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

Andrzej Pronobis
University of Washington
Seattle, WA, USA
Email: pronobis@cs.washington.edu

Francesco Riccio
Sapienza University of Rome
Rome, Italy
Email: riccio@diag.uniroma1.it

Rajesh P. N. Rao
University of Washington
Seattle, WA, USA
Email: rao@cs.washington.edu

***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.**

I. 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 [22, 23, 9]. While significant progress have been made in the area of robot control, largely thanks to the success of deep learning [13], 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), 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 as well as 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 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 structure of DASH corresponds to a hierarchical decomposition of the planning problem. Additionally, 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.

DASH includes both instance knowledge about the specific robot environment as well as default knowledge about typical human environments. The latter is modeled using a recently proposed Deep Generative Spatial Model (DGSM) [19]. 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 latent or missing spatial knowledge defined at various levels of abstraction.

In this paper, we describe the architecture of DASH and present its initial realization 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).

II. Related Work

The problem of representing spatial knowledge for mobile robots has received significant attention. However, most of the proposed approaches are specific to a particular sub-problem and focus on a fraction of the broad spectrum of spatial knowledge [11]. Moreover, the proposed solutions rarely answer the question of how spatial knowledge should be structured and used to support the behavior of the robot.

At the same time, several, more general frameworks have been proposed, that try to address this question directly. One of the first such frameworks was the Spatial Semantic Hierarchy [12], which concentrates on lower levels of spatial knowledge abstraction and does not support higher-level conceptualization. Other early frameworks were designed to build a representation from both spatial and semantic perspective [5, 27], but relied on traditional AI reasoning techniques unable to incorporate uncertainty that is crucial to achieve robustness in the real world. Several more recent approaches incorporated probabilistic models into the representation [25, 26, 18, 1, 9]. However, these works either modeled uncertainty for only certain aspects of the world or relied on an assembly of in dependent spatial models, which exchange information in a limited fashion.

On the other hand, recent deep learning revolution has demonstrated that replacing multiple representations with a single integrated model can lead to a drastic increase in performance [13]. In this work, we propose an approach that is comprehensive, designed specifically to support planning, while at the same time leveraging a new deep model of general knowledge spanning all levels of abstraction. In contrast to the method in [24], which utilizes a deep convolutional network for semantic mapping, our representation is fully probabilistic and generative. Another recently proposed end-to-end, deep architecture [8] learns to build metric maps and navigate towards goals in the environment directly from visual observations. While this approach focuses on a specific sub-problem, it demonstrates the benefits of deep integration between spatial understanding an hierarchical planning.

III. Analysis of the Problem

We begin with an analysis of roles and desired properties of a spatial knowledge representation. We focus specifically on scenarios involving large-scale, dynamic, human environments, such as office buildings, homes, and hospitals. We assume a mobile robot capable of moving around and sensing the environment, performing basic manipulation tasks (e.g. grasping objects or pushing buttons), as well as interacting and collaborating with humans (e.g. by asking for additional information or requesting help when a task cannot be accomplished by the robot alone).

We recognize that the ultimate purpose of a spatial knowledge representation is to enable and facilitate successful planning and execution of actions in the robot environment. More specifically, following [4], we can see a representation as:

a) A surrogate for the world that allows the robot to reason about 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 place or involving a spatial entity).

b) A set of commitments that determine the terms in which the robot thinks about space. A representation specifies which aspects of the world should be represented, at what level of abstraction they should exist, which spatial frame of reference should be used to relate them (absolute or relative, allo- or ego-centric), and how long their representation should persist. Importantly, these commitments significantly impact the ability of the robot to plan specific actions and can be seen as defining part of the action space of the robot.

c) A repository of general spatial knowledge that allows the robot to perform inferences about latent or missing information based on relations that typically occur in the real world. Such repository can be gathered from experienced or simply transferred to the robot.

d) A set of definitions that determine the reasoning that can be (and should be) performed. This includes reasoning about the location of the robot with respect to internal frames of reference, inferring abstract concepts from observations (e.g. affordances, semantic descriptions), or generating missing lower-level information from high-level descriptions (e.g. position of occluded objects based on room category).

e) A medium of communication between the robot and humans. Spatial knowledge is a natural common ground for human-robot communication and knowledge transfer. A representation supports this process by relating human spatial concepts to those internal to the robot.

f) A way of structuring spatial information so that it is computationally feasible to perform inferences and planning despite limited resources.

With that in mind and considering practical limitations of the scenario as well as experience resulting from existing approaches and robotic systems [12, 14, 23, 3, 9], we can identify several desired properties of a spatial knowledge representation.

First, spatial knowledge in realistic environments is inherently uncertain and dynamic. A very accurate representation is likely to be intractable and requires substantial effort to be kept up-to-date. Moreover, its usability remains constrained by the robot capabilities. Hence, our primary assumption is that the representation should be minimal and only as expressive as required to successfully act.

Complexity of planning increases exponentially with the number of considered spatial entities. However, decomposing the planning problem hierarchically can greatly reduce its complexity while maintaining highly optimal results. For this reason, hierarchical planners are used in the majority of existing robotic systems [14, 1, 9, 8]. Moreover, behavioral analyses found hierarchical spatial planning in humans [2]. Such an approach leads to a hierarchy of higher-level, long-term, global plans involving lower-level short-term, local behaviors. 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 often leads to discretization of continuous space, which reduces the number of states for planning [10] and provides a basis for higher-level conceptualization [27].

While abstracted knowledge is more likely to remain up-to-date, detailed, spatial information is required for executing actions in the robot peripersonal space. Yet, it can also be re-acquired through perception. A representation should correlate the levels of abstraction with the persistence of information, employing local working memory for integrating high-resolution spatial information. In other words, the robot should rely on 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, e.g. when planning for epistemic actions. Decision-theoretic planning relies on probabilistic representations of uncertainty, therefore, it is desirable for a representation to also be probabilistic in nature.

Finally, 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 [9]), trade exploration vs. exploitation [1], or drive learning.

IV. Deep Spatial Affordance Hierarchy (DASH)

Based on this analysis, we propose Deep Spatial Affordance Hierarchy (DASH) as well as its initial realization for a mobile robot equipped with a laser-range sensor. 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 are summarized in Table I and were chosen to simultaneously support both action planning and spatial understanding. 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.

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 hierarchical decomposition of the planning problem. A global

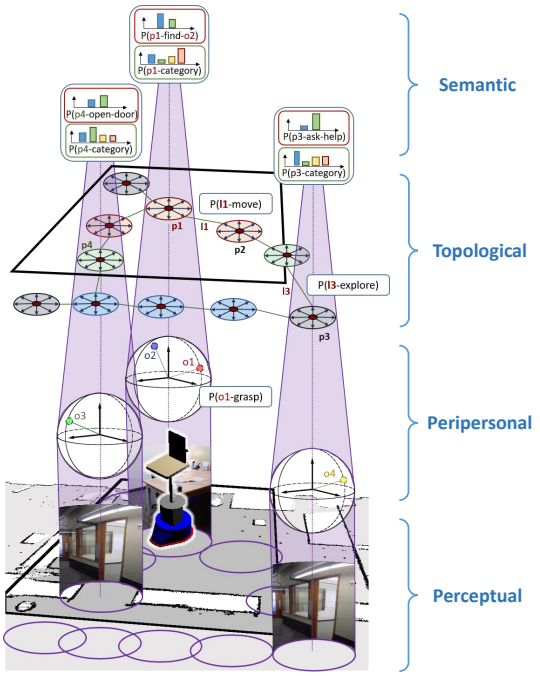


Fig. 1: The multi-layered architecture of DASH. The *perceptual layer* integrates perceptual information from the robot sensors. The *peripersonal layer* represents object and landmark information and affordances in the space immediately surrounding the robot. The *topological layer* encodes global topology,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

planner can derive a navigation plan relying only on the top layers for representing its beliefs, a local planner can be used to plan 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.

A. Perceptual Layer

The bottom, *perceptual layer* maintains an accurate representation of geometry and appearance of the environment obtained by short-term spatio-temporal integration of perceptual information. It relies on 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 robot location, and spans a radius roughly corresponding to the sensory horizon. The layer provides a more complete input for further abstractions with reduced occlusions and noise, and forms a basis for deriving low-level control laws. Missing observations (e.g. due to unresolved occlusions) are explicitly represented.

In our initial implementation, we realize the layer using a grid mapping approach based on Rao-Blackwellized particle

	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	Unexplored space	Novel semantic concepts

TABLE I: Characteristics of the four layers of DASH.

filter [7]. We crop the 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 environment is mapped correctly.

B. Peripersonal Layer

The *peripersonal layer* captures spatial information related to object and landmark instances from the perspective of an agent performing actions at different locations in the environment. The purpose of the layer is to represent object affordances related to actions that can be performed directly by the robot. This includes manipulation (e.g. reaching/grasping an object or pressing a button), interaction in relation to objects (e.g. pointing at an object), and epistemic action affordances (e.g. observing an object). Furthermore, the layer captures internal object and landmark descriptors as well as spatial relations between objects in relation to the robot (and therefore coarse local geometry). These descriptors are generated by the deep default knowledge model and serve as a basis for further abstractions.

To reflect the local and robo-centric nature of the represented information, the layer consists of a collection of ego-centric, metric reference frames, each focusing on the space immediately surrounding the robot at a different place 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 additional context. For instance, a reference frame centered in front of a desk might include information about shelves in the room, even beyond the reach of the robot. The peripersonal layer explicitly represents gaps in knowledge about the local space due to missing evidence (e.g. due to occlusions).

In our initial realization, the peripersonal representation for each place is built and updated from the current local occupancy grid in the perceptual layer. For each place, the representation retains information about the part of the environment roughly corresponding to the boundaries of the current room. To this end, we include grid cells that can be raytraced from the robot location and augment those with observations behind smaller obstacles. The selected grid cells are transformed into an ego-centric polar representation (see Fig. 4b for examples). This encodes high-resolution information about the geometry and objects nearby, and complements it with less-detailed context further away from the robot. From the perspective of spatial understanding, this provides sufficient context (e.g. the outline of a room) complemented with relevant details for understanding the semantics of the exact location (e.g. when the robot is in a doorway). From the perspective of planning, it is also in the vicinity of the robot that higher accuracy of spatial information is required. The polar grids in our implementation assume a maximum radius of 5m, with angle step of 6.4° and resolution decreasing with the distance from the robot. 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.

C. 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. The layer performs a bottom-up discretization of continuous space visited by the robot into a set of locations, called *places*, and a set of discrete headings, called *views*. The density of places should be sufficient for planning global navigation, while maintaining efficiency and robustness to dynamic changes. Together, views and places are used to represent the complete global pose of the robot. Additionally, they anchor knowledge in the representation, including robot-internal descriptors of each view and place derived from lower-levels using the deep default knowledge model, and a peripersonal representation describing the place in more detail.

Furthermore, the layer 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 (together with estimated uncertainty). The topological nature of the graph enables planning of complex navigational tasks. For instance, a place in an elevator

might afford navigating to places on different floors, depending on the information captured in the peripersonal layer (e.g. displayed floor number).

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* [21], which can be seen as candidate places 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 in the continual planning paradigm [9].

In our implementation, the topological layer is expanded incrementally. Placeholders are added at neighboring unexplored locations based on information in the perceptual layer, and connected with paths to existing places. Then, once the robot performs an exploration action, 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 associated with probability based on up-to-date information.

Similarly to [3], we formulate the problem of finding placeholder locations as sampling from a probability distribution that models their relevance and suitability. However, instead of sampling locations of all places at once, we incrementally add placeholders, within the scope of the perceptual layer. Specifically, the probability distribution is modeled as: $P(E|G,\mathcal{E}) = \frac{1}{Z}\prod_i \phi_S(E_i|G)\phi_N(E_i|\mathcal{E})$, where $E_i \in \{0,1\}$ represents the existence of a new place at location i in the perceptual layer, G is the perceptual occupancy grid, and $\mathcal{E}$ is the set of locations of all existing places. The potential ϕ_S ensures that placeholders are located in areas that are safe and preferred for navigation and constitute useful anchors. It is defined in terms of potentials calculated from G: $\phi_S(E_i) = \phi_O(E_i)(\phi_V(E_i) + \phi_P(E_i) - \phi_V(E_i)\phi_P(E_i))$, where:

- ϕ_O ensures that placeholders are created in areas safe from collisions with obstacles. It depends on the distance d_o to the nearest obstacle and is calculated similarly to the cost used for obstacle avoidance [14]. ϕ_o equals 0 for distance smaller than the radius r of the robot and $1-exp(-\alpha(d_o - r))$ otherwise.
- $\phi_V = exp(-\gamma d_c)$ depends on the distance d_c to the nearest node of a Voronoi graph of the 2D grid G. 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 the average width of a door.

The potential ϕ_N promotes positions at a certain distance d_n from existing places, and is defined as: $\phi_N(E_i|\mathcal{E}) = \sum_{p\in\mathcal{E}} exp(-\frac{(d(i,p)-d_n)^2}{2\sigma^2})$, where $d(i,p)$ is an Euclidean distance between the potential new place and an existing place. Final location of a new placeholder is chosen through MPE inference in $P(E|G,\mathcal{E})$. A path is then created which connects the placeholder to an existing place. Each path is associated with probability indicating navigability, which is estimated by performing A* search directly over the potential function ϕ_S and accumulating the potential along the trajectory.

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 the robot moves through the environment, the lattice is extended, while preserving consistency with existing lattice points. When performing MPE inference using $P(E|G,\mathcal{E})$, we assume that only one place can exist in a cell of a Voronoi tessellation established by the points of the lattice. The resulting set of placeholders 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 new place, we generate a set of eight *views* which are assumed to be vectors pointing from a point of the lattice to the eight immediately neighboring points.

D. Semantic Layer

On the 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. It is the role of the semantic layer to capture the knowledge that an object is likely to be a cup, or that certain places are likely to be located in a kitchen. Furthermore, it is the semantic layer that represents place affordances related to human interaction and actions characterized in terms of human concepts (e.g. asking a person for help with opening a door or finding a cup at a place). Finally, the layer stores asserted human knowledge passed to the robot, which can be used by the default knowledge model to infer lower-level information.

In our initial implementation, the semantic layer is a simple relational data structure that captures the information about semantic categories of places in the topological map. This includes categories of rooms in which places are located (e.g. an office or a corridor), but also a functional place category corresponding to a doorway. The relations are inferred by the default knowledge model from place descriptors and each relation is associated with a probability value. Additionally, for each place, the layer captures the likelihood of the information in the peripersonal representation under the default knowledge model. This likelihood is used to detect and explicitly represent that a place belongs to a novel, previously unseen category.

V. REPRESENTING DEEP DEFAULT KNOWLEDGE

While the four layers of the representation focus on instance knowledge about a specific robot environment, the *deep default knowledge model* captures general spatial knowledge about typical human environments. It provides definitions of spatial concepts and their relations across all levels of abstraction. This includes models of objects in terms of low-level perception, places in terms of objects, or semantic categories and attributes of objects and places. The role of the default

knowledge model is to permit inferences about missing or latent aspects of the environment based on the evidence collected across all the layers of the representation. This includes bottom-up inferences (e.g. about semantic descriptions based on perception) and top-down inferences (e.g. about place geometry based on semantic descriptions). The result is a more complete (albeit uncertain) belief state for the planner.

In our implementation, the default knowledge is modeled using a recently proposed Deep Generative Spatial Model (DGSM) [19], a probabilistic deep model which learns a joint distribution over spatial knowledge represented at multiple levels of abstraction. 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 in the opposite direction. The model can infer missing information in peripersonal layer resulting from partial observations and generate prototypical peripersonal representations based on semantic descriptions. To this end, DGSM leverages Sum-Product Networks (SPNs), a novel probabilistic deep architecture [17, 16], and a unique structure which matches the hierarchy of instance representations in DASH. Below, we give a primer on SPNs and describe the architecture of DGSM.

A. Sum-Product Networks

SPNs are a recently proposed probabilistic deep architecture with several appealing properties and solid theoretical foundations [16, 17, 6]. 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, and involve a polynomial number of sum and product operations, permitting exact inference. SPNs model joint or conditional distributions and can be learned generatively [17] or discriminatively [6] using Expectation Maximization (EM) or gradient descent. They are a deep, hierarchical representation, capable of representing context-specific independence.

As shown in Fig. 2, an SPN is a generalized directed acyclic graph of alternating layers of weighted sum and product operations. The sums can be seen as mixture models, over components defined using products, with weights of each sum representing mixture priors. The latent variables of such mixtures can be made explicit and their values inferred. In DGSM, such latent variables are used to represent internal descriptors of places and views and their semantic descriptions. The bottom layers effectively define features reacting to certain values of input variables. Not all possible architectures consisting of sums and products result in valid probability distributions and certain constraints [17, 16] must be followed to guarantee validity.

Evidence in SPNs is typically specified in terms of binary indicators corresponding to values of categorical variables.

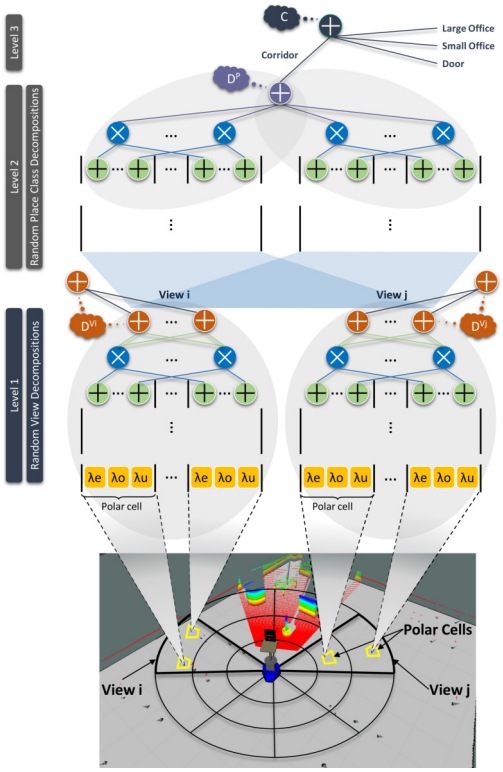


Fig. 2: The structure of the SPN implementing DGSM. Weighted sum nodes are marked with $+$, while product nodes are marked with $\times$. Explicit latent variables of the mixture models realized by the sums are indicated using callouts. The bottom image illustrates a robot in an environment and a robocentric polar grid of the peripersonal layer formed around the robot. The SPN is built on top of indicator variables (orange) representing the occupancy in each cell of the polar grid (one for empty, occupied and unknown space).

Partial or missing evidence can be expressed by setting all indicators of a variable to 1 (see [17] for a discussion about continuous variables). Inference is then accomplished by an upwards pass which calculates the probability of the evidence and a downwards pass which obtains marginal distributions or MPE state of the missing evidence [17, 6]. In this work, we learn the SPN using hard EM, which was shown to work well for generative learning [17] and overcomes the diminishing gradient problem. The reader is referred to [19] for details about the learning procedure.

B. Architecture of DGSM

The architecture of DGSM is based on a generative SPN illustrated in Fig. 2. The model learns a probability distribution $P(C, D_1^P, \ldots, D_{N_p}^P, D_1^{V_1}, \ldots, D_{N_v}^{V_8}, X_1, \ldots, X_{N_x})$, where C represents the semantic category of a place, $D_1^P, \ldots, D_{N_p}^P$ constitute an internal descriptor of the place, $D_1^{V_1}, \ldots, D_{N_v}^{V_8}$ are descriptors of eight views of the place, and $X_1, \ldots, X_C$ are input variables representing the occupancy in each cell of the polar grid of the peripersonal layer.

The structure of the model is partially static and partially

generated randomly according to the algorithm described in [19]. 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 generate a sub-SPN over the subset of X_i representing the peripersonal information within the view, as well as latent variables $D_1^{V_i}, \ldots, D_{N_v}^{V_i}$ serving as the internal view descriptor. The root of that sub-SPN can be seen as a mixture model consisting of 14 components in our implementation. On the second level, we use the 14 components for each view ($8*14$ in total) as inputs, and generate an SPN representing the complete place. Such place SPN is a distribution modeling a single semantic place category and is itself a mixture model with the latent variable D_i^P being part of the place descriptor. Finally, on the third level, we combine multiple place SPNs, each representing a single semantic category by a sum node forming the root of the complete network. The latent variable associated with the root node is C which is set to the appropriate semantic 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 descriptions.

VI. Experimental Evaluation

Our experimental evaluation consists of two parts. First, we demonstrate 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 deploy our complete implementation of DASH in order to build representations of large-scale environments.

A. Experimental Setup

The experiments were performed on laser range data from the COLD-Stockholm database [18]. 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. There are 9 different *large offices*, 8 different *small offices* (distributed across the floors), 4 long *corridors* (1 per floor, with varying appearance in different parts), and multiple examples of places in *doorways*. The dataset features several other room categories: an elevator, a living room, a meeting room, and a kitchen. However, with only one or two room instances in each. Therefore, we decided to designate those as novel when testing novelty detection and used the remaining four categories for the majority of the experiments. To ensure variability between the training and test sets, we split the data samples four times, each time training the DGSM model on samples from three floors and leaving one floor out for testing. The presented results are averaged over the four splits.

B. Bottom-up Inference

The bottom-up inference was tested for the task of inferring semantic place categories and detecting novel place categories given the information in the peripersonal layer. As a comparison, we used a well-established model based on an SVM and geometric features extracted from laser scans [15, 20]. The features were extracted from scans raytraced in the same local Cartesian grid maps used to form polar grids of the peripersonal layer. We raytraced 362 beams around the robot in high-resolution maps (2cm/pixel). To ensure the best SVM result, we used the RBF kernel and selected the kernel and learning parameters directly on the test set. We used C-SVC for classification and 1-class SVM for novelty detection.

The models were trained for places belonging to the four place categories from three floors, and evaluated on the fourth floor or using data from rooms designated as novel. The classification rate averaged over all classes (giving equal importance 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. For the novelty detection task, to decide whether the robot is located in a place belonging to a class known during training, we thresholded the likelihood produced by DGSM for the test peripersonal representations. We compared that to predictions of 1-class SVM. As a measure of performance, we used the area under the ROC curve (AUC). Here, again, DGSM performed significantly better, with AUC of 0.81 compared to 0.76 for SVM.

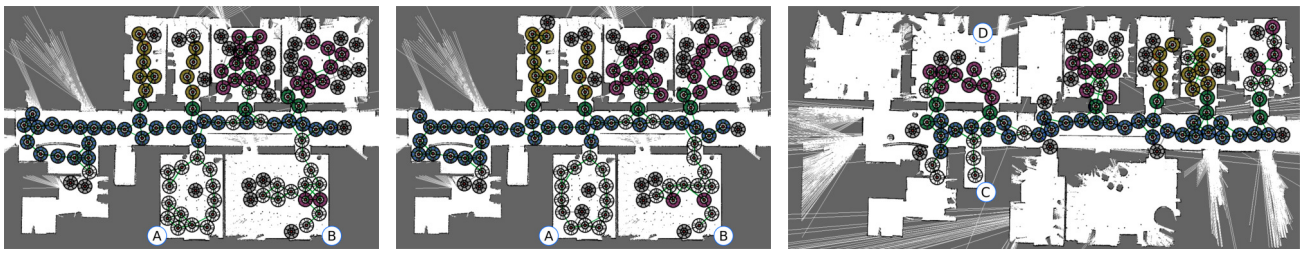
C. Top-down Inference

To test top-down inference, we used DGSM to infer values of cells in the peripersonal representation. First, we inferred complete, prototypical representations of places knowing only semantic categories. The generated polar occupancy grids are shown in Fig. 4a (compare to true data samples shown in Fig. 4b). We can see that each peripersonal representation is very characteristic of the class from which it was generated. The corridor is an elongated structure, the doorway is a narrow structure with empty space on both sides, and the offices clearly differ in shape and size.

Then, we used DGSM to generate missing values in partial peripersonal representations. To this end, we masked a random 90-degree view in each test polar grid (25% of the grid cells) and inferred the masked values. Fig. 4b shows examples of completed peripersonal representations. 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.

D. Representing Large-Scale Space

Finally, we deployed the complete implementation of DASH and evaluated its ability to build comprehensive, multi-layered representations of large-scale space. We incrementally built the representations based on the sensory data captured as the robot navigated through floors of the building in the COLD-dataset. We relied on the perceptual layer to perform low-level integration of observed laser-range data, on the 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



(a) Floor 5, run #1 (b) Floor 5, run #2 (c) Floor 7

Fig. 3: Contents of the topological and semantic layers after two different runs over 5-*th* floor (a-b) and a run over the 7-*th* floor (c). 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. Rooms marked with letters A-D belong to novel categories: A and B are meeting rooms, C is a living room and D is an elevator.

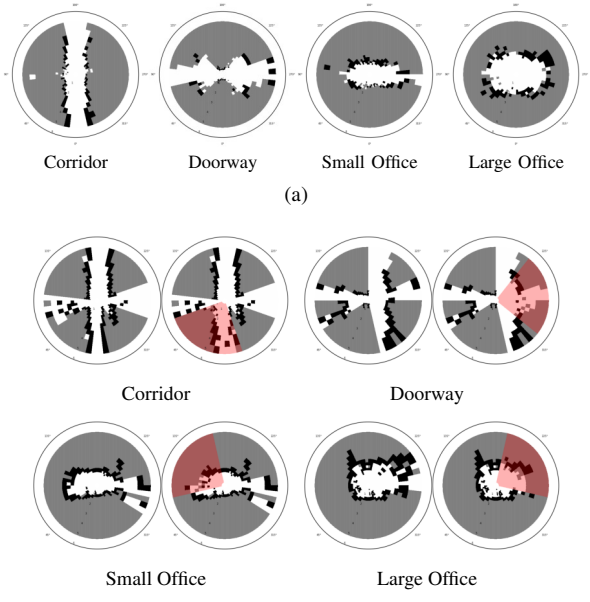


Fig. 4: (a) Prototypical peripersonal representations inferred from semantic place category. (b) Examples of completed peripersonal representations with masked data. The examples are grouped by true semantic category, with the left image showing original data, and the right image showing reconstructed data within the shaded area.

on the semantic layer to encode information about semantic categories of places, including detections of novel semantic categories.

Fig. 3a-b illustrates the state of the topological and semantic layers after two independent runs over the 5-*th* floor. While not visualized, the peripersonal representations were also built for each place. The figure presents the final graph of semantically annotated places visited by the robot, paths expressing navigability between them, as well as paths leading to placeholders representing possibility of further exploration. 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. Additionally, the semantic place categories inferred by DGSM agree with the ground truth when the category of the place was recognized as known. Places in the two rooms belonging to a novel category "meeting room" (marked as A and B) are largely correctly detected as novel, although both false positives and false negatives exist.

Fig. 3c shows a corresponding result for a different environment, the 7-*th* floor. Here the novelty detection is less accurate. DGSM correctly detects the places in the elevator (marked as D) as novel, but fails to detect novelty in the living room (marked as C), which instead is misclassified as a large office. This can be explained by the similarity between the living room and large offices in the dataset when observed solely using laser range sensor. Overall, for both floors, our implementation of DASH generates highly accurate and rich representations.

VII. Conclusions and Future Work

This paper presents the 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 perception 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 DASH with hierarchical planning to demonstrate its capacity to support autonomous robot behavior in complex realistic scenarios.

Acknowledgments

This work was supported by the Swedish Research Council (VR) project SKAEENet. The support is gratefully acknowledged.

References

[1] A. Aydemir et al. “Active Visual Object Search in Unknown Environments Using Uncertain Semantics”. In: *Transactions on Robotics* 29.4 (2013).

[2] J. Balaguer et al. “Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network”. In: *Neuron* 90.4 (2016).

[3] M. J.-Y. Chung et al. “Autonomous Question Answering with Mobile Robots in Human-Populated Environments”. In: *Proc. of IROS*. 2016.

[4] R. Davis, H. Shrobe, and P. Szolovits. “What is a Knowledge Representation”. In: *AI Magazine* 14.1 (1993).

[5] C. Galindo et al. “Multi-hierarchical semantic maps for mobile robotics”. In: *Proc. of IROS*. 2005.

[6] R. Gens and P. Domingos. “Discriminative Learning of Sum-product Networks”. In: *Proc. of NIPS*. 2012.

[7] G. Grisetti, C. Stachniss, and W. Burgard. “Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters”. In: *Transactions on Robotics* 23.1 (2007).

[8] S. Gupta et al. *Cognitive Mapping and Planning for Visual Navigation*. Feb. 2017. arXiv: 1702 . 03920 [cs.CV].

[9] M. Hanheide et al. “Robot Task Planning and Explanation in Open and Uncertain Worlds”. In: *Artificial Intelligence* (2016).

[10] N. Hawes et al. “Planning and Acting with an Integrated Sense of Space”. In: *International Workshop on Hybrid Control of Autonomous Systems*. 2009.

[11] I. Kostavelis and A. Gasteratos. “Semantic Mapping for Mobile Robotics Tasks: A Survey”. In: *Robotics and Autonomous Systems* 66 (2015).

[12] B. Kuipers. “The spatial semantic hierarchy”. In: *Artificial intelligence* 119.1-2 (2000).

[13] S. Levine et al. “End-to-End Training of Deep Visuomotor Policies”. In: *Journal of Machine Learning Research* 17.39 (2016).

[14] E. Marder-Eppstein et al. “The Office Marathon: Robust Navigation in an Indoor Office Environment”. In: *Proc. of ICRA*. 2010.

[15] O. M. Mozos, C. Stachniss, and W. Burgard. “Supervised learning of places from range data using AdaBoost”. In: *Proc. of ICRA*. 2005.

[16] R. Peharz et al. “On Theoretical Properties of Sum-product Networks”. In: *Proc. of AISTATS*. 2015.

[17] H. Poon and P. Domingos. “Sum-product Networks: A New Deep Architecture”. In: *Proc. of UAI*. 2011.

[18] A. Pronobis and P. Jensfelt. “Large-scale Semantic Mapping and Reasoning with Heterogeneous Modalities”. In: *Proc. of ICRA*. 2012.

[19] A. Pronobis and R. P. N. Rao. “Learning Deep Generative Spatial Models for Mobile Robots”. In: *Proc. of IROS*. 2017.

[20] A. Pronobis et al. “Multi-modal Semantic Place Classification”. In: *The International Journal of Robotics Research* 29.2-3 (Feb. 2010).

[21] A. Pronobis et al. “Representing Spatial Knowledge in Mobile Cognitive Systems”. In: *Proc. of IAS*. 2010.

[22] S. Rosenthal, J. Biswas, and M. Veloso. “An Effective Personal Mobile Robot Agent Through Symbiotic Human-robot Interaction”. In: *Proc. of AAMAS*. 2010.

[23] K. Sjöö et al. “The Explorer System”. In: *Cognitive Systems*. Ed. by H. Christensen, G.-J. Kruijff, and J. Wyatt. Vol. 8. 2010.

[24] N. Sünderhauf et al. “Place Categorization and Semantic Mapping on a Mobile Robot”. In: *Proc. of ICRA*. 2016.

[25] S. Vasudevan and R. Siegwart. “Bayesian Space Conceptualization and Place Classification for Semantic Maps in Mobile Robotics”. In: *RAS* 56.6 (2008).

[26] P. Viswanathan et al. “Automated Spatial-Semantic Modeling with Applications to Place Labeling and Informed Search”. In: *Proc. of CRV*. 2009.

[27] H. Zender et al. “Conceptual Spatial Representations for Indoor Mobile Robots”. In: *Robotics and Autonomous Systems* 56.6 (2008).