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.

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.

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.

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

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:

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

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:

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

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

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

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.

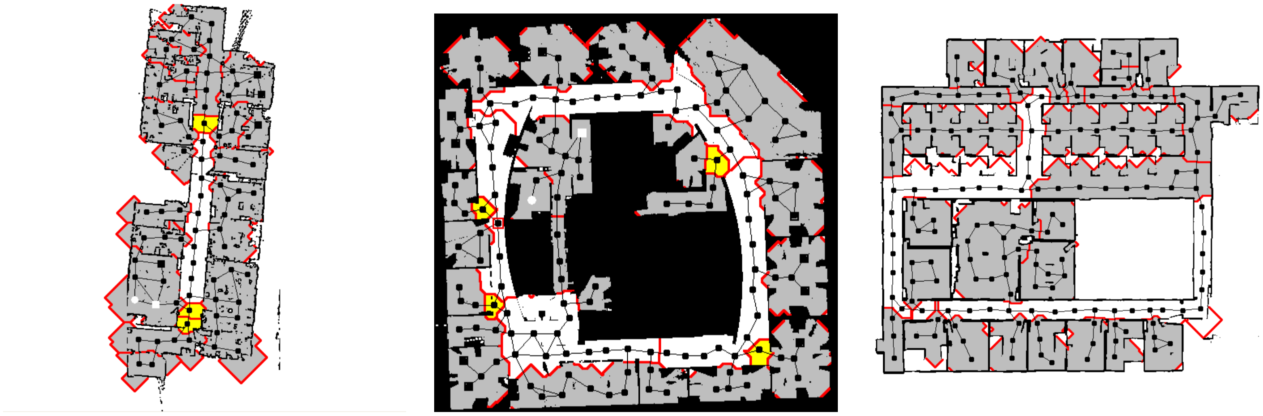
VI. CONCLUSIONS

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.

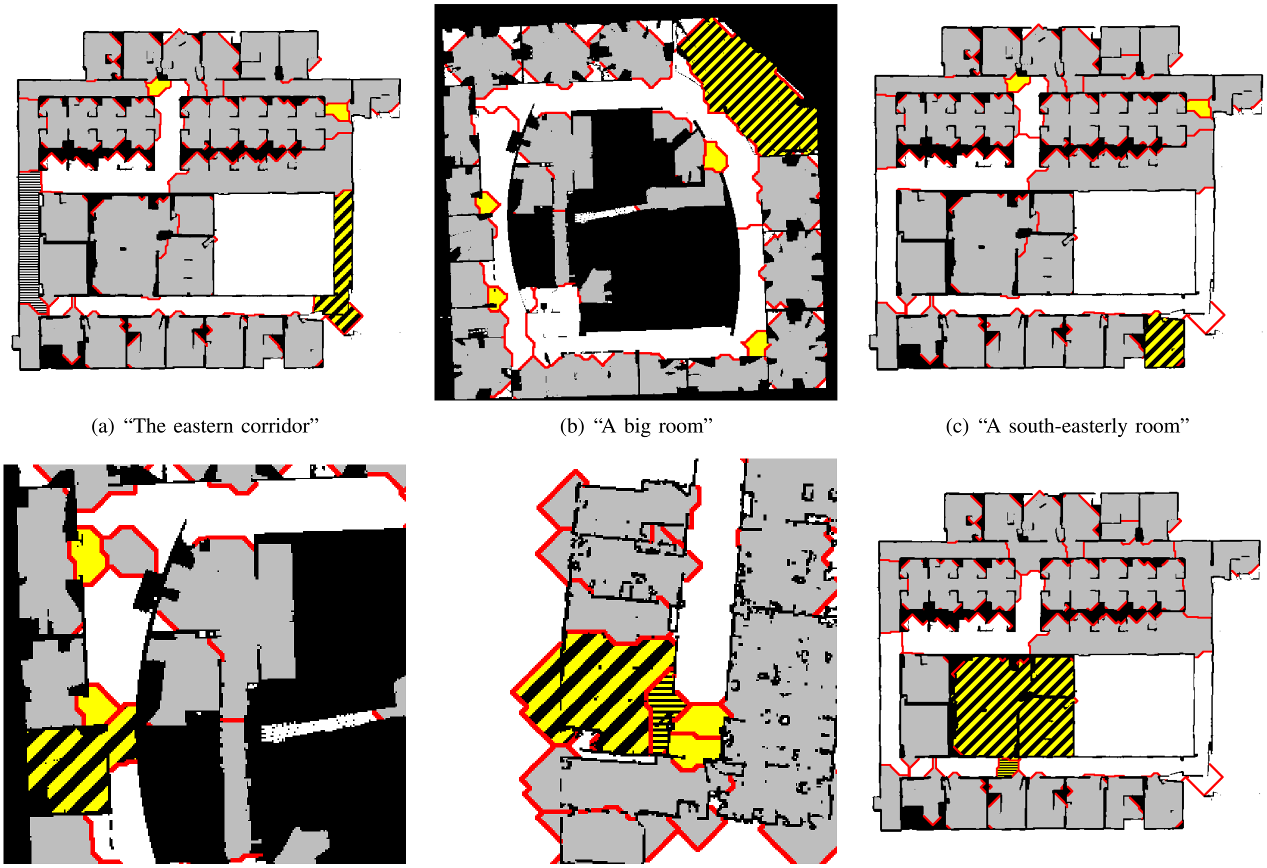


(a) FR079 map

(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)`

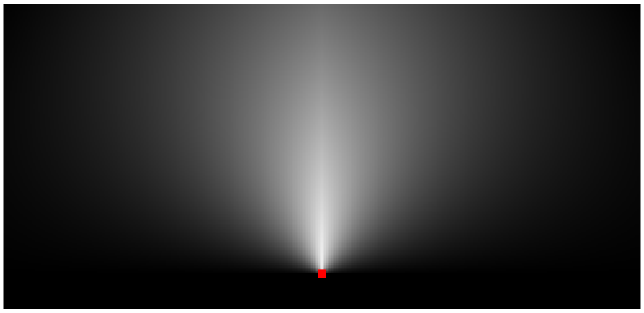


Figure 1: A visualisation of a the strength of a spatial relation across a region. The landmark is the red square and the direction vector used was $\langle 0, 1 \rangle$ (i.e. above of the landmark). The lighter the pixel the stronger the spatial relation is deemed to be at that point.

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).

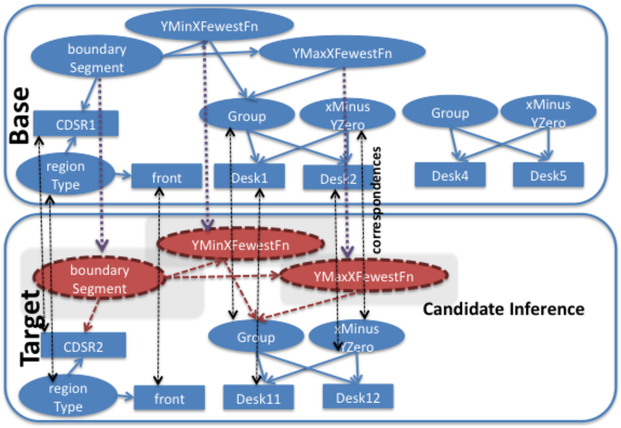


Figure 3: Analogical mapping between six base expressions and three target expressions.

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

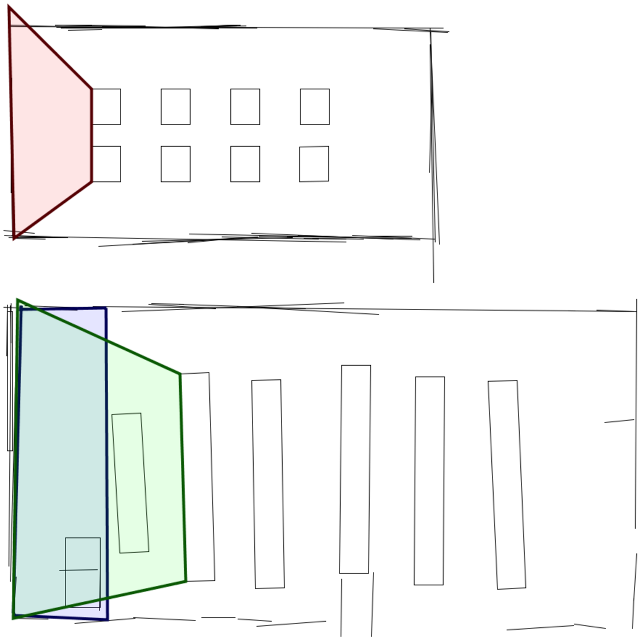


Figure 4: Maps of 2 real classrooms generated by our system. The lines around the perimeter are walls, the unfilled polygons are the outlines of objects and the filled polygons are CDSRs. The maps show an expert-annotated CDSR (red, top image), a subject-annotated CDSR (blue, bottom image) and a CDSR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

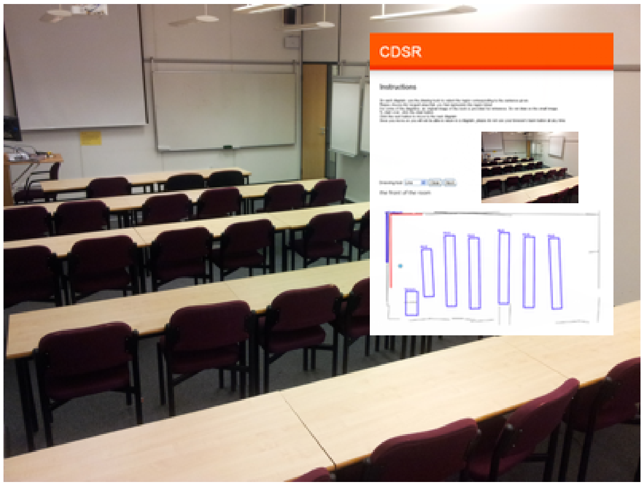


Figure 5: One of the classrooms used in our evaluation. This image was presented to subjects who were asked to annotate a copy of the image in the bottom half of Figure 4. The inset shows a screenshot from the data collection webpage.

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.

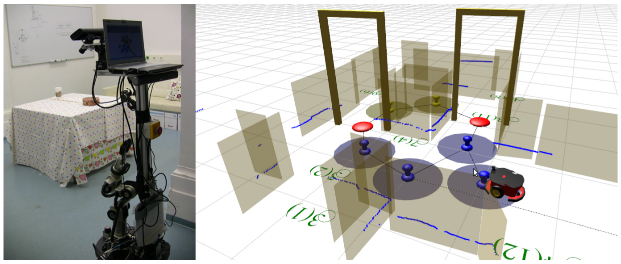


Fig. 1: Scenario and object search task at a glance, left: test scene with the robot at initial point, right: simulation/visualization of visual search task.

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 subpatterns of full pattern $\mathcal{T}^f$, and full pattern $\mathcal{T}^f$ is the superpattern of object pattern $\mathcal{T}^o$ and locality pattern $\mathcal{T}^l$.

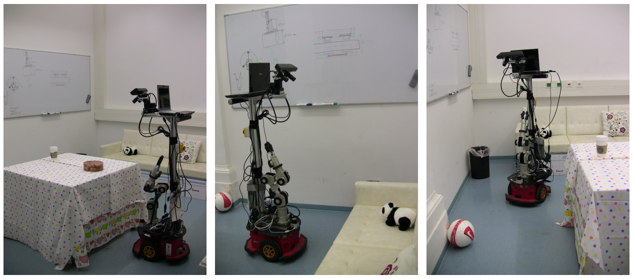
C. Pattern Retrieval

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 $t_n = S_1 \rightarrow S_h$ 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.

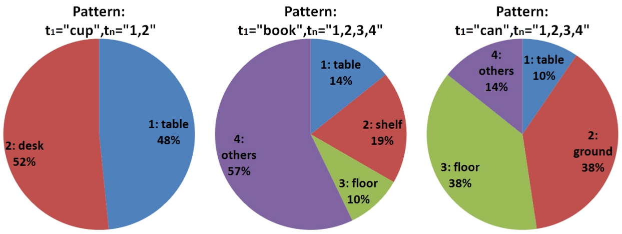


Fig. 3: The pattern retrieval results of three common objects – ”cup”, ”book” and ”can”, the source being the indoor-robot knowledge database OMICS, - only localities that exist in an office room are shown in the figure. Patterns containing ”cup”, ”book” and ”can” have absolute support values 31, 21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

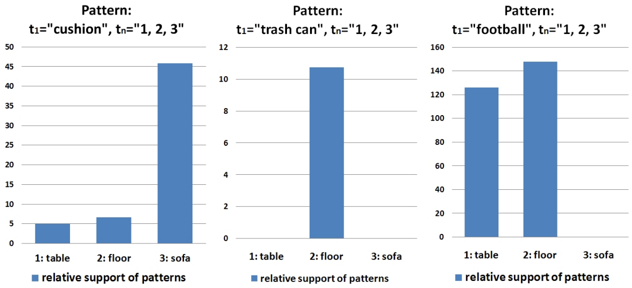


Fig. 4: The pattern retrieval results of three objects – ”cushion”, ”trash can” and ”football”, the source is the general Internet data accessed with Google advanced search, and only the localities that exist in an office room are searched for - these are displayed in the figure. Note that here we use 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.

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”.

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 ($\pm 90^\circ$) 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.

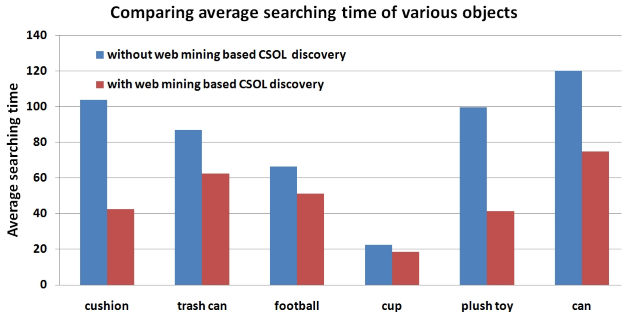


Fig. 5: Comparison of average visual search time for brute force search and the web content mining method proposed in this paper. The visual search of each object is repeated 10 times and the average processing time is recorded.

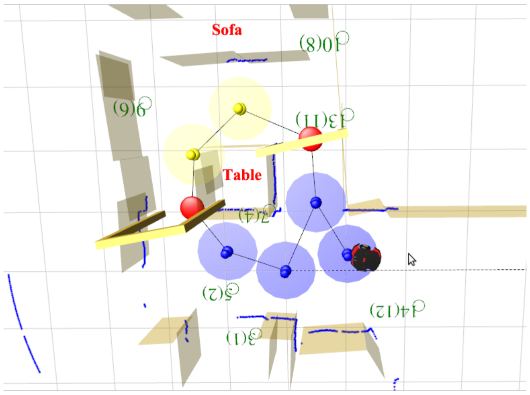


Fig. 6: Simulation/Visualization world from the top view.

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.

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*.