the feature vector.

V. Experiments

The evaluation of a context learner for indoor environments requires a large amount of diverse and real world data. For this reason, we first explain our data set and later on present qualitative and quantitative experimental results.

A. RGB-D Database

We have constructed our dataset from five different sites in Europe; the Technical University of Vienna (TuV), the University of Birmingham (UB), the Royal Institute of Technology (KTH), the German Center for Artificial Intelligence in Saarbrücken (DFKI) and the University of Ljubljana (UL) (see Figure 3). At each site, a Pioneer 3dx robot equipped with a Microsoft Kinect camera at 1.4 m height and $-20°$ tilt angle was used. The robot is controlled by a human operator using a joystick. It is important to note that the human operators did not know about the purpose of the method presented in this paper so as not to bias the collected data . Rather they were instructed to simply drive the robot in their work environment. The images are continuosly saved as the robot moves through the environment. The dataset can be used for other purposes such as testing 3D mapping and place recognition methods. The data set is available for download from
`http://www.cas.kth.se/rgb-d`.
Table I details the room types included from each site.

The dataset contains approximately 250,000 Kinect frames. We have annotated five object classes in the dataset:

- cup
- trashcan
- whiteboard marker
- wallplug
- water tap

In total, 1627 images were annotated. The objects in the list were chosen for being frequently found in typical indoor environments and for having diverse context, location and size. As an example, trashcans are on the floor typically near a wall whereas cups are on flat surfaces at a typical table height. Other objects such as whiteboard markers have a less clear contextual one-to-one mapping and can both be on tables and appear to be on a wall. Furthermore, whiteboard markers typically occupy very little space in images in contrast to bigger object such as trashcans. The object water tap has a very distinct 3D context, however the type of scenes it is usually found in are quite cluttered as can be seen from Figure 3e.

B. Evaluation

For the evaluation of the method we have selected the KTH, Birmingham and DFKI datasets as the training set and Ljubljana and Vienna as test sets. The training set corresponds to roughly 70% of the images.

1) Qualitative analysis: We will first go throught the set of example images and responses shown in Figure 6 in order to present a qualitative evaluation of the proposed method. From the top down, the rows of Figure 6 correspond to the objects cup, trashcan, whiteboard marker, wallplug and water tap

The top row of Figure 6 shows two example images for the object *cup*. We can see that the learned context for cups is flat surfaces at the height of a typical table. The method gives high response to these areas as can be seen in figure 6a. In figure 6b we can see that the large portions of the image can be ruled out. Interestingly, the method gives a high response for the bottom part of the whiteboard where there's a small flat surface to place whiteboard markers. In this case the method predicts that the small flat surface that is at a similar height of a table can also afford to support a cup.

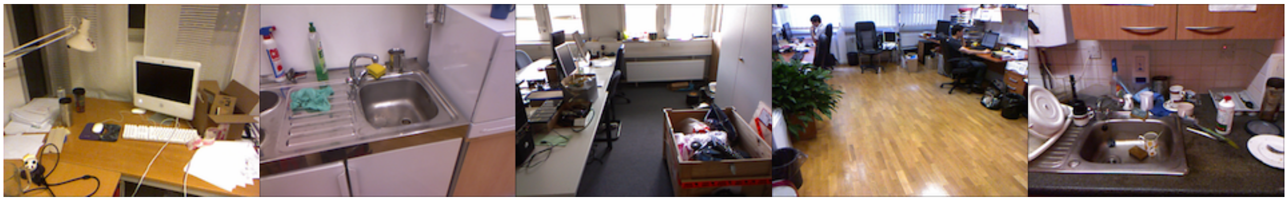

(a) KTH (b) TuV (c) DFKI (d) UB (e) UL

Fig. 3: Example images from the dataset.

In the second row, Figure 6f, we see that the method learned that a *trashcan* is typically found on the floor but even more specifically at the intersection of a floor and a wall. The high response over the area that corresponds to the object in Figure 6f and 6h comes from the fact that the sides of the trashcan also appear as a vertical surface on the floor thus resembling a wall.

In figures 6j and 6l, the two images show that the *whiteboard marker* appears in very distinct contexts, one on the table and the other on the whiteboard. It's promising that our approach has captured the multi-modal context in which the object appears in the training set. This is captured by our method without the need of a specialized algorithm for each case, such as a table detector and a wall detector as it would be needed in previous work [3].

Results for the object *wallplug* are displayed in the fourth row of Figure 6. Similar to whiteboard markers, there are mainly two types of wallplug placements in the dataset: at the intersection of a table and a wall and the other is on a flat wall surface near the floor or at the shoulder height. Figures 6n and 6p show that both contexts are successfully captured by the proposed method. The Figure 6p is a corridor scene consisting of flat surfaces. In this case, what prevents the method from predicting a high response over all of the walls is the height component in the feature vector.

Looking closely at a cluttered scene, Figure 5 shows a kitchen scene where the object *water tap* is present. We have picked this object since it has a very complex 3D context which can be disturbed a great deal as a result of clutter such as cups, plates, washing liquid, sponge, dirty dishes. However, certain aspects of its context is very persistent, there is almost always a sink in front of the water tap, which itself has a distinct 3D shape. Another cue is that typically there is a flat wall behind water taps. We can see that the method's highest response has a big overlap with the object itself. The area inside the sink, the counter top, the side wall and the oven has lower repsonses. However we also note that the method gives considerably high responses overall in the image. There are several reasons for this. One is that the high clutter in the training scenes makes it harder to extract the true 3D context of the object. Another is that, using the Kinect camera, scenes containing shiny surfaces result in large amount of noise in depth.

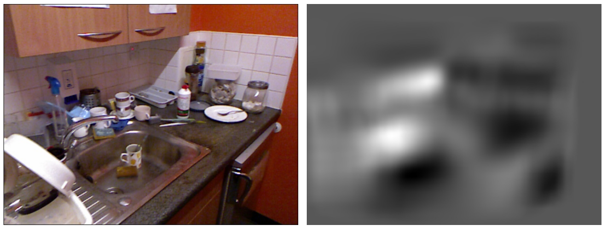
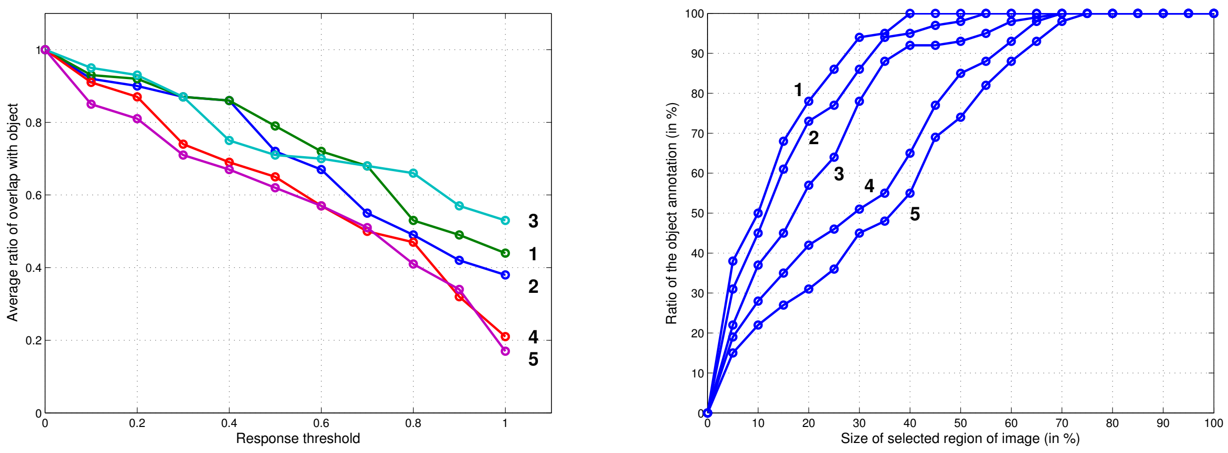

(a) (b)

Fig. 5: An example kitchen scene with a sink and water tap and clutter. The bright areas correspond to higher object presence as predicted by the local 3D context idea. Notice that even though the part of the image corresponding to sink is itself remains dark, the regions above it are computed as promising areas.

2) Quantitative analysis: We have evaluated our method quantitatively in three experimental setups to gain statistical insights. First, we have checked how much of the actual object is in the predicted region associated with a specific probability threshold. This tells us how accurate the method is in its location predictions. Second, for varying sizes of the search region represented as the percentage of the image we have computed the overlap of the object's bounding box with this region. This quantifies how much of the search region we can rule out without losing parts of the image that actually contains the object. Third, we have computed the average precision of a state-of-the-art object detection algorithm with and without utilizing 3D context in an object detection task.

In the first setup, we have thresholded the response image (here response image is the visualization of object likelihood as shown in Figure 1b) to obtain a binary image with white and black regions. The horizontal axis of Figure 4a corresponds to values of this threshold. We compute the overlap between the object's bounding polygon and its predicted location, i.e. the white regions in the image, which is shown in the vertical axis. The amount of overlap tells us if objects in question typically occur where the algorithm predicts them or not. This overlap is 100% when the threshold is zero, meaning when all of the image is selected. From these results, we can say that intuitively the method performs better for objects that blend with their 3D context. As an example, in depth, the whiteboard marker is almost indistinguishable from the table or the wall itself. This can be used as a complement to object recognition algorithms where an object of the size of whiteboard marker is usually very hard to



(a)

(b)

Fig. 4: Results quantifying the effect of 3D context as a function of (a) probability threshold (b) search region size for different objects (1 - cup, 2 - whiteboard marker, 3 - wallplug, 4 - trashcan, 5 - watertap).

detect. Furthermore, it is encouraging that even with high threshold values, the majority of the bounding box of cup, whiteboard marker and wallplug is included in the predicted region.

The performance drops faster for bigger objects such as trashcan as the threshold value increases. One reason for this is that these objects are typically observed from a wide variety of viewing angles. The result is that such objects can appear to be in very different backgrounds which are not part of the context. Consider the case where a chair is seen from a top and side view; from the top, it would seem that it is on the floor perhaps near a table, however the side view would show the distant scene in the background unrelated to the local 3D context.

Another reason is that such objects tend to *stick out* of their 3D context. With this we mean that a computer monitor's context consists largely of table surfaces, the average distance of the 3D points on the monitor to the table surface is much greater than for a whiteboard marker. This resulted in a negative bias in the type of evaluation we have selected in this paper.

In the second setup, we fixed a percentage of the image that is predicted as most likely to contain the object according to the method and check if the object falls in this region. As an example, we pick a number of the pixels that constitutes 10% of the image that are the most likely to contain the object. This selection criteria checks the method's performance for varying gains in efficiency (in this case 90% of the image is eliminated). We can see that for all objects, more than 70% of the target object's bounding box is still in the predicted region when we discard 50% of the image. For all object classes, in average at least one third of the image can be ommitted without missing the object. This is a promising result indicating that objects are embedded in their 3D context and much can be gained from exploiting it.

We have also investigated the effect of exploiting 3D context as a first step in an object detection application using the method presented in [19]. We have trained each object class using the implementation provided in [20]. After this we have provided two sets of test images to the object detection algorithm: raw images and images that are combined with the thresholded 3D context response. In order to construct the latter image, we have made an informed choice by looking at the results of the second experimental setup. We have thresholded the image such that the least promising 40% of the test images are omitted from the search region. The rationale behind this is that, as Figure 4b shows, we rule out very large portion of the image while still retaining nearly 90% of the object for all classes. This has the effect of eliminating false positives. The average precision results are shown in Table II where an increase in detection performance is observed on all classes, with variable success. We see that the objects that benefited most are small objects such as whiteboard marker and cup since the texture of these objects cannot be captured sufficiently due to their size and thus are most likely to be mistaken for other parts of the image. These results indicate that exploiting the 3D context of objects to predict likely locations greatly improves the performance of a state-of-the-art detector for all object classes. The object with the least amount of gain from incorporating 3D context is watertap and trashcan. We think this can be explained by the fact that mostly these objects occupy a large portion of the image in the data set and are harder to confuse with false positives.

In a robotics context, time performance is of crucial importance. The object detection algorithm used in this experimental setup [19] takes on average 23 seconds to compute a detection for object class. This is prohibitively slow for most robotics applications. The method presented in this paper takes between 0.6 to 2.1 seconds to run on a

TABLE II: Object detection results with and without the method presented in this paper.

Object label	AveP	AveP with
Cup	0.614	0.813
Whiteboard marker	0.332	0.516
Trashcan	0.541	0.774
Wallplug	0.214	0.519
Watertap	0.221	0.317

single image of 640x480 resolution on a computer with a 2.26 GHz CPU. Each weak classifier approximately takes 60-80 ms to run. We have chosen the parameter intervals of weak classifiers to cover almost all of the relevant part of the parameter space. The range of window sizes for R_f and R_l are between 10x10 to 100x100 pixels with 10 pixel increments and the offset is chosen in the interval ±50 pixels with 10 pixel increments. This results in training 250 weak classifiers trained for each object which covers a very large part of the parameter space for the objects types present in this paper. We then also rule out weak classifiers that has a weight lower than 0.01 as they contribute negligibly to the final result.

VI. Discussion and Future Work

In this paper, we have proposed to use local 3D structure as a strong cue in object placements in everyday scenes, we call this the *3D context* of an object.

We presented a method to extract the 3D context of everyday object and provide extensive quantitative evaluation on a large dataset collected from different natural work environments in Europe. Furthermore, we have quantified the benefit of exploiting local structure in an object detection application. The results that local structure surrounding objects is certainly a strong indicator of object placement in scenes and our method is able to accurately predict the location of the everyday objects included in the study.

It has been in previous work that humans possess strong priors about natural scenes [21], [22]. We don't expect objects to float in the middle of a room and we have strong expectations on what type of objects to expect in which scenes and where in the scene. We believe that embodied systems should also extract and exploit the structure of the world they perceive to consume information more efficiently. In most mobile robotics systems, a 3D representation of the world is built and maintained for safely navigating the world and manipulating objects. Therefore a robot equipped with a camera should make use of this information also when analyzing scenes. Entering a kitchen looking for a cup and being presented with the scene in Figure 1, we do not systematically scan the whole image including the floor and the walls, instead we almost instantly fixate on the table to get a higher resolution coverage of the table top. In this paper, we have argued that the local 3D structure of a scene, and specifically the 3D context of an object is an important cue in determining object placements in scenes.

A limitation of all context based approaches is that they are expected to perform poorly in uncommon scenes, where contextual expectations do not agree with the scene at hand. In this case, a global search over the whole image is needed, which is often more expensive than only searching the regions indicated by contextual cues. This is a penalty that is also observed in biological systems [21]. One problem however is that the system needs to detect that a scene is out-of-context, in order to make the decision to perform a global search. The authors in [23] exploit the notion that objects are physically supported by other objects in scenes, similar to the argument used in this paper and in our previous work [24]. Using this, the proposed method out-of-context objects and scenes.

Future work includes investigation other methods to capture the 3D context objects as with this work we have presented the idea and implemented one instantiation of it. Furthermore, we would like to combine an RGB-D camera with a high resolution photo camera to be able to obtain detailed views of regions predicted by the approach and test the effect of this in object detection results. Finally we are interested in employing the 3D context idea in an place categorization framework.

VII. Acknowledgements

We thank Johan Ekekrantz for his help with the development. The authors gratefully acknowledge the help of Peter Ursic, Hendrik Zender, Kai Zhou and Marc Hanheide with the RGB-D database collection. This work was supported by SSF through its Centre for Autonomous Systems and the EU FP7 project CogX.

References

[1] A. Aydemir, K. Sjöö, J. Folkesson, A. Pronobis, and P. Jensfelt, "Search in the real world: Active visual object search based on spatial relations," in *in Proc. of the IEEE International Conference on Robotics and Automation (ICRA'11)*, 2011.

[2] A. Aydemir, M. Göbelbecker, A. Pronobis, K. Sjöö, and P. Jensfelt, "Plan-based object search and exploration using semantic spatial knowledge in the real world," in *in Proc. of the European Conference on Mobile Robotics (ECMR'11)*, Örebro, Sweden, Sept. 2011.

[3] Z. C. Marton, R. B. Rusu, D. Jain, U. Klank, and M. Beetz, "Probabilistic categorization of kitchen objects in table settings with a composite sensor," in *The 22nd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, St. Louis, MO, USA, 10/2009 2009.

[4] D. Pangercic, M. Tenorth, D. Jain, and M. Beetz, "Combining perception and knowledge processing for everyday manipulation," in *in Proc. of IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS'10)*, 2010.

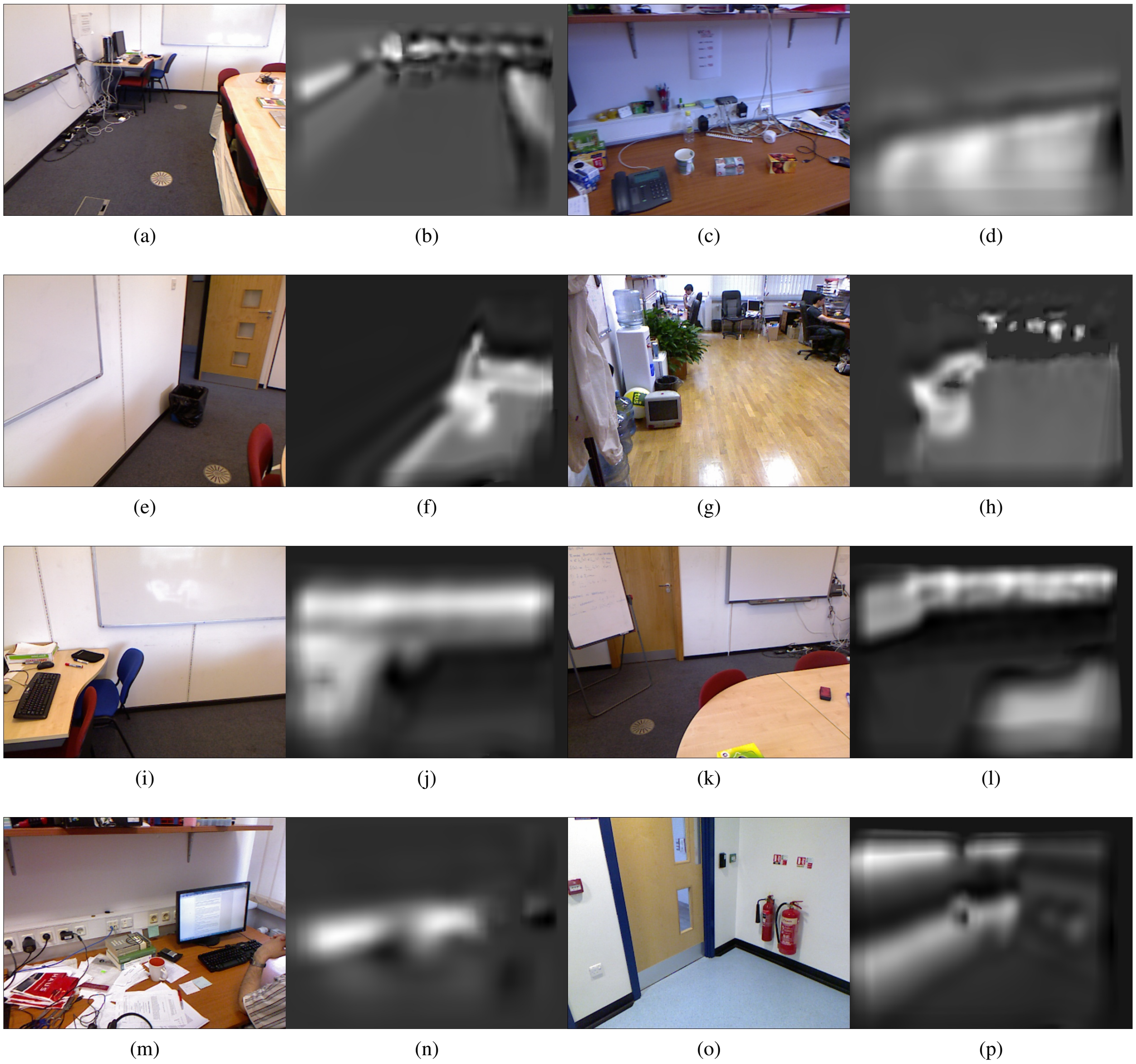
[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] S. Frintrop, "Vocus: A visual attention system for object detection and goal-directed search," Ph.D. dissertation, University of Bonn, July 2005.

[7] S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual SLAM," *IEEE Transactions on Robotics, special Issue on Visual SLAM*, vol. 24, no. 5, Oct. 2008.

[8] M. Björkman and J.-O. Eklundh, "Vision in the real world: Finding, attending and recognizing objects," *International Journal of Imaging Systems and Technology*, vol. 16, no. 5, pp. 189–208, 2006.

[9] S. Frintrop, A. Nuchter, H. Surmann, and J. Hertzberg, "Saliency-based object recognition in 3d data," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, sept.-2 oct. 2004, pp. 2167 – 2172 vol.3.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

Fig. 6: Example images from the dataset and responses for the objects cup, trashcan, whiteboard marker, wallplug and water tap respectively. Brighter areas correspond to a higher likelihood of object presence.

[10] A. Torralba, “Contextual priming for object detection,” *Int. J. Comput. Vision*, vol. 53, pp. 169–191, July 2003.

[11] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, May 2011.

[12] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *NIPS*, 2001, pp. 841–848.

[13] J. Lasserre, C. Bishop, and T. Minka, “Principled hybrids of generative and discriminative models,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, june 2006, pp. 87 – 94.

[14] T. Deselaers, G. Heigold, and H. Ney, “Object classification by fusing svms and gaussian mixtures,” *Pattern Recognition*, vol. 43, no. 7, pp. 2476–2484, 2010.

[15] V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman, “Generalized RBF feature maps for efficient detection,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

[16] O. Chapelle, P. Haffner, and V. Vapnik, “Support vector machines for histogram-based image classification,” *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1055 –1064, sep 1999.

[17] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[18] R. B. Rusu, N. Blodow, and M. Beetz, “Fast point feature histograms (fpfh) for 3d registration,” in *The IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 05/2009 2009.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, Sept. 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, “Discriminatively trained deformable part models, release 4,” http://people.cs.uchicago.edu/ pff/latent-release4/.

[21] Biederman, “On the semantics of a glance at a scene,” in *Perceptual organization*, 1981.

[22] M. Bar, “Visual objects in context,” *Nature Reviews: Neuroscience*, vol. 5, pp. 617–629, August 2004.

[23] A. S. W. Myung Jin Choi, Antonio Torralba, “Context models and out-of-context objects,” *Pattern Recognition Letters*, vol. to appear, 2012.

[24] K. Sjöö, A. Aydemir, T. Mörwald, K. Zho, and P. Jensfelt, “Mechanical support as a spatial abstraction for mobile robots,” in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS’10)*, Oct. 2010.

Kinect@Home: Crowdsourcing a Large 3D Dataset of Real Environments

Alper Aydemir, Daniel Henell and **Patric Jensfelt**
CVAP, KTH, Stockholm, Sweden
aydemir, dhenell,patric@kth.se

Roy Shilkrot
Fluid Interfaces Group, Media Lab, MIT
roy.shil@gmail.com

Abstract

We present Kinect@Home, aimed at collecting a vast RGB-D dataset from real everyday living spaces. This dataset is planned to be the largest real world image collection of everyday environments to date, making use of the availability of a widely adopted robotics sensor which is also in the homes of millions of users, the Microsoft Kinect camera.

Introduction

Robotics has a long-standing aim to build robots that can function in complex man-made environments. The long term vision (which is rapidly becoming a short term goal) of robotics is to help humans with tedious and hard tasks, e.g. assisting elderly in everyday tasks, providing care for disabled persons for increased ability or performing hard, hazardous and tedious tasks that are unfit for human health.

In order to determine and accomplish such tasks, the robotics researcher usually *guesses* the tasks needed or the environments used by a typical user of such robots in the real world and tries to come up with various problems and solutions regarding perception, action and planning in robotics. The proposed solutions generally lacks the basis for the robustness as they are not tested in complex real environments with the intended end user. This leads a mismatch between what is promised in publications and their actual performance which is a growing concern as the pressure on robotics as a field to provide working products increases. For this reason, we present the Kinect@Home project.

Kinect@Home

The Kinect@Home project is aimed at collecting a vast dataset of Microsoft Kinect images of real everyday living spaces such as offices, homes and alike. The project location is at http://www.kinectathome.com. We have chosen the Microsoft Kinect camera because it provides both an RGB image and a depth value for each pixel of the image. Thanks to its high quality 3D data for its low price, the Kinect camera has been rapidly adopted as a robotics sensor. Most importantly, it has since entered the homes of some 20 million users therefore fit for a crowdsourcing task. The significance of this is being, never before a highly used robotics sensor was at the home of millions of people, therefore it presents ample opportunity for a crowdsourcing application.

Datasets in computer vision and robotics are widely used for testing and benchmarking various algorithms such as object recognition and detection, mapping and image segmentation. Already there exists several Kinect datasets (Kevin et al. 2011; Min et al. 2010; Garage 2011; Silberman and Fergus 2011; Koppula et al. 2011) mainly on the topic of object recognition and detection in scenes. We welcome these efforts and find them very encouraging. Closest to our approach is (Janoch 2011) where individual images of indoor scenes are being collected. However none of these datasets aims to capture the challenging real world scenes that a robot shipped to a home today might face. We believe we can make a big impact by collecting a large dataset of real world environments for developing better methods.

In robotics, various research groups have opted to recreate the man-made environments that these robots are intended to work in by building mock versions of living spaces such as kitchens and living rooms in their laboratories. These environments certainly serve as an initial testbed for algorithms and methods as a way of validating the plausibility of the proposed approach. However, there are several shortcomings regarding evaluating robot performances in simulated of living spaces. First, since only a few instances of the said home environments can be built, the evaluation of the proposed methods tends to include only a few cases of a general problem. Second, the environments tend not to be realistic and instead become over simplified, as no human lives and uses these spaces on a daily basis. We therefore propose the Kinect@Home project as a way to collect large amounts of 3D data from ordinary people's everyday environments. With this project, we will amass a large dataset of everyday indoor environments such as offices, kitchens, living room spaces. This data will be used for various applications such as object detection, recognition, 3D mapping and various other robotic applications. The dataset will be available publicly at the interest of all interested researchers.

In order to construct such a dataset, the software implementation should have certain specifications. We will continue by briefly describing our software architecture.

Software architecture and usage

The software architecture consists of two parts: clients which are ordinary people uploading Kinect frames and the server which collects the uploaded data. There are several considerations for building the software implementation that realizes the dataset. First of all, we want to minimize the number of steps a user has to take in order to accomplish the task. Therefore we avoid hefty downloads, installation guides or tedious tutorials. This means we cannot simply ask the user to download and install a program, record the Kinect frames to file (which would take a few gigabytes of data) and send over to us.

We have chosen a browser plug-in as the client since it provides a much more light-weight installation compared to a stand alone program both technically and in the minds of regular internet user. Furthermore by doing this the user interface will be HTML-based and by default cross platform. The plug-in is programmed using the FireBreath cross platform browser plugin framework (Firebreath 2012).

We want the server to be as simple as possible and general enough to accept any type of client that may be realized in the future. Furthermore, the bandwidth and heavy hard disk file operations involving receiving large amounts of images need to be considered. For this reason, we have opted for an HTTP RESTful API using the Django web framework. We have considered frameworks such as ZeroMQ, Apache thrift, rpclib (Arslan 2012; Hintjens 2010). We will skip over the detailed discussion for the lack of space in this extended abstract, however they all seemed to need a significant amount of infrastructure, front-end code and a complete user-interface. Instead, HTTP REST calls are a fairly basic and almost ubiquitous standard used throughout the internet.

The raw Kinect data is too big to be uploaded without compression, we assume the typical user would not wait for the whole upload period. Therefore we compress the data stream with near-lossless video encoding. We compress and upload the data in chunks. We have tried several compression techniques cite. The RGB data is compressed using x264 codec and the depth stream is encoded lossless using FFV1 for 16bit depth images. This way the amount of HTTP calls and computational overhead is reduced compared to uploading every frame individually.

Upon reaching the website, the user will be prompted to connect their Kinect devices and install the plug-in. Once this is done, the website starts showing the live Kinect images on the browser as a confirmation that the software is working accordingly. This also helps to display the user the currently captured data. A *Record* button and an optional email adress text box is also displayed the purpose of which we will explain in more detail. Once the button is pressed, the plug-in starts uploading captured frames to server. After a set period of time or when the user hits the *Stop* button, the recording stops and the user is prompted with an optional text box for metadata about the video. A progress bar indicates how much of the data is sent to the server.

Privacy and control of the data

In order to alleviate any user trust and user related problems we give full control to the data uploader. If the user provides an email address, we email the participant with a PIN code after each recording and the unique identification number of the specific upload. With these credentials, the user can view or delete the uploaded files anytime, with no questions asked. Our code base is entirely open source. As part of addressing the privacy concerns, we don't keep any user-related data whatsoever. The users however need to agree a terms of service agreement, which basically states that the data uploaded will be used for scientific purposes.

Conclusion

We have presented a crowdsourcing platform for collecting Kinect camera images. We will share our findings about the software architecture and the wider public's reactions in the coming months during the symposium. The system is open source and the data will be completely anonymous and publicly available. We expect a high participation.

Acknowledgements

The authors thank Javier Romero and Burak Arslan for their suggestions.

References

Arslan, B. 2012. Remote procedure call library.

Firebreath. 2012. Cross-platform browser plugin framework.

Garage, W. 2011. Solutions in perception challenge.

Hintjens, P. 2010. ZeroMQ: The Guide.

Janoch, A. 2011. A Category-Level 3-D Object Dataset: Putting the Kinect to Work. In *ICCV Workshop on Consumer Depth Cameras for Computer Vision*.

Kevin, L.; Bo, L.; Xiaofend, R.; and Fox, D. 2011. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 1817–1824.

Koppula, H.; Anand, A.; Joachims, T.; and Saxena, A. 2011. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*.

Min, S.; Bradski, G.; Bing-Xing, X.; and Savarese, S. 2010. Depth-encoded hough voting for joint object detection and shape recovery. In *Proceedings of European Conference on Computer Vision*.

Silberman, N., and Fergus, R. 2011. Indoor Scene Segmentation using a Structured Light Sensor. In *ICCV Workshop on 3D Representation and Recognition*.