Deep Spatial Affordance Hierarchy: Spatial Knowledge Representation for Planning in Large-scale Environments

Andrzej Pronobis, Francesco Riccio, Rajesh P. N. Rao*

Abstract

Domain-specific state representations are a fundamental component that enables planning of robot actions in unstructured human environments. In case of mobile robots, it is the spatial knowledge that constitutes the core of the state, and directly affects the performance of the planning algorithm. Here, we propose Deep Spatial Affordance Hierarchy (DASH), a probabilistic representation of spatial knowledge, spanning multiple levels of abstraction from geometry and appearance to semantics, and leveraging a deep model of generic spatial concepts. DASH is designed to represent space from the perspective of a mobile robot executing complex behaviors in the environment, and directly encodes gaps in knowledge and spatial affordances. In this paper, we explain the principles behind DASH, and present its initial realization for a robot equipped with laser-range sensor. We demonstrate the ability of our implementation to successfully build representations of large-scale environments, and leverage the deep model of generic spatial concepts to infer latent and missing information at all abstraction levels.

1 Introduction

Many recent advancements in the fields of robotics and artificial intelligence have been driven by the ultimate goal of creating artificial agents able to perform service tasks in real environments in collaboration with humans (Aydemir et al. 2013; Hanheide et al. 2016). While significant progress have been made in the area of robot control, largely thanks to the success of deep learning (Levine et al. 2016), we are still far from solving more complex scenarios that require forming plans spanning large spatio-temporal horizons.

In such scenarios, domain-specific state representations play a crucial role in determining the capabilities of the agent and the tractability of the solution. In case of mobile robots operating in large-scale environments, it is the spatial knowledge that constitutes the core of the state. As a result, the way in which it is represented directly affects the actions the robot can plan for, the performance of the planning algorithm, and ultimately, the ability of the robot to successfully reach the goal. For complex tasks involving interaction with humans, the relevant spatial knowledge spans multiple levels of abstraction and spatial resolutions, including detailed geometry and appearance, global environment structure, and high-level semantic concepts. Representing such knowledge is a difficult task given uncertainty and partial observability governing real applications in human environments.

In this work, we propose Deep Spatial Affordance Hierarchy (DASH, ref. Fig. 1), a probabilistic representation of spatial knowledge designed to support and facilitate planning and execution of complex behaviors by a mobile robot. The representation encodes the belief about the state of the world. However, more importantly, it also provides information about spatial affordances, i.e. the possibilities of actions on objects or locations in the environment. It does so by leveraging a hierarchy of sub-representations (layers), which directly correspond to a hierarchical decomposition of the planning problem. The layers represent multiple spatial knowledge abstractions (from geometry and appearance to semantic concepts), using different spatial resolutions (from voxels to places), frames of reference (allo- or ego-centric), and spatial scopes (from local to global). The goal is to represent spatial knowledge in a way that directly corresponds to how it will be utilized by the robot and its planning algorithm.

DASH includes both instance knowledge about the specific robot environment as well as default knowledge about generic human environments. The latter is modeled using a recently proposed Deep Generative Spatial Model (DGSM) (Pronobis and Rao 2017). Specifically, DGSM leverages recent developments in deep learning, providing fully probabilistic, generative model of spatial concepts learned directly from raw sensory data. DGSM unifies the layers of our representation, enabling upwards and downwards inferences about spatial concepts defined at different levels of abstraction. Finally, DASH is designed to explicitly represent and fill gaps in spatial knowledge due to uncertainty, unknown concepts, missing observations or unexplored space. This brings the possibility of using the representation in open-world scenarios, involving active exploration and learning.

*A.Pronobis and R. Rao are with Computer Science & Engineering, University of Washington, Seattle, WA, USA. A. Pronobis is also with Robotics, Perception and Learning Lab, KTH Royal Institute of Technology, Stockholm, Sweden. F. Riccio is with Dept. of Computer, Control, and Management Engineering, Sapienza University of Rome, Rome, Italy. {pronobis,rao}@cs.washington.edu, riccio@diag.uniroma1.it. This work was supported by the Swedish Research Council (VR) project SKAEENet.

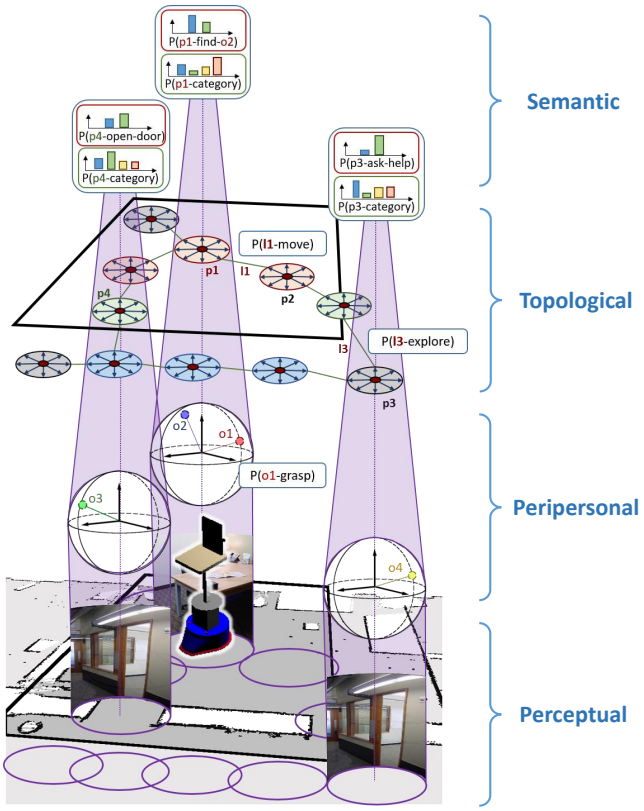


Fig. 1: The multi-layered architecture of Deep Spatial Affordance Hierarchy. The *perceptual layer* integrates perceptual information from the robot sensors. The *peripersonal layer* represents object and landmark information and affordances in the space immediately surrounding the robot. The *topological layer* encodes global topology and coarse geometry and navigation action affordances. Finally, the *semantic layer* relates the internal instance knowledge to human semantic concepts. The four layers are connected by the probabilistic *deep default knowledge model* (shaded purple columns), which provides definitions of generic spatial concepts and their relations across all levels of abstraction.

In this paper, we describe the general architecture of DASH and present an initial realization of the representation for a mobile robot equipped with a laser range sensor. We perform a series of experiments demonstrating the ability of the representation to perform different types of inferences, including bottom-up inferences about semantic spatial concepts and top-down inferences about geometry of the environment. We then showcase its ability to build semantic representations of large-scale environments (e.g. floors of an office building).

We begin the presentation of DASH with a description of the scenario, an analysis of roles and desired properties of a spatial knowledge representation (Sec. 2). Then, we describe the architecture of DASH (Sec. 3), present its initial realization (Sec. 4) and describe the details of the deep generative model of default spatial knowledge (Sec. 5). We follow with the experimental evaluation in Sec. 6.

2 Analysis of the Problem

We recognize that the ultimate purpose of a spatial knowledge representation for an autonomous mobile robot is to enable and facilitate successful planning and execution of actions in the robot environment. Here, we focus specifically on scenarios involving large-scale, dynamic, human environments, such as office buildings, homes, and hospitals. We assume that a mobile robot is physically capable of sensing the environment using on-board sensors. The sensors are likely to have limited field of view, and might be attached to actuators, such as pan-tilt units. Furthermore, the robot is capable of moving around the environment and performing basic manipulation tasks (e.g. grasping objects or pushing buttons). Finally, we assume that the robot can interact and collaborate with humans in order to accomplish its tasks (e.g. by asking for additional information or requesting help when a task cannot be accomplished by the robot itself). We follow with an analysis of roles of a spatial knowledge representation in the context of the considered scenarios as well as a discussion of its desired properties.

Role of a Spatial Knowledge Representation

Referring to the discussion of roles of a knowledge representation in (Davis, Shrobe, and Szolovits 1993), and a more specific analysis for spatial knowledge in (Pronobis et al. 2010b), we formulate a set of roles of a domain-specific spatial knowledge representation for a mobile robot. Such a representation can be seen as:

a) A substitution (surrogate) for the world that allows the robot to reason about actions involving parts of the environment beyond its sensory horizon. The surrogate can either represent the belief about the state of the world (what the world looks like), or more directly, the belief about affordances (what the robot can do at a specific place or involving a specific spatial entity). It is important to note that it is inherently imperfect, i.e. it is incomplete (some aspects of the world are not represented), inaccurate (captured with uncertainty), and likely to become invalid (e.g. due to dynamics of the world).

b) A set of commitments that determine the terms in which the robot thinks about space. The representation defines which aspects of the world are relevant, and specifies the formalism used to represent and relate them. To this end, it defines the levels of abstraction at which spatial entities exist, spatial frames of reference used to relate them (absolute or relative, allo- or ego-centric) as well as their persistence. It is worth noting that these commitments significantly affect the ability of the robot to plan and execute specific actions. Furthermore, the representation does not have to be more expressive than required to successfully act. Therefore, we can think of the commitments in the representation as defining part of the action space of the robot.

c) A set of definitions that determine the reasoning that can be (and that should be) performed within the framework. This includes reasoning about the location of the robot with respect to the internal frames of reference (whether metric, topological or semantic), inferring more abstract concepts from observations (e.g. affordances, semantic descriptions),

or generating missing lower-level information from high-level descriptions (e.g. expected position of occluded objects in rooms of known functional category).

d) A medium of communication between the robot and humans. In scenarios involving human-robot collaboration, spatial knowledge provides a common ground for communication and knowledge transfer. The representation must therefore be capable of relating human spatial concepts to those internal to the robot.

e) A way of structuring the spatial information so that it is computationally feasible to perform inferences and action planning in a specified time (e.g. in real time) despite limited resources.

Desired Properties of the Representation

Having in mind the specifics of the scenario, the roles of a representation, practical limitations, and experience resulting from existing approaches and robotic systems (Thrun et al. 1998; Kuipers 2000; Marder-Eppstein et al. 2010; Hanheide et al. 2016), we identify several desired properties of a spatial knowledge representation for mobile robots.

Spatial knowledge in realistic environments is inherently uncertain and dynamic. Given the local nature of the robot's sensing, it is futile to represent the environment as accurately as possible. A very accurate representation is likely to be intractable and will require a substantial effort to be kept up-to-date. Moreover, its usability will remain constrained by robot capabilities. Hence, our primary assumption is that the representation should instead be minimal and the spatial knowledge should be represented only as accurately as it is required to support the functionality of the robot.

Planning is a computationally demanding process and its complexity increases exponentially with the size of the environment and number of considered spatial entities. However, due to the way real-world environments are structured and limitations of robot sensors and actuators, decomposing the planning problem hierarchically can greatly reduce its complexity while maintaining highly optimal results. This naturally leads to a hierarchy of higher-level, long-term, global plans involving lower-level short-term, local behaviors. In fact, hierarchical planners are used in the majority of existing robotic systems (Marder-Eppstein et al. 2010; Aydemir et al. 2013; Hanheide et al. 2016) due to their tractability. Moreover, behavioral analyses found hierarchical spatial planning in humans (Balaguer et al. 2016). In order to support such strategies, a spatial representation should perform knowledge abstraction, providing symbols corresponding to spatial phenomena of gradually increasing complexity, anchored to reference frames of increasing spatial scope and decreasing resolution. This leads to discretization of continuous space, which significantly reduces the number of states for planning (Hawes et al. 2009) and provides a basis for higher-level conceptualization (Zender et al. 2008).

Due to the dynamic properties of the real world, abstracted knowledge is more likely to remain valid over time. At the same time, high-resolution up-to-date spatial information is required for executing actions in the robot peripersonal space. Yet, it can also be re-acquired through perception. Therefore, the representation should correlate the levels of abstraction with the persistence of information, employing local working-memory representations for integrating high-resolution spatial information (visual servoing being the extreme example). In other words, the robot should use the world as an accurate representation whenever possible.

Representing uncertainty in the belief state is crucial for the robot to make informed decisions in the real-world, including planning for epistemic actions and anticipating future uncertainty. In this context, decision-theoretic planning algorithms rely on probabilistic representations of uncertainty, therefore, it is desirable for a knowledge representation to also be probabilistic in nature.

Furthermore, a representation should not only represent what is known about the world, but also what is unknown. This includes explicit representation of missing evidence (e.g. due to occlusions), unexplored space (e.g. exploration frontiers) or unknown concepts (e.g. unknown object categories). Representing knowledge gaps can be exploited to address the open-world problem (in the continual planning paradigm (Hanheide et al. 2016)), trade exploration vs exploitation, or drive learning.

3 Deep Spatial Affordance Hierarchy (DASH)

As a result of the problem analysis, we propose Deep Spatial Affordance Hierarchy (DASH). A general overview of the architecture of the representation is shown in Fig. 1. DASH represents the robot environment using four sub-representations (layers) focusing on different aspects of the world, encoding knowledge at different levels of abstraction and spatial resolutions as well as in different frames of reference of different spatial scope. The characteristics of the layers were chosen to simultaneously support both action planning and spatial understanding for the purpose of localization and human-robot interaction. In particular, the former objective is realized by directly representing spatial affordances, which we define as the possibilities of actions on objects or locations in the environment relative to the capabilities and state of the robot. The characteristics of the layers are summarized in Table 1.

DASH is organized as a hierarchy of spatial concepts, with higher-level layers providing a coarse, global representation comprised of more abstract symbols, and lower-level layers providing a more fine-grained representation of parts of the environment anchored to the higher-level entities. The layers are connected by a crucial component of the representation, the probabilistic *deep default knowledge model*, which provides definitions of generic spatial concepts and their relations across all levels of abstraction.

The hierarchy directly relates to a similar, hierarchical decomposition of the planning problem. A global planner can derive a navigation plan relying only on the top layers for representing its beliefs, a local planner can be used to plan specific manipulation actions using intermediate layers, with a controller realizing them base on knowledge in the lowest-level representation. Below, we provide details about each component of the representation.

	Perceptual	Peripersonal	Topological	Semantic
World Aspects Captured	Detailed geometry and appearance	Object/landmark info, coarse local geometry	Large-scale topology, coarse global geometry	Human semantic descriptions
Reference Frame	Metric (allo-centric, sliding window)	Collection of: Metric (epi-centric)	Topological (allo-centric) Metric (allo-centric)	Relational
Spatial Scope	Sensory horizon	Local	Global	Global
Spatial Entities	Voxels	Objects/landmarks	Places, paths, views	Relations to human concepts
Affordances	—	Manipulation and epistempic actions	Navigation and epistemic actions	Human interaction actions Tasks involving human concepts
Robot Pose	Center of the window	Relative to objects/landmarks	Place/view ID	Described semantically
Knowledge Gaps	Missing observations	Missing evidence Unknown objects	Unexplored space Unknown places	Novel semantic concepts

Table 1: Characteristics of the four layers of DASH.

Perceptual Layer

At the bottom level of the representation is the *perceptual layer*. The layer maintains an accurate representation of the geometry and appearance of the local environment obtained by short-term spatio-temporal integration of perceptual information from (possibly multiple and directional) sensors with finite horizon. Spatial information in perceptual layer is represented in an allo-centric metric reference frame, which facilitates integration of perception from multiple viewpoints and sensors. However, the representation is always centered at the current location of the robot, and spans a radius roughly corresponding to the maximum range of the robot sensors (essentially a sliding window). Information outside the spatial scope is forgotten, which makes the layer akin to a working memory, and enables consistent large-scale higher-level representations without the need to maintain low-level global consistency. The layer provides a more complete input for further abstractions with reduced occlusions and noise. It enables tracking of the relative movements of the robot, and forms a basis for deriving low-level control laws for manipulation and obstacle avoidance. Missing observations (e.g. due to unresolved occlusions) are explicitly represented.

Peripersonal Layer

Above the perceptual layer is the *peripersonal layer*, which captures spatial information related to object and landmark instances from the perspective of an agent performing actions at different locations in the environment. To support planning, the layer represents object affordances related to actions that can be performed directly by the robot. This includes manipulation (e.g. possibility of reaching/grasping an object or pressing a button), interaction in relation to objects (e.g. possibility of pointing at an object), and epistemic affordances (e.g. possibility of observing an object). Furthermore, the layer captures object and landmark descriptors that are internal to the robot as well as spatial relations between objects and landmarks in relation to the robot (and therefore coarse local geometry). Finally, it serves as an intermediate layer of the deep default knowledge model, used to generate descriptions of locations in terms of higher-level concepts (e.g. room categories or place affordances).

To reflect the local and robo-centric nature of the captured information, the peripersonal layer relies on a collection of ego-centric, metric reference frames, each focusing on the space immediately surrounding the robot at a different location in the environment (see Fig. 1). The spatial scope of each of the reference frames is defined primarily by the peripersonal space of the robot, within which objects can be grasped and manipulated. However, to support epistemic affordances, interaction about objects, and higher-level conceptualization, the scope can be extended to include context in the form of knowledge about objects that directly relates to the functionality of the location. For instance, a reference frame centered in front of a desk might include information about shelves and books in the room, even beyond the reach of the robot. While recent results from neuropsychology suggest existence of local, body-centered representations in animals and humans (Holmes and Spence 2004), our motivation for such decomposition is primarily the efficiency of the planning problem.

The peripersonal layer explicitly represents gaps in knowledge about the local space due to missing evidence (e.g. resulting from occlusions) and unknown objects. The latter occurs when the default knowledge model is not familiar with an object, and cannot produce a certain object descriptor or affordance information.

Topological Layer

The topological layer provides an efficient representation of large-scale space, including coarse geometry and topology, and serves several key roles in DASH. First, it provides a way to express the global pose of the robot. Second, it captures navigation and exploration action affordances associated with locations in the environment. Third, it is a global counterpart to the local peripersonal representations and anchors them in the large-scale space. Finally, it captures internal descriptors of places and serves as an intermediate layer

of the deep default knowledge model used to derive semantic place descriptions.

To this end, the layer performs a bottom-up discretization of continuous space into a set of locations called *places*. Places correspond to locations in the environment previously visited by the robot, and are meant to represent space at a resolution sufficient for action execution, while maintaining efficiency and robustness to dynamic changes. In other words, the resolution is selected to ensure that high-level navigation can be planned using the topological layer only, with local behaviors planned using the knowledge in the peripersonal layer at the destination. Places are spatially related to other, neighboring places, which encodes coarse global geometry of the environment and allows for path integration.

For each place, the topological layer maintains a set of discrete headings, called *views*. Together with places, views can be used to efficiently represent the complete global pose of the robot. Moreover, views and places are used to anchor knowledge in the representation. First, the topological layer captures robot-internal descriptors of each view and place. The descriptors are derived from lower-level representations using the deep default knowledge model and serve as an intermediate layer of the model. Second, each visited place anchors a peripersonal representation describing the place in more detail.

Besides places and views, the layer also defines *paths* connecting neighboring places into a topological graph. The semantics of a path between two places is the possibility of navigating directly from one place to the other. Thus, essentially, paths represent navigation place affordances, which can be associated with probability indicating uncertainty estimated based on the current, detailed information in the peripersonal layer (e.g. based on visible obstacles). Furthermore, the topological nature of the graph of places and paths, enables planning of complex navigational tasks, such as involving elevators. The place in the elevator might afford navigating to places on different floors, depending on the information captured in the peripersonal layer (e.g. displayed floor number) or additional state information.

Existence of a path in the graph does not necessarily imply that it has previously been traveled by the robot. In fact, a path can indicate the possibility of navigating towards unexplored space. To this end, the topological layer utilizes the concept of *placeholders* (Pronobis et al. 2010b), which can be seen as candidate places, and are used to explicitly represent unexplored space. As a result, paths that lead to placeholders express the possibility of epistemic exploration actions. This can be used to address the open world problem, for instance, in the continual planning paradigm (Hanheide et al. 2016).

Semantic Layer

On top of DASH is the semantic layer, a probabilistic relational representation relating the spatial entities in the other layers to human semantic spatial concepts defined in the deep default knowledge model. This includes such concepts as object categories and attributes, place attributes, room categories, or the concept of a room itself. It is the semantic layer that captures the knowledge that an object is likely to be a cup, or that certain places are likely to be located in a kitchen. Furthermore, the layer plays an important role in planning complex tasks, by representing place affordances related to human interaction as well as actions characterized in terms of human concepts. For instance, it is the sensory layer that defines the affordance expressing the possibility of asking a person for help with making coffee or the possibility of finding a cup at a certain place. Finally, the layer enables transfer of knowledge from humans to the robot (e.g. capturing object category information provided by the user). Such knowledge can be utilized by the default knowledge model to generate lower-level information stored in other layers.

Deep Default Knowledge

The four layers representing knowledge about the specific robot environment are linked by the *deep default knowledge model*. The model provides definitions of generic spatial concepts, valid for typical human environments, and their relations across all levels of abstraction (from sensory input to high-level concepts). This includes robot-internal models of objects in terms of low-level perception, places in terms of objects, place and object affordances, or models of semantic categories and attributes of objects and places. In other words, the four layers can be seen as defining the traditional ABox of our spatial knowledge base, while the deep default knowledge model represents its TBox.

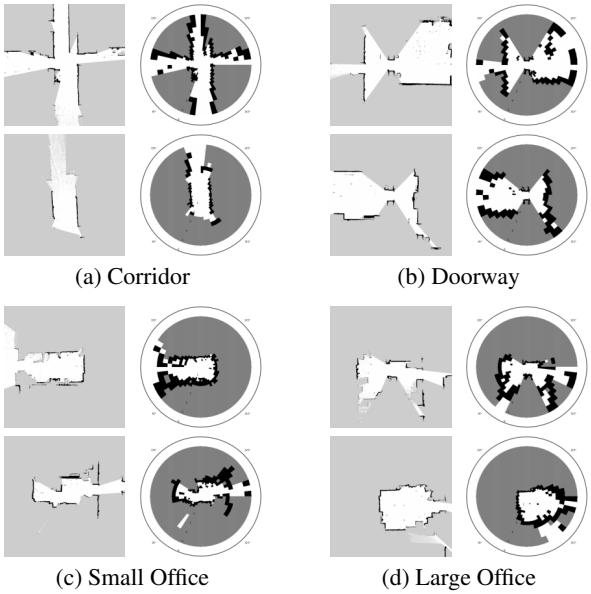
The role of the default knowledge model is to permit inferences about missing or latent aspects of the environment in each layer, based on the knowledge available in other layers. This includes bottom-up inferences (e.g. about semantic descriptions based on perception) and top-down inferences (e.g. about object presence or place affordances based on semantic descriptions). The resulting knowledge base constitutes a more complete (albeit uncertain) belief state for the planner. In this work, we implement this component using a deep generative probabilistic model based on Sum-Product Networks (see Sec. 5).

4 Realization of DASH for Laser-Range Data

In order to evaluate the architecture of DASH in practice, we provide its initial realization for a mobile robot equipped with a laser-range sensor. We utilize laser-range data to simplify the initial implementation, however the proposed algorithms can be easily extended to include 3D and visual information.

Perceptual Layer

To integrate local laser-range observations in the perceptual layer, we use a common occupancy grid representation. Specifically, we utilized a grid mapping approach based on Rao-Blackwellized particle filters (Grisetti, Stachniss, and Burgard 2007). We crop the resulting grid map to only retain a rectangular fragment of size 10x10m, centered at the current position of the robot. Consequently, we do not require global consistency of the grid map, as long as the local



(a) Corridor (b) Doorway

(c) Small Office (d) Large Office

Fig. 2: Visualization of spatial knowledge represented in the peripersonal layer for sample places of different semantic categories, expressed as both Cartesian and polar occupancy grids.

environment is mapped correctly. This will still result in partial maps (especially when the robot enters a new room), but it will help to accumulate observations over time. During our experiments, the robot was exploring the environment driving with a constant speed, while continuously gathering data and performing inferences based on the current state of the perceptual layer.

Peripersonal Layer

The peripersonal representation for each place is constructed from the current local occupancy grid in the perceptual layer. However, since the scope of the peripersonal representation is limited to the space immediately surrounding the robot and relevant context, we only retain information about the parts of the environment visible from the robot (grid cells that can be raytraced from the robot location). As a result, walls occlude the view and the resulting grid will mostly contain objects present in a single room. In order to include a more complete appearance of the objects, we additionally include observations behind small obstacles, and a small vicinity around every occupied cell visible from the robot (e.g. corners of furniture). Examples of such local occupancy grids can be seen in Fig. 2.

Next, every local grid map is transformed into an egocentric polar representation (compare polar and Cartesian grids in Fig. 2). This encodes high-resolution information about the geometry and objects nearby, and complements it with less-detailed context further away from the robot. Encoding spatial knowledge closer to the robot in more detail is important for understanding the semantics of the exact robot location (for instance when the robot is in a doorway). However, it also relates to how spatial information is used by a robot when planning and executing actions. It is in the vicinity of the robot that higher accuracy of spatial information is required. The polar grids in our implementation assumed radius of $5m$, with angle step of 6.4 degrees and resolution decreasing with the distance from the robot. It is worth noting that lack of evidence resulting from occlusions is explicitly represented in the cells of the polar representation. Such representation of peripersonal layer is clearly a simplification, however one that results from the nature of the laser-range data.

Topological Layer

The topological layer is maintained by a mapping algorithm discretizing continuous space into sets of *places*, *placeholders*, *views*, and *paths*. The goal is to generate an efficient discretization, which supports all the roles of the topological layer, including expression of the global robot pose, representation of affordances related to navigation and exploration, and anchoring of local spatial knowledge to the global space.

The mapping algorithm expands the topological layer incrementally, adding placeholders at neighboring unexplored locations, and connecting them with paths to existing places. Then, once the robot performs an exploration action associated with a specific path, a new place is generated to which a peripersonal representation, as well as place and view descriptors are anchored. At this point, the path between the two places signifies navigation affordance, and is associated with probability based on current, up-to-date information. In order to choose the location for a new placeholder, the algorithm relies upon information contained in the perceptual layer, including detailed local geometry and obstacles.

Similarly to (Chung et al. 2016), we formulate the problem of finding placeholder locations using a probability distribution that models their relevance and suitability. However, instead of sampling locations of all places in the environment at once, we incrementally add placeholders as the robot explores the environment, within the scope of the perceptual layer. Specifically, the probability distribution is modeled as a combination of two components:

$$P(E \mid G) = \frac{1}{Z}\prod_i \phi_I(E_i)\phi_N(\mathscr{E}), \tag{1}$$

where $E_i \in \{0,1\}$ determines the existence of a place at a location i in the perceptual layer, G is the perceptual occupancy grid, and $\mathscr{E}$ is a set of locations of all existing places within the scope of the perceptual representation.

The potential function ϕ_I models suitability of a specific location, and is defined in terms of three potentials calculated from G:

$$\phi_I(E_i) = \phi_O(E_i)(\phi_V(E_i) + \phi_P(E_i) - \phi_V(E_i)\phi_P(E_i)), \tag{2}$$

where:

- ϕ_O ensures that placeholders are created in areas that are safe from collisions with obstacles. It depends on the distance d_o to the nearest obstacle and is calculated similarly to the cost map used on our robot for obstacle avoidance (Marder-Eppstein et al. 2010). ϕ_o equals 0 for distance smaller than the radius r of the robot base and $1 - exp(-\alpha(d_o - r))$ otherwise.

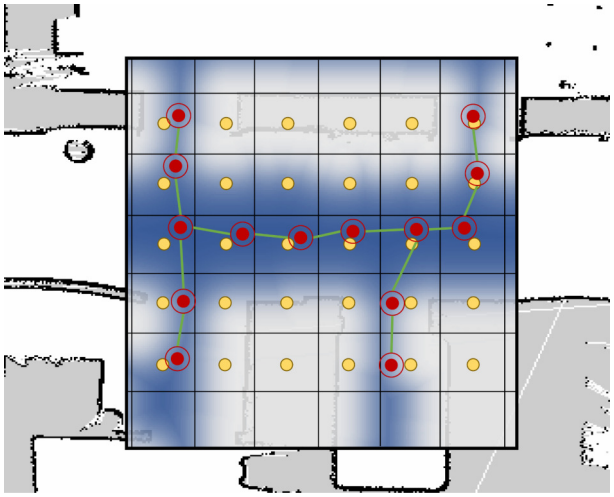


Fig. 3: Visualization of generated places and paths on top of the knowledge in the perceptual layer. The highlighted region corresponds to the spatial scope of the perceptual representation and displays the value of the potential ϕ_I. The low-resolution lattice is illustrated using yellow points, and red points indicate the final, optimized locations of places. Paths highlighted in green afford navigability throughout the environment.

- $\phi_V = exp(-\gamma d_c)$ depends on the distance d_c to the nearest node of a Voronoi graph of the 2D map. This promotes centrally located places that are often preferred for navigation.
- ϕ_P promotes places inside narrow passages (e.g. doors). The potential is generated by convolving the local map with a circular 2D filter of a radius corresponding to an average width of a door.

Overall, ϕ_I ensures that placeholders are located only in areas that are safe and preferred for navigation, and constitute useful anchors for information stored in other layers of the representation. The potential ϕ_N, models the neighborhood of a place and guarantees that places are evenly spread throughout the environment. To this end, the potential function promotes positions at a certain distance d_n from existing places:

$$\phi_N(E_i) = \sum_{p \in \mathscr{E}} e^{-\frac{(d(i,p)-d_n)^2}{2\sigma^2}},$$

where $d(i,p)$ is a Euclidean distance between the potential new place and an existing place.

Final location of new placeholders is chosen through MPE inference in $P(E \mid G)$. However, before adding a new placeholder to the map it is important to verify whether the robot will be able to navigate to it. To this end, we perform an A* search directly over the potential function, and quantify the navigability based on the accumulated potential. Only then, a *path* is created between an existing place and a placeholder. Similarly, the accumulated potential is used to quantify navigability of paths between existing places.

In order to incorporate knowledge about coarse global geometry into the topological representation, we further relate placeholders and places to a global low-resolution lattice (0.8m distance between points in our experiments), as illustrated in Fig. 3. As the robot moves through the environment, the lattice is extended, while preserving consistency with existing points. We assume that a place must be associated with a point of the lattice, and each lattice point can be associated with only one place. As a result, when performing MPE inference using $P(E \mid G)$, we assume that only one place might exist in a cell of a Voronoi tessellation established by the points of the lattice. The resulting set of placeholders (and eventually places) will uniquely correspond to lattice points, yet be created only in locations which are suitable, and can serve as navigation goals for the lower-level controller.

For each place that is created from a placeholder, we generate a set of eight *views*. The views are a discrete representation of the heading of the robot when located at a place, and are assumed to be vectors pointing from a point of the lattice to the eight immediately neighboring points. Since, places are associated uniquely with lattice points, each view will naturally point in the direction of only one neighboring place. As a result, each *path* connecting a place to another place or placeholder will be associated with a specific view.

Semantic Layer

In our initial implementation, the semantic layer captures the information about semantic categories of places in the topological map. This includes categories of rooms in which places are located, such as an office or a corridor, but also a functional place category corresponding to places located in a doorway. The layer is implemented as a simple relational data structure assigning place instances to semantic categories in the ontology of the deep default knowledge model. Each such relation is associated with probability value. Additionally, for each place, the layer captures the likelihood of the peripersonal representation of the place being observed for any of the semantic categories. That likelihood is used to detect and explicitly represent that a place belongs to a novel category, i.e. one that is not recognized by the default knowledge model.

5 Representing Default Knowledge

In our implementation, default knowledge is modeled using a recently proposed Deep Generative Spatial Model (DGSM) (Pronobis and Rao 2017), a probabilistic deep model which learns a joint distribution over spatial knowledge represented at multiple levels of abstraction. We apply the deep model to capture generic spatial concepts and relations between knowledge represented in peripersonal, topological, and semantic layers. Once learned, it enables a wide range of probabilistic inferences. First, based on the knowledge in the peripersonal layer, it can infer descriptors of views and places, as well as semantic categories of places. Moreover, it can detect that a place belongs to a novel category, not known during training. Inference can also be performed over the contents of the peripersonal representation. The model can infer missing geometry information resulting

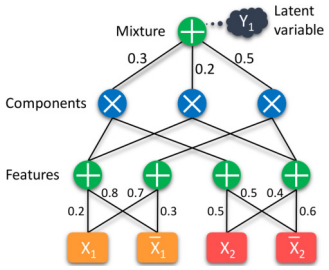


Fig. 4: An SPN for a naive Bayes mixture model $P(X_1, X_2)$, with three components over two binary variables. The bottom layer consists of indicators for each of the two variables. Weights are attached to inputs of sums. Y_1 represents a latent variable marginalized out by the top sum node.

from partial observations and generate prototypical peripersonal representations based on semantic information.

To this end, DGSM leverages Sum-Product Networks (SPNs), a novel probabilistic deep architecture (Poon and Domingos 2011; Peharz et al. 2015), and a unique structure matching the hierarchy of representations in DASH. Below, we give a primer on Sum-Product Networks and describe the details of the architecture of the DGSM model.

Sum-Product Networks

Sum-product networks are a recently proposed probabilistic deep architecture with several appealing properties and solid theoretical foundations (Peharz et al. 2015; Poon and Domingos 2011; Gens and Domingos 2012). One of the primary limitations of probabilistic graphical models is the complexity of their partition function, often requiring complex approximate inference in the presence of non-convex likelihood functions. In contrast, SPNs represent probability distributions with partition functions that are guaranteed to be tractable, involve a polynomial number of sums and product operations, permitting exact inference. While not all probability distributions can be encoded by polynomial-sized SPNs, recent experiments in several domains show that the class of distributions modeled by SPNs is sufficient for many real-world problems, offering real-time efficiency.

SPNs model a joint or conditional probability distribution and can be learned both generatively (Poon and Domingos 2011) and discriminatively (Gens and Domingos 2012) using Expectation Maximization (EM) or gradient descent. They are a deep, hierarchical representation, capable of representing context-specific independence. As shown in Fig. 4 on a simple example of a naive Bayes mixture model, the network is a generalized directed acyclic graph of alternating layers of weighted sum and product nodes. The sum nodes can be seen as mixture models, over components defined using product nodes, with weights of each sum representing mixture priors. The latent variables of such mixtures can be made explicit and their values inferred. This technique is often used for classification models where the root sum is a mixture of sub-SPNs representing multiple classes. The bottom layers effectively define features reacting to certain values of indicators for the input variables. Not all possible architectures consisting of sums and products will result in a valid probability distribution. However, following simple constraints on the structure of an SPN will guarantee validity (see (Poon and Domingos 2011; Peharz et al. 2015) for details).

Inference in SPNs is accomplished by an upward pass through the network. Once the indicators are set to represent the evidence, the upward pass will yield the probability of the evidence as the value of the root node. Partial evidence (or missing data) can easily be expressed by setting all indicators for a variable to 1. Moreover, it can be shown (Poon and Domingos 2011) that MPE inference can be performed by replacing all sum nodes with max nodes, while retaining the weights. Then, the indicators of the variables for which the MPE state is inferred are all set to 1 and a standard upward pass is performed. A downward pass then follows which recursively selects the highest valued child of each sum (max) node, and all children of a product node. The indicators selected by this process indicate the MPE state of the variables.

In this work, we learn the SPN using hard EM, which was shown to work well for generative learning (Poon and Domingos 2011) and overcomes the diminishing gradient problem. The reader is referred to (Pronobis and Rao 2017) for details about the learning procedure.

Architecture of DGSM

The architecture of DGSM is based on a generative SPN illustrated in Fig. 5. The model learns a probability distribution $P(C, D^P_1, \ldots, D^P_{N_p}, D^{V_1}_1, \ldots, D^{V_8}_{N_v}, X_1, \ldots, X_{N_x})$, where C represents the semantic category of a place, $D^P_1, \ldots, D^P_{N_p}$ constitute an internal descriptor of the place, $D^{V_1}_1, \ldots, D^{V_8}_{N_v}$ are descriptors of eight views, and $X_1, \ldots, X_C$ are input variables representing the occupancy in each cell of the polar grid of the peripersonal layer. Each occupancy cell is represented by three indicators in the SPN (for empty, occupied and unknown space). These indicators constitute the bottom of the network (orange nodes).

The structure of the model is partially static and partially generated randomly according to the algorithm described in (Pronobis and Rao 2017). The resulting model is a single SPN, which is assembled from three levels of sub-SPNs. First, we begin by splitting the polar grid of the peripersonal layer equally into eight 45 degree parts, corresponding to the *views* defined in the topological layer. For each view, we randomly generate a sub-SPN over the subset of X_i representing the occupancy within the view, as well as latent variables $D^{V_i}_1, \ldots, D^{V_i}_{N_v}$ serving as an internal view descriptor. The sub-SPN can be seen as a mixture model consisting of 14 components in our implementation. In the second level, we use the distributions defining the components from each view ($8 * 14$ in total) as inputs, and generate random SPNs representing each of the semantic place classes in the ontology. Each of such SPNs is itself a mixture model with the latent variable D^P_i being part of the place descriptor. Finally, in the third level, the sub-SPNs for place classes are combined by a sum node (mixture) forming the root of the whole network. The latent variable associated with the root node is

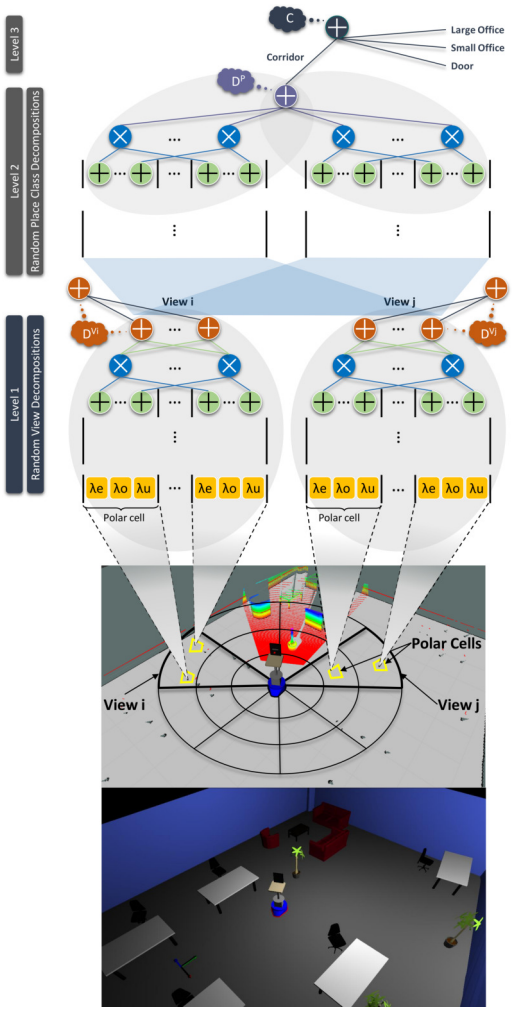


Fig. 5: The structure of the SPN implementing our spatial model. The bottom images illustrate a robot in an environment and a robocentric polar grid formed around the robot. The SPN is built on top of the variables representing the occupancy in the polar grid.

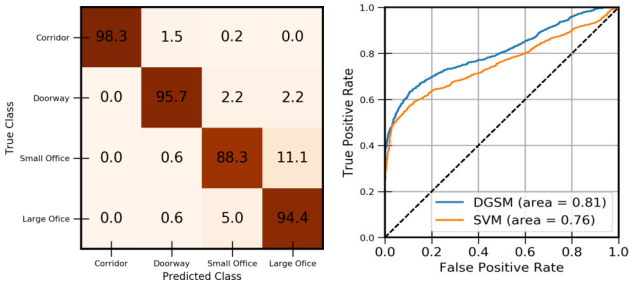
C and is set to the appropriate class label during learning. Overall, such decomposition allows us to use networks of different complexity for representing lower-level features of each view and for modeling the top composition of views into place classes.

6 Experimental Evaluation

Our experimental evaluation consists of two parts. First, we evaluated the ability of the deep default knowledge model implemented with DGSM to perform both top-down and bottom-up inferences across the layers of the representation. Then, we deployed our complete implementation of DASH in order to build representations of large-scale environments.

Experimental Setup

Our experiments were performed on laser range data from the COLD-Stockholm database (Pronobis and Jensfelt



(a) (b)

Fig. 6: Results of experiments with bottom-up inference: (a) normalized confusion matrices for semantic place categorization; (b) ROC curves for novelty detection (inliers are considered positive, while novel samples are negative).

2012). The database contains multiple data sequences captured using a mobile robot navigating with constant speed through four different floors of an office building. On each floor, the robot navigates through rooms of different semantic categories. Four of the room categories contain multiple room instances, evenly distributed across floors. There are 9 different *large offices*, 8 different *small offices*, 4 long *corridors* (1 per floor, with varying appearance in different parts), and multiple examples of observations captured when the robot was moving through *doorways*. The dataset features several other room categories: an elevator, a living room, a meeting room, a large meeting room, and a kitchen. However, with only one or two room instances in each. Therefore, we decided to use the four categories with multiple room instances for the majority of the experiments and designated the remaining classes as novel when testing novelty detection.

To ensure variability between the training and testing sets, we split the samples from the four room categories four times, each time training the model on samples from three floors and leaving one floor out for testing. The presented results are averaged over the four splits.

Bottom-up Inference

First, we evaluated the ability of DGSM to infer semantic place categories given information in the peripersonal layer. As a comparison, we used a well-established model based on an SVM and geometric features (Mozos, Stachniss, and Burgard 2005; Pronobis et al. 2010a). The features were extracted from laser scans raytraced in the same local Cartesian grid maps used to form polar grids of the peripersonal layer. We raytraced the scans in high-resolution maps (2cm/pixel), to obtain 362 beams around the robot. To ensure the best SVM result, we used an RBF kernel and selected the kernel and learning parameters directly on the test sets.

The models were trained with peripersonal representations obtained for locations on three floors in places belonging to four place categories, and evaluated on the fourth floor or using data from rooms designated as novel. The classification rate averaged over all classes (giving equal importance

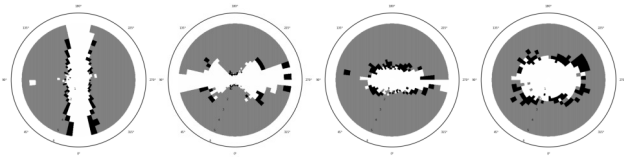
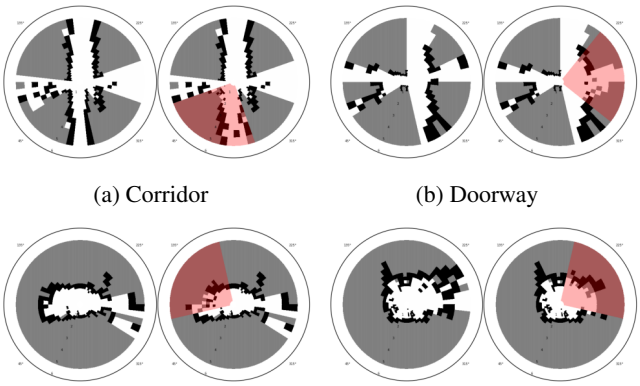


Fig. 7: Prototypical peripersonal representations inferred from semantic place category.



(a) Corridor (b) Doorway

(c) Small Office (d) Large Office

Fig. 8: Examples of completions of peripersonal representations with missing data grouped by true semantic category.

to each class) and data splits was $85.9\% \pm 5.4$ for SVM and $92.7\% \pm 6.2$ for DGSM, with DGSM outperforming SVM for every split. The normalized confusion matrix for DGSM is shown in Fig. 6(a). Most of the confusion exists between the small and large office classes. Offices in the dataset often have complex geometry that varies greatly between room instances.

Additionally, we evaluated the quality of the uncertainty measure produced by DGSM and its applicability to detecting novel concepts. To this end, we thresholded the likelihood of the test peripersonal representations produced by DGSM to decide whether the robot is located in a place belonging to a class known during training. We compared to a one-class SVM with an RBF kernel trained on the geometric features. The cumulative ROC curve for the novelty detection experiments over all data splits is shown in Fig. 6(b). We see that DGSM offers a significantly more reliable novelty signal, with AUC of 0.81 compared to 0.76 for SVM.

Top-down Inference

In the second experiment, we used DGSM to perform inference in the opposite direction, and infer values of cells in the peripersonal representation. First, we inferred complete, prototypical peripersonal representations of places knowing only place semantic categories. The generated polar occupancy grids are shown in in Fig. 7a-d. We can compare the plots to the true examples depicted in Fig. 2. We can see that each polar grid is very characteristic of the class from which it was generated. The corridor is an elongated structure with walls on either side, and the doorway is depicted as a narrow structure with empty space on both sides. Despite the fact that, as shown in Fig. 2, large variability exists between the instances of offices within the same category, the generated observations of small and large offices clearly indicate a distinctive size and shape.

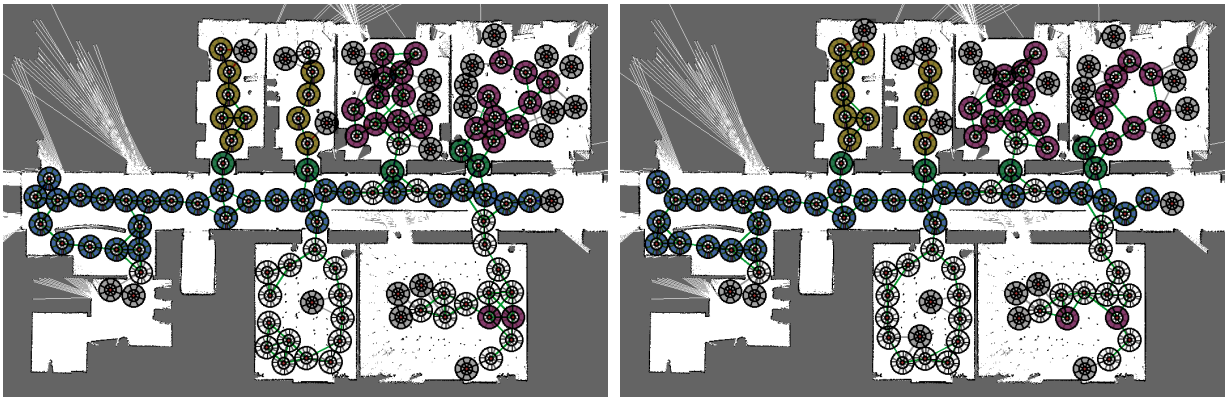
Then, we used DGSM to generate missing values in partial observations of places. To this end, we masked a random 90-degree view in each test polar grid (25% of the grid cells). All indicators for the masked polar cells were set to 1 to indicate missing evidence and MPE inference followed. Fig. 8 shows examples of peripersonal representations filled with predicted information to replace the missing values. Overall, when averaged over all test examples and data splits, DGSM correctly reconstructed $77.14\% \pm 1.04$ of masked cells. This demonstrates its generative potential.

Representing Large-Scale Space

In our final experiment, we deployed the complete implementation of DASH and evaluated its ability to build comprehensive, multi-layered representations of large-scale space. Specifically, we tasked it with representing the 5-*th* and 7-*th* floor of the office building in the COLD-dataset, which measure respectively 298 and 435 square meters. In each case, we incrementally built the representation based on the sensory data captured as the robot navigated through the environment. We relied on the perceptual layer to perform low-level integration of observed laser scans, on peripersonal layer to capture local place information, the topological layer to maintain a consistent topological graph expressing navigability and knowledge gaps related to unexplored space, and finally on the semantic layer to encode information about semantic categories of places, including detections of novel semantic categories.

Fig. 9 illustrates the state of the representation after two completed runs over the 5-*th* floor. The figure presents the final topological graph of places visited by the robot, paths expressing navigability between them, as well as paths leading to placeholders representing possibility of further exploration. For each place, we use color to illustrate the inferred semantic category, or detection of a novel category. First, we can observe that places are evenly distributed across the environment and exist in locations which are relevant for navigation or significant due to their semantics (e.g. in doorways). Moreover, the graphs created during different runs are similar and largely consistent. Second, the semantic place categories inferred by DGSM agree with the ground truth when the category of the place was recognized as known. To detect novel classes, we again thresholded the estimates of the likelihood of the peripersonal representations provided by DGSM. On the 5-*th* floor, the novel category was "meeting room" and two meeting rooms are shown in the bottom part of the map. Although both false positives and false negatives exist, places in both meeting rooms are largely correctly classified as belonging to novel categories.

Fig. 10 shows results for a different environment, the 7-*th* floor. Similar observations can be made as for the 5-*th* floor. However, here the novelty detection is less accurate. DGSM correctly detects the places in the elevator as novel (marked



(a) Run #1 (b) Run #2

Fig. 9: Contents of the topological and semantic layers after two different runs over 5-*th* floor. Gray nodes represent placeholders, while blank nodes indicate places detected as belonging to novel categories. Colors indicate recognized semantic place categories: blue for a corridor, green for a doorway, yellow for a small office, and magenta for a large office. The two large bottom rooms belong to a novel category: "meeting room".

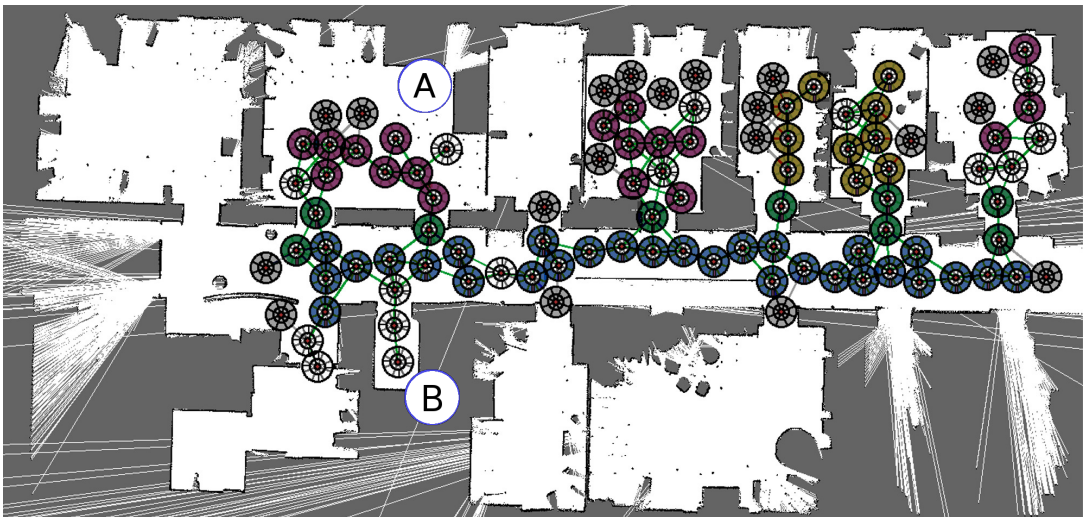


Fig. 10: Contents of the topological and semantic layers after a single run over the 7-*th* floor. Gray nodes represent placeholders, while blank nodes indicate places detected as belonging to novel categories. Colors indicate recognized semantic place categories: blue for a corridor, green for a doorway, yellow for a small office, and magenta for a large office. The rooms marked with letters A and B belong to novel categories: "living-room" and "elevator".

with "B" in the figure), but fails to detect novelty in the living room ("A" in the figure), which instead is misclassified as "large office". While not a desirable outcome, it is not surprising, given the similarity between the living room and large offices in the dataset when observed solely using laser range sensors.

7 Conclusions and Future Work

This paper presented Deep Spatial Affordance Hierarchy, a representation of spatial knowledge, designed specifically to represent the belief about the state of the world and spatial affordances for a planning algorithm on a mobile robot. We demonstrated that an implementation following the principles of DASH can successfully learn general spatial concepts at multiple levels of abstraction, and utilize them to obtain a complete and comprehensive model of the robot environment, even for a relatively simple sensory input. The natural direction for future work is to extend our implementation to include more complex perceptions provided by visual and depth sensors. Additionally, we intend to train the deep model of default knowledge to directly predict complex place affordances related to human-robot interaction. Finally, we are working to integrate our implementation of DASH with a deep hierarchical planning approach to evaluate its capacity to support autonomous robot behavior in complex realistic scenarios.

References

Aydemir, A.; Pronobis, A.; Gbelbecker, M.; and Jensfelt, P. 2013. Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics* 29(4):986–1002.

Balaguer, J.; Spiers, H.; Hassabis, D.; and Summerfield, C. 2016. Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron* 90(4):893 – 903.

Chung, M. J.-Y.; Pronobis, A.; Cakmak, M.; Fox, D.; and Rao, R. P. N. 2016. Autonomous question answering with mobile robots in human-populated environments. In *Proc. of IROS*.

Davis, R.; Shrobe, H.; and Szolovits, P. 1993. What is a knowledge representation. *AI Magazine* 14(1).

Gens, R., and Domingos, P. 2012. Discriminative learning of sum-product networks. In *Proc. of NIPS*.

Grisetti, G.; Stachniss, C.; and Burgard, W. 2007. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Transactions on Robotics* 23(1).

Hanheide, M.; Göbelbecker, M.; Horn, G. S.; Pronobis, A.; Sjöö, K.; Aydemir, A.; Jensfelt, P.; Gretton, C.; Dearden, R.; Janicek, M.; Zender, H.; Kruijff, G.-J.; Hawes, N.; and Wyatt, J. L. 2016. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence*.

Hawes, N.; Zender, H.; Sj, K.; Brenner, M.; Kruijff, G.-J.; and Jensfelt, P. 2009. Planning and acting with an integrated sense of space. In *Proc. of the International Workshop on Hybrid Control of Autonomous Systems*.

Holmes, N. P., and Spence, C. 2004. The body schema and multisensory representation(s) of peripersonal space. *Cognitive processing* 5(2).

Kuipers, B. 2000. The spatial semantic hierarchy. *Artificial intelligence* 119(1-2).

Levine, S.; Finn, C.; Darrell, T.; and Abbeel, P. 2016. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* 17(39):1–40.

Marder-Eppstein, E.; Berger, E.; Foote, T.; Gerkey, B.; and Konolige, K. 2010. The office marathon: Robust navigation in an indoor office environment. In *Proc. of ICRA*.

Mozos, O. M.; Stachniss, C.; and Burgard, W. 2005. Supervised learning of places from range data using AdaBoost. In *Proc. of ICRA*.

Peharz, R.; Tschiatschek, S.; Pernkopf, F.; and Domingos, P. 2015. On theoretical properties of Sum-product Networks. In *Proc. of AISTATS*.

Poon, H., and Domingos, P. 2011. Sum-product networks: A new deep architecture. In *Proc. of UAI*.

Pronobis, A., and Jensfelt, P. 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Proc. of ICRA*.

Pronobis, A., and Rao, R. P. N. 2017. Learning deep generative spatial models for mobile robots. arXiv:1610.02627 [cs.RO].

Pronobis, A.; Mozos, O. M.; Caputo, B.; and Jensfelt, P. 2010a. Multi-modal semantic place classification. *The International Journal of Robotics Research* 29(2-3).

Pronobis, A.; Sjöö, K.; Aydemir, A.; Bishop, A. N.; and Jensfelt, P. 2010b. Representing spatial knowledge in mobile cognitive systems. In *Proc. of the International Conference on Intelligent Autonomous Systems (IAS-11)*.

Thrun, S.; Bücken, A.; Burgard, W.; Fox, D.; Fröhlinghaus, T.; Henning, D.; Hofmann, T.; Krell, M.; and Schmidt, T. 1998. Map learning and high-speed navigation in RHINO. In Kortenkamp, D.; Bonasso, R.; and Murphy, R., eds., *AI-based Mobile Robots: Case Studies of Successful Robot Systems*. MIT Press.

Zender, H.; Mozos, O. M.; Jensfelt, P.; Kruijff, G.; and Burgard, W. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems* 56(6). Special Issue "From Sensors to Human Spatial Concepts".