Building Multiversal Semantic Maps for Mobile Robot Operation

Jose-Raul Ruiz-Sarmiento[a,*], Cipriano Galindo[a], Javier Gonzalez-Jimenez[a]

[a] *Machine Perception and Intelligent Robotics Group*
System Engineering and Auto. Dept., Instituto de Investigación Biomédica de Málaga (IBIMA), University of Málaga, Campus de Teatinos, 29071, Málaga, Spain.

Abstract

Semantic maps augment metric-topological maps with meta-information, *i.e. semantic knowledge* aimed at the planning and execution of high-level robotic tasks. Semantic knowledge typically encodes human-like concepts, like types of objects and rooms, which are connected to sensory data when symbolic representations of percepts from the robot workspace are grounded to those concepts. This *symbol grounding* is usually carried out by algorithms that individually categorize each symbol and provide a crispy outcome – a symbol is either a member of a category or not. Such approach is valid for a variety of tasks, but it fails at: (i) dealing with the uncertainty inherent to the grounding process, and (ii) jointly exploiting the contextual relations among concepts (*e.g.* microwaves are usually in kitchens). This work provides a solution for *probabilistic symbol grounding* that overcomes these limitations. Concretely, we rely on Conditional Random Fields (CRFs) to model and exploit contextual relations, and to provide measurements about the uncertainty coming from the possible groundings in the form of beliefs (*e.g.* an object can be categorized (grounded) as a microwave or as a nightstand with beliefs 0.6 and 0.4, respectively). Our solution is integrated into a novel semantic map representation called *Multiversal Semantic Map* ($\mathcal{M}v\mathcal{S}map$), which keeps the different groundings, or universes, as instances of ontologies annotated with the obtained beliefs for their posterior exploitation. The suitability of our proposal has been proven with the Robot@Home dataset, a repository that contains challenging multi-modal sensory information gathered by a mobile robot in home environments.

Keywords: mobile robots, symbol grounding, semantic maps, conditional random fields, ontologies, probabilistic inference

1. Introduction

A mobile robot intended to operate within human environments needs to create and maintain an internal representation of its workspace, commonly referred to as a *map*. Robotic systems rely on different types of maps depending on their goals. For example, *metric maps* are purely geometric representations that permit robot self-localization with respect to a given reference frame [1, 2]. *Topological maps* consider a graph structure to model areas of the environment and their connectivity, hence straightforwardly supporting navigational planning tasks [3, 4]. In its turn, *Hybrid maps* come up from the combination of the previous ones by maintaining local metric information and a graph structure to perform basic but core robotic skills as localization and global navigation [5, 6]. A pivotal requirement for the successful building of these types of maps is to deal with uncertainty coming, among other sources, from errors in the robot perception (limited field of view and range of sensors, noisy measurements, etc.), and inaccurate models and algorithms. This issue is addressed in state-of-the-art approaches through probabilistic techniques [7].

Despite the possibilities of these representations, planning and executing high-level robotic tasks within human-like environments demand more sophisticated maps to enable robots, for example, to deal with user commands like *"hey robot! I am leaving, take care of the oven while I am out, please"* or *'Guide the customer through the aisle with garden stuff and show him the watering cans"*. Humans share a common-sense knowledge about concepts like *oven*, or *garden stuff*, which must be transferred to robots in order to successfully face those tasks. *Semantic maps* emerged to cope with this need, providing the robot with the capability to *understand*, not only the spatial aspects of human environments, but also the meaning of their elements (objects, rooms, etc.) and how humans interact with them (*e.g.* functionalities, events, or relations). This feature is distinctive and traversal to semantic maps, being the key difference with respect to maps that simply augment metric/topological models with labels to state the category of recognized objects or rooms [8, 9, 10, 11, 12]. Contrary, semantic maps handle meta-information that models the properties and relations of relevant concepts therein the domain at hand, codified into a *Knowledge Base* (KB), stating that, for example, microwaves are box-shaped objects usually found in kitchens and useful for heating food. Building and maintaining semantic maps involve the symbol grounding problem [13, 14, 15], *i.e.* linking portions of the sensory data gathered by the robot (percepts), represented by symbols, to concepts in the KB by means of some categorization and tracking method.

Semantic maps generally support the execution of reasoning engines, providing the robot with inference capabilities for efficient navigation, object search [16], human-robot interac-

*Corresponding author
Email addresses: jotaraul@uma.es (Jose-Raul Ruiz-Sarmiento), cipriano@ctima.uma.es (Cipriano Galindo), javiergonzalez@uma.es (Javier Gonzalez-Jimenez)

tion [17] or pro-activeness [18] among others. Typically, such engines are based on logical reasoners that work with crispy[1] information (*e.g.* a percept is identified as a microwave or not). The information encoded in the KB, along with that inferred by logical reasoners, is then available for a task planning algorithm dealing with this type of knowledge and orchestrating the aforementioned tasks [19]. Although crispy knowledge-based semantic maps can be suitable in some setups, especially in small and controlled scenarios [20], they are also affected by uncertainty coming from both, the robot perception, and the inaccurate modeling of the elements within the robot workspace. Moreover, these systems usually reckon on off-the-shelf categorization methods to individually ground percepts to particular concepts, which disregard the contextual relations between the workspace elements: a rich source of information intrinsic to human-made environments (for example that nigh-stands are usually in bedrooms and close to beds).

In this work we propose a solution for addressing the symbol grounding problem from a probabilistic stance, which permits both exploiting contextual relations and modeling the aforementioned uncertainties. For that we employ a Conditional Random Field (CRF), a particular type of Probabilistic Graphical Model [21], to represent the symbols of percepts gathered from the workspace as nodes in a graph, and their geometric relations as edges. This representation allows us to jointly model the symbol grounding problem, hence exploiting the relations among the elements in the environment. CRFs support the execution of probabilistic inference techniques, which provide the beliefs about the grounding of those elements to different concepts (*e.g.* an object can be a bowl or a cereal box with beliefs 0.8 and 0.2 respectively). In other words, the uncertainty coming both from the robot perception, and from the own symbol grounding process, is propagated to the grounding results in the form of beliefs.

The utilization of CRFs also leads to a number of valuable advantages:

- *Fast inference*: probabilistic reasoning algorithms, resorting to approximate techniques, exhibit an efficient execution that permits the retrieval of inference results in a short time [22, 23].
- *Multi-modal information*: CRFs easily integrate percepts coming from different types of sensors, *e.g.* RGB-D images and 2D laser scans, related to the same elements in the workspace [21].
- *Spatio-temporal coherence*: they can be dynamically modified to mirror new information gathered by the robot, also considering previously included percepts. This is done in combination with an anchoring process [14].
- *Life-long learning*: CRFs can be re-trained in order to take into account new concepts not considered during the initial training, but that could appear in the current robot workspace [24].

[1] For the purpose of this work, the term *crispy* takes the same meaning as in classical logic: it refers to information or processes dealing with facts that either are true or not.

In order to accommodate the probabilistic outcome of the proposed grounding process, a novel semantic map representation, called *Multiversal Semantic Map* ($\mathcal{M}v\mathcal{S}map$), is presented. This map extends the previous work by Galindo *et al.* [25], and considers the different combinations of possible groundings, or *universes*, as instances of ontologies [26] with belief annotations on their grounded concepts and relations. According to these beliefs, it is also encoded the probability of each ontology instance being the right one. Thus, $\mathcal{M}v\mathcal{S}maps$ can be exploited by logical reasoners performing over such ontologies, as well as by probabilistic reasoners working with the CRF representation. This ability to manage different semantic interpretations of the robot workspace, which can be leveraged by probabilistic conditional planners (*e.g.* those in [27] or [28]), is crucial for a coherent robot operation.

To study the suitability of our approach, we have conducted an experimental evaluation focusing on the construction of $\mathcal{M}v\mathcal{S}maps$ from facilities in the novel Robot@Home dataset [29]. This repository consists of 81 sequences containing 87,000+ timestamped observations (RGB-D images and 2D laser scans), collected by a mobile robot in different ready to move apartments. Such dataset permits us to intensively analyze the semantic map building process, demonstrating the claimed representation virtues. As an advance on this study, a success of $\sim 81.5\%$ and $\sim 91.5\%$ is achieved while grounding percepts to object and room concepts, respectively.

The next section puts our work in the context of the related literature. Sec. 3 introduces the proposed Multiversal Semantic Map, while Sec. 4 describes the processes involved in the building of the map for a given environment, including the probabilistic symbol grounding. The suitability of our approach is demonstrated in Sec. 5, and Sec. 6 discuses some of its potential applications. Finally, Sec. 7 concludes the paper.

2. Related work

This section reviews the most relevant related works addressing the symbol grounding problem (Sec. 2.1), aiming to put into context our probabilistic solution, as well as the most popular approaches for semantic mapping that can be found in the literature (Sec. 2.2).

2.1. Symbol grounding

As commented before, the symbol grounding problem consists of linking symbols that are meaningless by themselves to concepts in a Knowledge Base (KB), hence retrieving a notion of their meanings and functionalities in a given domain [13]. In the semantic mapping problem, symbols are typically abstract representations of percepts from the robot workspace, namely objects and rooms [15, 30]. Therefore, a common approach to ground those symbols is their processing by means of categorization systems, whose outcomes are used to link them to concepts in the KB. The remaining of this section provides a brief overview of categorization approaches for both objects and rooms, and concludes with our proposal for a probabilistic grounding.

In its beginnings, the vast literature around object categorization focused on the classification of isolated objects employing their geometric/appearance features. A popular example of this is the work by Viola and Jones [31], where an integral image representation is used to encode the appearance of a certain object category, and is exploited by a cascade classifier over a sliding window to detect occurrences of such object type in intensity images. A limiting drawback of this categorization method is the lack of an uncertainty measurement about its outcome. Another well known approach, which is able to provide such uncertainty, is the utilization of image descriptors like Scale-Invariant Feature Transform (SIFT) [32] or Speeded-Up Robust Features (SURF) [33] to capture the appearance of objects, and its posterior exploitation by classifiers like Supported Vector Machines (SVMs) [34] or Bag-of-Words based ones [35, 36]. The work by Zhang *et al.* [37] provides a comprehensive review of methods following this approach. It is also considerable the number of works tackling the room categorization problem through the exploitation of their geometry or appearance, like the one by Mozos *et al.* [38] which employs range data to classify spaces according to a set of geometric features. Also popular are works resorting to global descriptors of intensity images, like the *gist* of the scene proposed by Oliva and Torralba [39], those resorting to local descriptors like the aforementioned SIFT and SURF [40, 41], or the works combining both types of cues, global and local, pursuing a more robust performance [42, 43]. Despite the acceptable success of these *traditional* approaches, they can produce ambiguous results when dealing with objects/rooms showing similar features to two or more categories [44]. For example, these methods could have difficulties to categorize a white, box-shaped object as a microwave or a nightstand.

For that reason, modern categorization systems also integrate contextual information of objects/rooms, which has proven to be a rich source of information for the disambiguation of uncertain results [45, 46, 47]. Following the previous example, if the object is located in a bedroom and close to a bed, this information can be used to determine that it will likely be a nightstand. Probabilistic Graphical Models (PGMs) in general, and Undirected Graphical Models (UGMs) in particular, have became popular frameworks to model such relations and exploit them in combination with probabilistic inference methods [21]. Contextual relations can be of different nature, and can involve objects and/or rooms.

On the one hand, objects are not placed randomly, but following configurations that make sense from a human point of view, *e.g.* faucets are on sinks, mouses can be found close to keyboards, and cushions are often placed on couches or chairs. These object–object relations have been exploited, for example, by Anand *et al.* [48], which reckon on a model isomorphic to a Markov Random Field (MRF) to leverage them in home and office environments, or by Valentin *et al.* [49], which employ a Conditional Random Field (CRF), the discriminant variant of MRFs, to classify the faces of mesh-based representations of scenes compounded of objects according to their relations. Other examples of works also resorting to CRFs are the one by Xiong and Huver [50], which employs them to categorize the main components of facilities: clutters, walls, floors and ceilings, and those by Ruiz-Sarmiento *et al.* [22, 51, 52], where CRFs and ontologies [26] work together for achieving a more efficient and coherent object categorization.

On the other hand, object–room relations also supposes a useful source of information: objects are located in rooms according to their functionality, so the presence of an object of a certain type is a hint for the categorization of the room and, likewise, the category of a room is a good indicator of the object categories that can be found therein. Thus, recent works have explored the joint categorization of objects and rooms leveraging both, object–object and object–room contextual relations. CRFs have proven to be a suitable choice for modelling this holistic approach, as it has been shown in the works by Rogers and Christensen [53], Lin *et al.* [54], or Ruiz-Sarmiento *et al.* [55].

In this work we propose the utilization of a CRF to jointly categorize the percepts of objects and rooms gathered during the robot exploration of an environment, as well as its integration into a symbol grounding system. This CRF is exploited by a probabilistic inference method, namely Loopy Belief Propagation (LBP) [56, 57], in order to provide uncertainty measurements in the form of beliefs about the grounding of the symbols of these percepts to categories. Such categories correspond to concepts codified within an ontology, stating the typical properties of objects and rooms, and giving a semantic meaning to those symbols. Additionally, to make the symbols and their groundings consistent over time, we rely on an anchoring process [14]. To accommodate the outcome of this probabilistic symbol grounding, a novel semantic map representation is proposed.

2.2. Semantic maps

In the last decade, a number of works have appeared in the literature contributing different semantic map representations. One of the earliest works in this regard is the one by Galindo *et al.* [25], where a multi-hierarchical representation models, on the one hand, the concepts of the domain of discourse through an ontology, and on the other hand, the elements from the current workspace in the form of a spatial hierarchy that ranges from sensory data to abstract symbols. NeoClassic is the chosen system for knowledge representation and reasoning through Description Logics (DL), while the employed categorization system is limited to the classification of simple shape primitives, like boxes or cylinders, as furniture, *e.g.* a red box represents a couch. The potential of this representation was further explored in posterior works, *e.g.* for improving the capabilities and efficiency of task planners [19], or for the autonomous generation of robot goals [18]. A similar approach is proposed in Zender *et al.* [20], where the multi-hierarchical representation is replaced by a single hierarchy ranging from sensor-based maps to a conceptual abstraction, which is encoded in a Web Ontology Language (OWL)–DL ontology defining an office domain. To categorize objects, they rely on a SIFT-based approach, while rooms are grounded according to the objects detected therein. In Nüchter and Hertzberg [58] a constraint network implemented in Prolog is used to both codify the properties and relations among the different planar surfaces in a building (wall, floor, ceiling, and door) and classify them, while two different approaches are considered for object categorization: a SVM-based classifier relying on contour-based features, and a Viola and Jones cascade of classifiers reckoning on range and reflectance data.

These works set out a clear road for the utilization of ontologies to codify semantic knowledge [59], which has been further explored in more recent research. An example of this is the work by Tenorth *et al.* [60], which presents a system for the acquisition, representation, and use of semantic maps called KnowRob-Map, where Bayesian Logic Networks are used to predict the location of objects according to their usual relations. The system is implemented in SWI-Prolog, and the robot's knowledge is represented in an OWL-DL ontology. In this case, the categorization algorithm classifies planar surfaces in kitchen environments as tables, cupboards, drawers, ovens or dishwashers [11]. The same map type and categorization method is employed in Pangercic *et al.* [61], where the authors focus on the codification of object features and functionalities relevant to the robot operation in such environments. The paper by Riazuelo *et al.* [62] describes the RoboEarth cloud semantic mapping which also uses an ontology for codifying concepts and relations, and rely on a Simultaneous Localization and Mapping (SLAM) algorithm for representing the scene geometry and object locations. The categorization method resorts to SURF features (like in Reinaldo *et al.* [63]), and performs by only considering the object types that are probable to appear in a given scene (the room type is known beforehand). In Günther *et al.* [64], the authors employ an OWL-DL ontology in combination with rules defined in the Semantic Web Rule Language (SWRL) to categorize planar surfaces.

It has been also explored the utilization of humans for assisting during the semantic map building process through a situated dialogue. Examples of works addressing this are those by Bastianelli *et al.* [65], Gemignani *et al.* [66], or the aforementioned one by Zender *et al.* [20]. The main motivation of these works is to avoid the utilization of categorization algorithms, given the numerous challenges that they must face. However, they themselves argue that the more critical improvement of their proposals would arise from a tighter interaction with cutting-edge categorization techniques. The interested reader can refer to the survey by Kostavelis and Gasteratos [67] for an additional, comprehensive review of semantic mapping approaches for robotic tasks.

The semantic mapping techniques discussed so far rely on crispy categorizations of the perceived spatial elements, *e.g.* an object is either a cereal box or not, a room is a kitchen or not, etc., which are typically exploited by (logical) reasoners and planners for performing a variety of robotic tasks. As commented before, these approaches: (i) can lead to an incoherent robot operation due to ambiguous grounding results, and (ii) exhibit limitations to fully exploit the contextual relations among spatial elements. In this work we propose a solution for probabilistic symbol grounding to cope with both, the uncertainty inherent to the grounding process, and the contextual relations among spatial elements. Perhaps the closet work to ours is the one by Pronobis and Jensfelt [16], which employs a Chain Graph (a graphical model mixing directed and undirected relations) to model the grounding problem from a probabilistic stance, but that fails at fully exploiting contextual relations. We also present a novel representation called Multiversal Semantic Map ($\mathcal{M}v\mathcal{S}map$), in order to accommodate and further exploit the outcome of the probabilistic symbol grounding.

3. The Multiversal Semantic Map

The proposed *Multiversal Semantic Map* ($\mathcal{M}v\mathcal{S}map$) (see Fig. 1) is inspired by the popular, multi-hierarchical semantic map presented in Galindo *et al.* [25]. This map considers two separated but tightly related hierarchical representations containing: (i) the semantic, meta-information about the domain at hand, *e.g.* refrigerators keep food cold and are usually found in kitchens, and (ii) the factual, spatial knowledge acquired by the robot and its implemented algorithms from a certain workspace, *e.g.* `obj-1` is perceived and categorized as a refrigerator. These hierarchies are called terminological box (*T-Box*) and spatial box (*S-Box*), respectively, names borrowed from the common structure of hybrid knowledge representation systems [68].

$\mathcal{M}v\mathcal{S}maps$ enhance this representation by including uncertainty, in the form of *beliefs*, about the groundings (categorizations) of the spatial elements in the S-Box to concepts in the T-Box. For example, a perceived object, represented by the symbol `obj-1`, could be grounded by the robot as a microwave or a nightstand with beliefs 0.65 and 0.35, respectively, or it might think that a room (`room-1`) is a kitchen or a bedroom with beliefs 0.34 and 0.67. Moreover, in this representation the relations among the spatial elements play a pivotal role, and they have also associated compatibility values in the form of beliefs. To illustrate this, if `obj-1` was found in `room-1`, $\mathcal{M}v\mathcal{S}maps$ can state that the compatibility of `obj-1` and `room-1` being grounded to microwave and kitchen respectively is 0.95, while to microwave and bedroom is 0.05. These belief values are provided by the proposed probabilistic inference process (see Sec. 4.4).

Furthermore, $\mathcal{M}v\mathcal{S}maps$ assign a probability value to each possible set of groundings, creating a *multiverse*, *i.e.* a set of universes stating different explanations of the robot environment. A universe codifies the joint probability of the observed spatial elements being grounded to certain concepts, hence providing a global sense of certainty about the robot's understanding of the environment. Thus, following the previous example, a universe can represent that `obj-1` is a microwave and `room-1` is a kitchen, while a parallel universe states that `obj-1` is a nightstand and `room-1` is a bedroom, both explanations annotated with different probabilities. Thereby, the robot performance is not limited to the utilization of the most probable universe, like traditional semantic maps do, but it can also consider other possible explanations with different semantic interpretations, resulting in a more coherent robot operation.

The next sections introduce the terminological box (Sec. 3.1), the spatial box (Sec. 3.2), and the multiverse (Sec. 3.3) in more detail, as well as the formal definition of $\mathcal{M}v\mathcal{S}maps$ (Sec. 3.4).

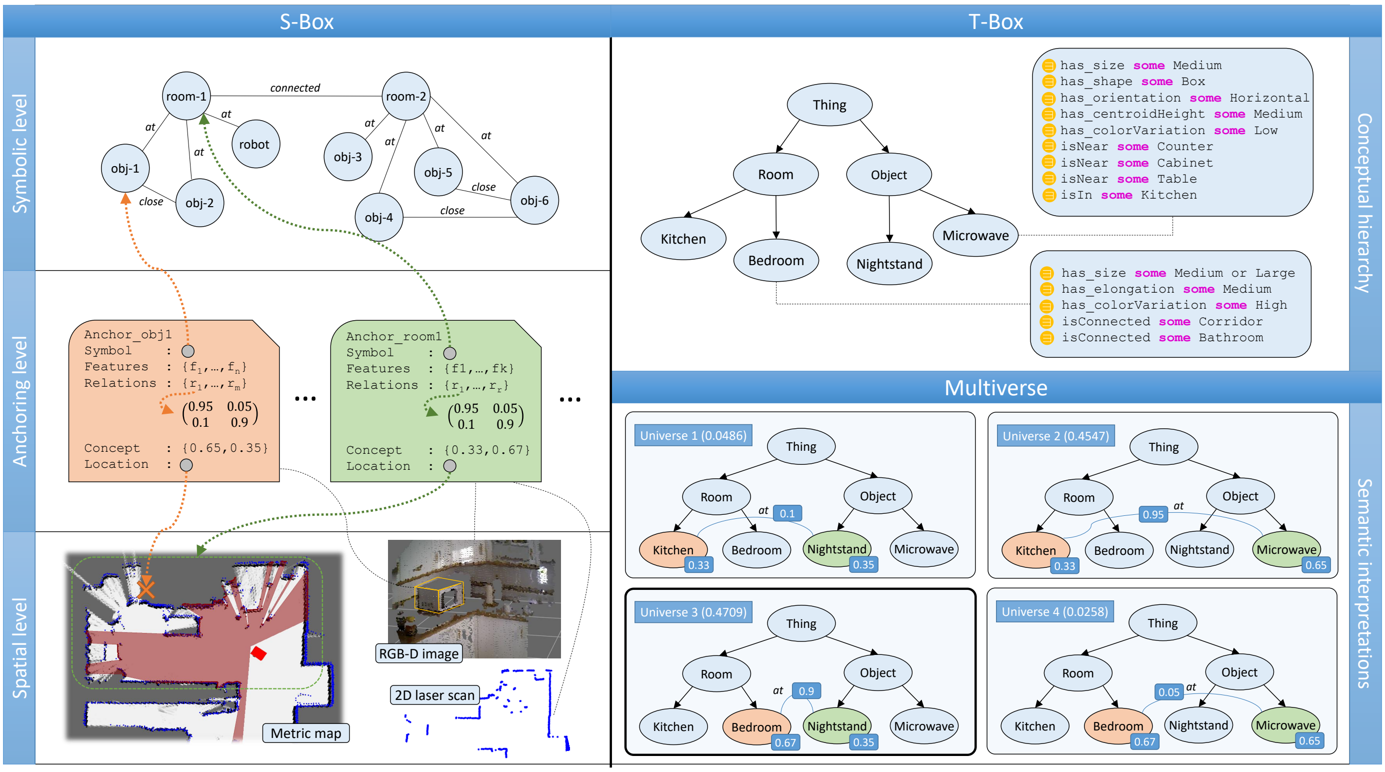

Figure 1: Example of Multiversal Semantic Map representing a simple domestic environment.

In its turn, Sec. 4 describes how a *MvSmap* for a given robot workspace is built from scratch.

3.1. Representing semantic knowledge: the T-Box

The terminological box, or T-Box, represents the semantic knowledge of the domain where the robot is to operate, modeling relevant information about the type of elements that can be found there. Semantic knowledge has been traditionally codified as a hierarchy of concepts (*e.g.* `Microwave` *is-a* `Object` or `Kitchen` *is-a* `Room`), properties of that concepts (`Microwave` *hasShape* `Box`), and relations among them (`Microwave` *isIn* `Kitchen`). This hierarchy is often called *ontology* [26], and its structure is a direct consequence of its codification as a taxonomy. The T-Box gives meaning to the percepts in the S-Box through the grounding of their symbolic representations to particular concepts. For example, a segmented region of a RGB-D image, symbolized by `obj-1`, can be grounded to an instance of the concept `Microwave`.

The process of obtaining and codifying semantic knowledge can be tackled in different ways. For example, web mining knowledge acquisition systems can be used as mechanisms to obtain information about the domain of discourse [69]. Available common-sense Knowledge Bases, like ConceptNet [70] or Open Mind Indoor Common Sense [71], can be also analyzed to retrieve this information. Another valuable option is the utilization of internet search engines, like Google's image search [72], or image repositories like Flickr [73], for extracting knowledge from user-uploaded information. In this work we have codified the semantic knowledge through a human elicitation process, which supposes a truly and effortless encoding of a large number of concepts and relations between them. In contrast to online search or web mining-engine based methodologies, this source of semantic information (a person or a group of people) is trustworthy, so there is less uncertainty about the validity of the information being managed. Moreover, the time required by this approach is usually tractable, as reported in [52], although it strongly depends on the complexity of domain at hand. For highly complex domains the web mining approach – under human supervision – could be explored.

The left part of the T-Box in Fig. 1 depicts an excerpt of the ontology used in this work, defining rooms and objects usually found at homes. The top level sets the root, abstract concept `Thing`, with two children grouping the two types of elements that we will consider during the building of the map, namely `Rooms` and `Objects`. Rooms can belong to different concepts like `Kitchen`, `Bedroom`, etc., while examples of types of objects are `Microwave`, `Nightstand`, etc. The right part of the T-Box illustrates the simplified definitions of the concepts `Bedroom` and `Microwave`, codifying some of their properties and relations with other concepts.

3.2. Modeling space: the S-Box

The spatial box (S-Box) contains factual knowledge from the robot workspace, including the morphology and topology of the space, geometric/appearance information about the perceived spatial elements, symbols representing those elements, and beliefs concerning their grounding to concepts in the T-Box. The S-Box also adopts a hierarchical structure, ranging from sensory-like knowledge at the ground level to abstract symbols at the top one (see S-Box in Fig. 1). This representation is the common choice in the robotics community when dealing with large environments [74].

At the bottom of this hierarchy is the *spatial level*, which builds and maintains a metric map of the working space. $\mathcal{M}v\mathcal{S}maps$ do not restrict the employed metric map to a given one, but any geometric representation can be used, *e.g.* point-based [75], feature-based [76], or occupancy grid maps [1]. This map permits the robot to self-localize in a global frame, and also to locate the perceived elements in its workspace.

The top level of the S-Box is the *symbolic level*, envisioned to maintain an abstract representation of the perceived elements through *symbols*, including the robot itself (*e.g.* `obj-2`, `room-1`, `robot-1`, etc.), which are modeled as nodes. Arcs between nodes state different types of relations, as for example, objects connected by a relation of proximity (see *close* relations in the *symbolic level* in Fig. 1), or an object and a room liked by a relation of location (*at* relations). In this way, the symbolic level constitutes a topological representation of the environment, which can be used for global navigation and task planning purposes [77].

Finally, the intermediate level maintains the nexus between the S-Box and the T-Box. This level stores the outcome of an *anchoring process*, which performs the critical function of creating and maintaining the correspondence between percepts of the environment and symbols that refer to the same physical elements [14, 78]. The result is a set of the so-called *anchors*, which keep geometric/appearance information about the percepts (location, features, relations, etc.) and establish links to their symbolic representation. Additionally, in a $\mathcal{M}v\mathcal{S}map$ anchors are in charge of storing the beliefs about the grounding of their respective symbols, as well as their compatibility with respect to the grounding of related elements.

For illustrative purposes, the middle level in Fig. 1 exemplifies two anchors storing information of a percept from a microwave (in orange) and from a kitchen (in green). The coloured doted lines are pointers to their location in the metric map and their associated symbols, while the black doted lines point at the percepts of these elements from the environment. As an example, the outcome of a symbol grounding process is shown (field `Concept` within the anchor), which gives a belief for `obj-1` being grounded to `Microwave` and `Nightstand` of 0.65 and 0.35 respectively, while those for `room-1` are 0.33 for `Kitchen` and 0.67 for `Bedroom`. It is also shown the beliefs, or compatibility, for the symbols `obj-1` and `room-1` (related through the connection r_1) being grounded to certain pairs of concepts, *e.g.* 0.95 for `Microwave` and `Kitchen`, while 0.05 for `Microwave` and `Bedroom`.

3.3. Multiple semantic interpretations: the Multiverse

$\mathcal{M}v\mathcal{S}maps$ define the possible sets of symbols' groundings as *universes*. For example, by considering only the elements represented by `obj-1` and `room-1` in Fig. 1, four universes are possible: U_1:{(`obj-1` *is-a* `Nightstand`), (`room-1` *is-a* `Kitchen`)}, U_2:{(`obj-1` *is-a* `Microwave`), (`room-1` *is-a* `Kitchen`)}, U_3:{(`obj-1` *is-a* `Nightstand`), (`room-1` *is-a* `Bedroom`)}, and U_4:{(`obj-1` *is-a* `Microwave`), (`room-1` *is-a* `Bedroom`)}. This multiverse considers the possible explanations to the elements in the robot workspace. Additionally, $\mathcal{M}v\mathcal{S}maps$ annotate universes with their probability of being the plausible one, computed as the joint probability of grounding the symbols to the different concepts, giving a measure of certainty abut the current understanding of the robot about its workspace. Thus, a universe can be understood as an instance of the codified ontology with a set of grounded symbols and annotated probabilities.

To highlight the importance of the multiverse, let's us consider the simplified scenario depicted in Fig. 1. Under the title *Multiverse*, the four possible universes are displayed, with their probabilities annotated in brackets along with their names. The coloured (green and orange) concepts in those universes state the symbols that are grounded to them. We can see how the most plausible universe, *i.e.* , combination of groundings, is *Universe 3 (U_3)* (represented with a bold border), which sets `obj-1` as a nightstand and `room-1` as a bedroom. Suppose now that the robot is commanded to store a pair of socks in the nightstand. If the robot relies only on the most probable universe, we could end up with our socks heated in the microwave. However, if the robot also considers other universes, it could be aware that *Universe 2 (U_2)* is also a highly probable one, considering it as a different interpretation of its knowledge. In this case the robot should disambiguate both understandings of the workspace by, for example, gathering additional information from the environment, or in collaboration with humans.

It is worth mentioning that the information encoded in the Multiverse can be exploited, for example, by probabilistic conditional planners (*e.g.* those in [27] or [28]) for achieving a more coherent robot operation. Also, when a certain universe reaches a high belief, it could be considered as the ground, categorical truth, hence enabling the execution of logical inference engines like Pellet [79], FaCT++ [80], or Racer [81].

3.4. Formal description of $\mathcal{M}v\mathcal{S}maps$

Given the ingredients of $\mathcal{M}v\mathcal{S}maps$ provided in the previous sections, a *Multiversal Semantic Map* can be formally defined by the quintuple $\mathcal{M}v\mathcal{S}map = \{\mathcal{R}, \mathcal{A}, \mathcal{Y}, \mathcal{O}, \mathcal{M}\}$, where:

- $\mathcal{R}$ is the metric map of the environment, providing a global reference frame for the observed spatial elements.
- $\mathcal{A}$ is a set of anchors internally representing such spatial elements, and linking them with the set of symbols in $\mathcal{Y}$.
- $\mathcal{Y}$ is the set of symbols that represent the spatial elements as instances of concepts from the ontology $\mathcal{O}$.
- $\mathcal{O}$ is an ontology codifying the semantic knowledge of the domain at hand.
- $\mathcal{M}$ encodes the multiverse, containing the set of universes.

Notice that the traditional T-Box and S-Box are defined in a $\mathcal{M}v\mathcal{S}map$ by $\mathcal{O}$ and $\{\mathcal{R}, \mathcal{A}, \mathcal{Y}\}$ respectively. Since the robot is usually provided with the ontology $\mathcal{O}$ beforehand, building a

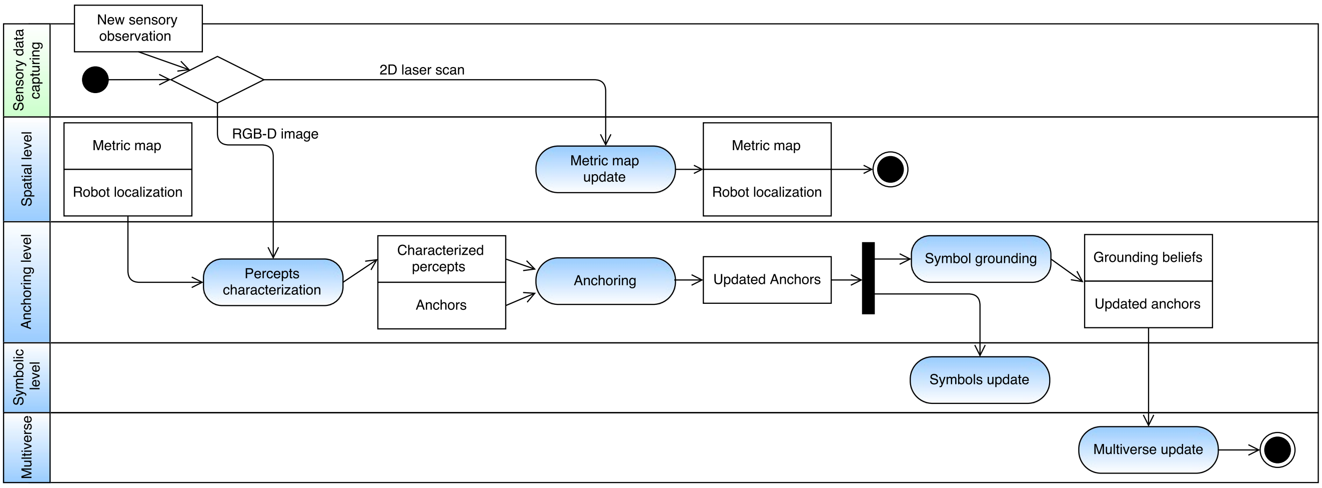

Figure 2: UML activity diagram illustrating the pipeline for the building and maintaining of a $\mathcal{M}v\mathcal{S}map$ according to the sensory information gathered during the robot exploration. Blue rounded boxes are processes, while white shapes stand for consumed/generated data. The processes or data related to the same component of the semantic map are grouped together.

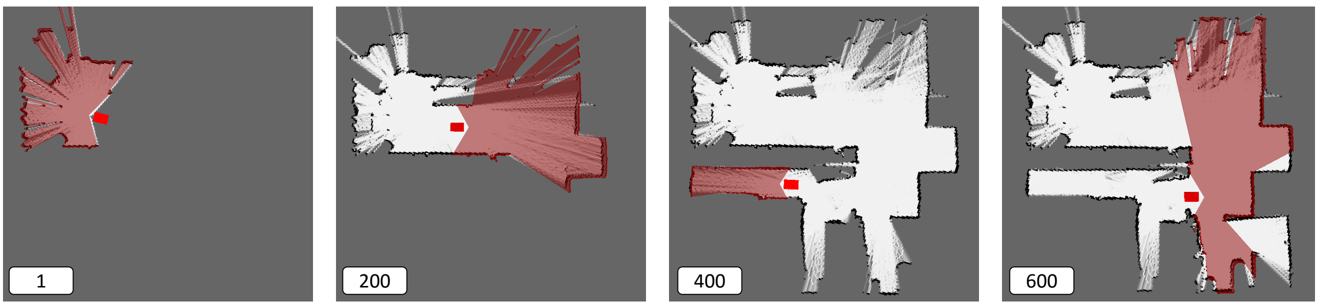

Figure 3: Example of the progressive building of an occupancy grid map from a home environment. The 2D laser scans in red are the scans currently being aligned with the map, while the red boxes represent the estimated robot location. White cells in the map stand for free space, while black ones are occupied areas. Grey cells represent unknown space. Quantities in boxes are the number of scans registered so far to build the corresponding map.

$\mathcal{M}v\mathcal{S}map$ consists of creating and maintaining the remaining elements in the map definition, as described in the next section.

4. Building the Map

This section describes the processes involved in the building of a $\mathcal{M}v\mathcal{S}map$ for a given environment according to the sensory information gathered by a mobile robot (see Fig. 2). In our discussion, we assume that the robot is equipped with a 2D range laser scanner and a RGB-D camera, two sensors commonly found in robotic platforms, although they could be replaced by any other sensory system able to survey the spatial elements in the environment.

In a nutshell, when a new 2D laser scan is available, it triggers the update of the 2D metric map $\mathcal{R}$ in the *spatial level* (see Sec. 4.1). In its turn, if a new RGB-D observation is collected, it is processed in order to characterize the percepts of the surveyed room and the objects therein, as well as their contextual relations (see Sec. 4.2). The characterized percepts fed an anchoring process that compares them with those from previously perceived elements, which are stored in the form of *anchors* in the *anchoring level* (see Sec. 4.3). When a percept is matched with a previous one, its corresponding anchor is updated, otherwise a new anchor, including a new symbol in the *symbolic level*, is created. Finally, the information encoded in the *anchoring level* is used to build a Conditional Random Field, which is in charge of grounding the symbols of the spatial elements to concepts in the T-Box, also providing a measure of the uncertainty concerning such groundings in the form of beliefs (see Sec. 4.4). These beliefs are stored in the anchors, and are employed to update the multiverse $\mathcal{M}$. The next sections describe the core processes of this pipeline in detail.

4.1. Building the underlying metric map

During the robot exploration, the collected 2D laser scans are used to build a metric representation of the environment in the form of an occupancy grid map [1]. For that, we rely on standard Simultaneous Localization and Mapping (SLAM) techniques to jointly build the map and estimate the robot pose [82].

Thus, the building process is based on an Iterative Closet Point (ICP) algorithm [83], which aligns each new scan to the current reference map. Once aligned, the scan measurements are inserted into the map, hence building it incrementally. Given that the robot is also localized in the map at any moment, the spatial information coming from the sensors mounted on it (*e.g.* RGB-D cameras) can be also located. For that, those sensors have to be extrinsically calibrated, that is, the sensors' position in the robot local frame must be known. Fig. 3 shows an example of the incremental building of a metric map from an apartment in the Robot@Home dataset [29].

4.2. Characterizing percepts

Concurrently with the metric map building, when a RGB-D observation is collected it is processed in order to characterize the percepts of the spatial elements therein. This information is required by the posterior anchoring process, so it can decide which percepts correspond to elements previously observed and which ones are perceived for the first time, being consequently incorporated to the semantic map.

Typically, a RGB-D observation contains a number of percepts corresponding to objects, while the whole observation itself corresponds to the percept of a room (see Fig. 6-left). On the one hand, objects' percepts are characterized through geometric (planarity, linearity, volume, etc.) and appearance features (*e.g.* hue, saturation, and value means). On the other hand, room percepts are prone to not cover the entire room, *i.e.* it is common to not survey the whole room with a single RGB-D observation, so the extracted geometric and appearance features (footprint, volume, hue, saturation and value histograms, etc.) are, in addition, averaged over time by considering those from past room percepts. Moreover, the metric map hitherto built for that room is also considered and characterized, since it supposes a rich source of information for its posterior categorization [38]. The upper part of Tab. 1 lists the features used to describe those percepts.

In addition to objects and rooms, the contextual relations among them are also extracted and characterized. We have considered two types of relationships, one linking objects that are placed closer than a certain distance (*close*), and another one relating an object and its container room (*at*). The lower part of Tab. 1 lists the features employed to characterize such relations. It is worth mentioning the function of the *bias* feature characterizing the object–room relations, which is a fixed value that permits the CRF to automatically learn the likelihood of finding a certain object type into a room of a certain category (see Sec. 4.4.1). The outcome of this characterization process is known as the *signature* of the percept.

4.3. Modeling and keeping track spatial elements: Anchoring

Once characterized, the percepts feed an *anchoring process* [14], which establishes the correspondences between the symbols of the already perceived spatial elements (*e.g.* `obj-1` or `room-1`) and their percepts. For that, it creates and maintains internal representations, called anchors, which include: the features of the spatial elements and their relations, their geometric location[2], their associated symbols, the beliefs about

Table 1: Features used to characterize the percepts (objects and rooms) and contextual relations among them (object-object and object-room). These features are grouped according to their type, geometric or appearance, stating in parentheses the type of information from where they come, RGB-D images or metric maps. Values in parentheses in the features' names give the number of features grouped under the same name (for example the centroid of an object has x, y and z coordinates).

Object	**Room**
Geometric (RGB-D)	*Geometric (RGB-D)*
Planarity	Scatter (2)
Scatter	Footprint (2)
Linearity	Volume (2)
Min. height	*Appearance (RGB-D)*
Max. height	H, S, V, means (6)
Centroid (3)	H,S,V, Stdv. (6)
Volume	H, S, V, histograms (30)
Biggest area	*Geometric (Metric map)*
Orientation	Elongation
Appearance (RGB-D)	Scatter
H, S, V, means (3)	Area
H, S, V, Stdv. (3)	Compactness
H, S, V, histograms (15)	Linearity

Object-Object	**Object-Room**
Geometric (RGB-D)	Bias
Perpendicularity	
Vertical distance	
Volume ratio	
Is on relation	
Appearance (RGB-D)	
H, S, V, mean diff.	
H, S, V, Stdv. diff.	

the groundings of those symbols, and their compatibility with the groundings of related elements. The content of an anchor was previously illustrated in the *anchoring level* in Fig. 1. In its turn, the sub-components of the anchoring process are depicted in Fig. 4.

Let $\mathcal{S}_{in} = \{s_1, \ldots, s_n\}$ be the set of characterized percepts surveyed in the last RGB-D observation. Then, the signatures of these percepts are compared with those of anchors already present in the semantic map, which produces two disjoint sets: the set $\mathcal{S}_{update}$ of percepts of spatial elements that have been previously observed in the environment, and the set $\mathcal{S}_{new}$ of percepts of elements detected for the first time. We have considered a simple but effective matching algorithm that checks the location of two percepts, the overlapping of their bounding boxes, and their appearance to decide if they refer to the same physical element.

The two sets of percepts resulting from the matching step are processed differently: while the set $\mathcal{S}_{update}$ triggers the update of their associated anchors, *i.e.* their locations, features, and relations are revised according to the new available information, the set $\mathcal{S}_{new}$ produces the creation of new anchors. As a consequence, the content of the *symbolic level* is also revised: the symbols representing updated anchors are checked for possible changes in their relations, while new symbols are created for the new anchors. As an example, Fig. 5 shows two point clouds

[2] Notice that although the underlying metric map is 2D, the extrinsic calibration of sensors can be used to locate an element in 6D (3D position and 3D orientation).

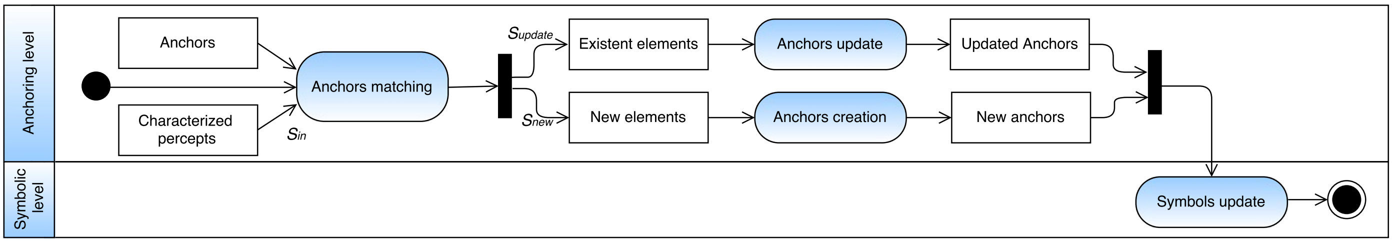

Figure 4: UML activity diagram showing the sub-processes (blue rounded boxes) and consumed/produced data (white shapes) involved in the anchoring process.

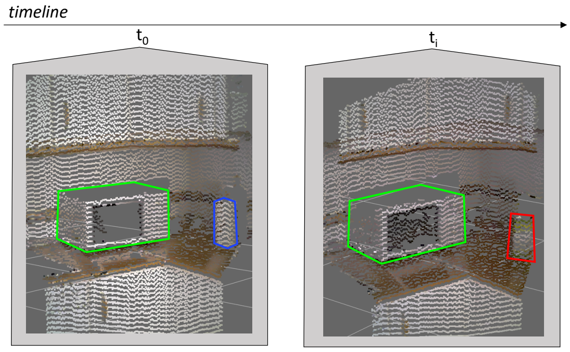

Figure 5: Example of the matching step within the anchoring process, showing two point clouds gathered from a kitchen at different time instants. The green shapes contain percepts that are matched as belonging to the same spatial element, while the percepts enclosed in the blue and red ones have been correctly considered as corresponding to different elements due to their different appearance (they contain a paper roll and a milk bottle respectively).

representing RGB-D images gathered from the same kitchen at different time instants. At time t_0, two new anchors are created for accommodating the information from the two percepts (highlighted in green and blue). Then, at time t_1, the signature of the percept in green is matched with the one with the same color at t_0, while the percept in red, despite their similar location and size, is considered different from the one in blue at t_0 due to their appearance, and a new anchor is created. Notice that to complete the aforementioned content of anchors the beliefs about the grounding of their symbols, as well as the compatibility with the groundings of related elements, must be computed. This is carried out by the probabilistic techniques in the next section.

Although the described anchoring process could appear similar to a tracking procedure, it is more sophisticated regarding the information that is stored/managed. For example, in typical tracking problems, it is usually not needed to maintain a symbolic representation of their tracks, nor to ground them to concepts within a knowledge base. Further information in this regard can be found in the work by Coradeschi and Saffiotti [14].

4.4. Probabilistic symbol grounding

We holistically model the symbol grounding problem employing a Conditional Random Field (CRF) (see Sec. 4.4.1), a probabilistic technique first proposed by Lafferty *et al.* [84] that, in addition to exploiting the relations among objects and rooms, also provides the beliefs about such groundings through a probabilistic inference process (see Sec. 4.4.2). These belief values are the main ingredients for the generation and update of the multiverse in the $\mathcal{M}v\mathcal{S}map$ (see Sec. 4.5).

4.4.1. CRFs to model the symbol grounding problem

The following definitions are required in order to set the problem from this probabilistic stance:

- Let $\boldsymbol{s} = [s_1, .., s_n]$ be a vector of n of spatial elements, stating the observed objects or rooms in the environment, which are characterized by means of the features in their associated anchors.
- Define $L_o = \{l_{o_1}, .., l_{o_k}\}$ as the set of the k considered object concepts (*e.g.* `Bed`, `Oven`, `Towel`, etc.).
- Let $L_r = \{l_{r_1}, .., l_{r_j}\}$ be the set of the j considered room concepts (*e.g.* `Kitchen`, `Bedroom`, `Bathroom`, etc.).
- Define $\boldsymbol{y} = [y_1, .., y_n]$ to be a vector of discrete random variables assigning a concept from L_o or L_r to the symbol associated with each element in $\boldsymbol{s}$, depending on whether such symbol represents an object or a room.

Thereby, the grounding process is jointly modeled by a CRF through the definition of the probability distribution $P(\boldsymbol{y} \mid \boldsymbol{s})$, which yields the probabilities of the different assignments to the variables in $\boldsymbol{y}$ conditioned on the elements from $\boldsymbol{s}$. Since its exhaustive definition is unfeasible due to its high dimensionality, CRFs exploit the concept of independence to break this distribution down into smaller pieces. Thus, a CRF is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of nodes $\mathcal{V}$ models the random variables in $\boldsymbol{y}$, and the set of undirected edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ links contextually related nodes. Notice that this graph can be built directly from the codified information within the *symbolic level*. Thus, mimicking the representation in that level, the same types of edges are considered in the CRF: proximity of two objects, and presence of an object into a room. Intuitively, this means that, for a certain object, only the nearby objects in the environment and its container room have a direct influence on its grounding, while the grounding of a room is affected by the objects therein. Fig. 6-right shows an example of a CRF graph built from the spatial elements in the observation depicted in Fig. 6-left, also including elements that were perceived in previous observations of the same room and were stored in the S-Box.

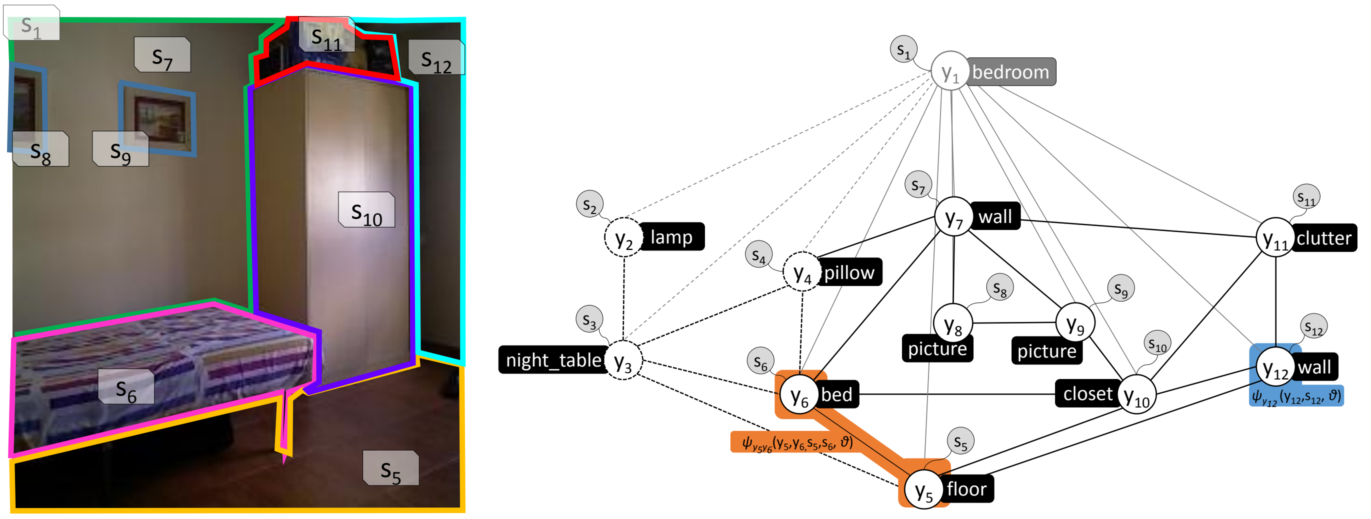

Figure 6: Left, RGB image from a RGB-D observation of a sequence where the robot is exploring a bedroom. The objects' percepts are enclosed in coloured shapes and represented by s_5-s_{12}, while the whole image is considered the room percept and is represented by s_1. Right, CRF graph representing the spatial elements and relations in such image as random variables and edges respectively (solid lines), as well as the elements and relations from previously surveyed objects (doted lines, represented as $s_2 - s_4$). The area highlighted in blue states the scope of an unary factor, while the one in orange stands for the scope of a pairwise factor.

According to the Hammersley-Clifford theorem [85], the probability $P(\boldsymbol{y} \mid \boldsymbol{s})$ can be factorized over the graph $\mathcal{G}$ as a product of *factors* $\psi(\cdot)$:

$$p(\boldsymbol{y}|\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{s},\boldsymbol{\theta})} \prod_{c \in C} \psi_c(y_c, s_c, \boldsymbol{\theta}) \tag{1}$$

where C is the set of maximal cliques[3] of the graph $\mathcal{G}$, and $Z(\cdot)$ is the also called partition function, which plays a normalization role so $\sum_{\xi(y)} p(\boldsymbol{y}|\boldsymbol{s};\boldsymbol{\theta}) = 1$, being $\xi(\boldsymbol{y})$ a possible assignment to the variables in $\boldsymbol{y}$. The vector $\boldsymbol{\theta}$ stands for the model parameters (or weights) to be tuned during the training phase of the CRF. Factors can be considered as functions encoding pieces of $P(\boldsymbol{y} \mid \boldsymbol{s})$ over parts of the graph. Typically, two kind of factors are considered: *unary factors* $\psi_i(y_i, s_i, \boldsymbol{\theta})$, which refer to nodes and talk about the probability of a random variable y_i belonging to a category in L_o or L_r, and *pairwise factors* $\psi_{ij}(y_i, y_j, s_i, s_j, \boldsymbol{\theta})$ that are associated with edges and state the compatibility of two random variables (y_i, y_j) being tied to a certain pair of categories. As a consequence, the cliques used in this work have at most two nodes (see Fig. 6-right). The expression in Eq.1 can be equivalently expressed for convenience through log-linear models and exponential families as [86]:

$$p(\boldsymbol{y}|\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{s},\boldsymbol{\theta})} \prod_{c \in C} \exp(\langle \phi(s_c, y_c), \boldsymbol{\theta} \rangle) \tag{2}$$

being $\langle \cdot, \cdot \rangle$ the inner product, and $\phi(s_c, y_c)$ the sufficient statistics of the factor over the clique c, which comprises the features extracted from the spatial elements (recall Tab. 1). Further information about this representation can be found in [55].

[3] A maximal clique is a fully-connected subgraph that can not be enlarged by including an adjacent node.

Training a CRF model for a given domain requires the finding of the parameters in $\boldsymbol{\theta}$, in such a way that they maximize the likelihood in Eq.2 with respect to a certain i.i.d. training dataset $\mathcal{D} = [d^1, \ldots d^m]$, that is:

$$\max_{\boldsymbol{\theta}} \mathcal{L}_p(\boldsymbol{\theta} : \mathcal{D}) = \max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p(\boldsymbol{y}^i \mid \boldsymbol{s}^i; \boldsymbol{\theta}) \tag{3}$$

where each training sample $d^i = (\boldsymbol{y}^i, \boldsymbol{s}^i)$ consists of a number of characterized spatial elements ($\boldsymbol{s}^i$) and the corresponding ground truth information about their categories ($\boldsymbol{y}^i$). If no training dataset is available for the domain at hand, the codified ontology can be used to generate synthetic samples for training, as we have shown in our previous work [51, 55]. The optimization in Eq.3 is also known as Maximum Likelihood Estimation (MLE), and requires the computation of the partition function $Z(\cdot)$, which in practice turns this process into a $\mathcal{NP}$-hard, hence intractable problem. To face this in the present work, the calculus of $Z(\cdot)$ is estimated by an approximate inference algorithm during the training process, concretely the *sum-product* version of the *Loopy Belief Propagation* (LBP) method [56], which has shown to be a suitable option aiming at categorizing objects [23].

4.4.2. Performing probabilistic inference

Once the CRF representation modeling a given environment is built, it can be exploited by probabilistic inference methods to perform different probability queries. At this point, two types of queries are specially relevant: the *Maximum a Posteriori* (MAP) query, and the *Marginal* query. The goal of the MAP query is to find the most probable assignment $\hat{\boldsymbol{y}}$ to the variables in $\boldsymbol{y}$, *i.e.* :

$$\hat{\boldsymbol{y}} = arg \max_{\boldsymbol{y}} p(\boldsymbol{y} \mid \boldsymbol{s}; \boldsymbol{\theta}) \tag{4}$$

Once again, the computation of the partition function $Z(\cdot)$ is needed, but since given a certain CRF graph its value remains constant, this expression can be simplified by:

$$\hat{\boldsymbol{y}} = arg\max_{\boldsymbol{y}} \prod_{c \in C} \exp(\langle \phi(s_c, y_c), \boldsymbol{\theta} \rangle) \tag{5}$$

Nevertheless, this task checks every possible assignment to the variables in $\boldsymbol{y}$, so it is still unfeasible. An usual way to address this issue is the utilization of approximate methods, like the *max-product* version of LBP [87]. The alert reader may think that, in the end, the MAP assignment provides crispy results. Although this is undoubtedly true, the computation of those results considers both the relations among the spatial elements in the environment, and the belief about their belonging to different categories, so it is clearly differentiated from the crispy results given by an off-the-shelf categorization method working on individual elements. The black boxes in Fig. 6-right show an example of the outcome of a MAP query over the defined CRF graph.

In its turn, the Marginal query, which can be performed by the aforementioned *sum-product* version of LBP, provides us the beliefs about the possible groundings. In other words, this query yields the marginal probabilities for each symbol being grounded to different concepts, as well as the compatibility of these groundings with respect to the grounding of contextually related symbols. Therefore, it is also possible to retrieve the probability of a certain assignment to the variables in $\boldsymbol{y}$, which is of interest for managing universes (see Sec. 4.5). Recall that, in a $\mathcal{M}v\mathcal{S}map$, these beliefs are stored in their corresponding anchors for their posterior exploitation during the robot operation (see anchors in Fig. 1). Sec. 5 will show both MAP and Marginal queries in action.

4.5. Managing the Multiverse

To conclude the building of the $\mathcal{M}v\mathcal{S}map$, the outcome of the marginal query is exploited to generate and update the multiverse. The probability for each possible universe can be retrieved by means of Eq.1, replacing the factors $\psi(\cdot)$ by the provided beliefs $b(\cdot)$, and the partition function $Z(\cdot)$ by its approximation $Z_{LBP}(\cdot)$ computed by the LBP algorithm, that is:

$$p(\boldsymbol{y}|\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{Z_{LBP}(\boldsymbol{s}, \boldsymbol{\theta})} \prod_{c \in C} b_c(y_c, s_c) \tag{6}$$

The exhaustive definition of such multiverse, that is, to compute and store the probabilities and groundings in each possible universe, highly depends on the complexity of the domain at hand. The reason for this is that the number of possible universes depends on both, the number of spatial elements, and the number of concepts defined in the ontology. For example, let's suppose a domain with 3 types of rooms and 4 types of objects. During the robot exploration, 5 objects have been observed within 2 rooms, so a total of $4^5 \times 3^2 = 9,216$ possible interpretations, or universes, exist. This is a large number for a small scenario, but it supposes a reduced size in memory since each universe is defined by: (i) its probability, and (ii) its grounded symbols. Concretely, in this case each universe can be codified through a *float* number for its probability (4 bytes) and 7 *char* numbers for the groundings (7 bytes in total, supposing that each concept can be identified by a *char* number as well), so the size of the multiverse is $11 \times 9,216 = 99kB$. Notice that such a size grows exponentially with the number of spatial elements, so in crowded environments this exhaustive definition is unpractical, or even unfeasible.

In those situations, the exhaustive definition can be replaced by the generation of the more relevant universes for a given task and environment. Thus, for example, the MAP grounding yielded by a MAP query permits the definition of the most probable universe. Recall that the probability of this or other universes of interest can be retrieved by inserting their respective groundings and stored beliefs in Eq.6. Other probable universes can be straightforwardly identified by considering the ambiguous groundings. For example, if an object is grounded to concepts with the following beliefs {`Bowl 0.5`, `Milk-bottle 0.45`, `Microwave 0.05`}, and the MAP query grounds it to `Bowl`, it makes sense to also keep the universe where the object is grounded to `Milk-bottle`, and *vice versa*. As commented before, the set of relevant universes is task and domain dependant so, if needed, they should be defined strategies for their generation in order to keep the problem tractable.

To tackle this issue we propose a simple but practical strategy based on the utilization of a threshold, or *ambiguity factor*, that determines when a grounding result is ambiguous. For that, if the ratio between the belief about a symbol being grounded to a certain concept (b_i) and the highest belief for that symbol (b_h) is over this threshold (α), then these two possible groundings are considered ambiguous. Mathematically:

$$ambiguous(b_i, b_h) = \begin{cases} 1 \text{ (true)} & \text{if } b_i/b_h > \alpha \\ 0 \text{ (false)} & \text{otherwise} \end{cases} \tag{7}$$

Therefore, if a pair of grounding values are ambiguous according to this strategy, their associated universes are considered relevant, being consequently stored in the multiverse. Continuing with the previous example, the ratio between the beliefs for `Milk-bottle` and `Bowl` is $0.45/0.5 = 0.9$, while between `Microwave` and `Bowl` is $0.05/0.5 = 0.1$. Thus, with a value for α higher than 0.1 and lower than 0.9, this strategy would consider the first pair of groundings as ambiguous, but not the second one. The efficacy of this strategy for keeping the number of universes low, without disregarding relevant ones, is shown in Sec. 5.3.

5. Experimental Evaluation

To evaluate the suitability of both, the proposed probabilistic symbol grounding as well as the novel semantic map, we have carried out a number of experiments using the challenging Robot@Home [29] dataset, which is briefly described in Sec. 5.1. More precisely, to test the symbol grounding capabilities of our approach (see Sec. 5.2), it has been analyzed its performance both (i) when grounding object and rooms symbols in isolation, *i.e.* using the traditional categorization approach that works with the individual features of each spacial

Table 2: Performance of baseline methods individually grounding objects and rooms. Rows index the results employing features of different nature, while columns index the different methods (CRF: Conditional Random Fields, SVM: Supported Vector Machines, NB: Naive Bayes, DT: Decision Tress, RF, Random Forests, NN: Nearest Neighbors). Please refer to App. A for a description of the used performance metrics.

	CRF			SVM	NB	DT	RF	NN
Objects	Macro p./r.	Micro p.	Micro r.	Macro p./r.	Macro p./r.	Macro p./r.	Macro p./r.	Macro p./r.
Geometric	72.86%	52.12%	42.41%	62.84%	66.67%	71.61%	73.20%	40.69%
Appearance	34.08%	18.50%	14.58%	33.72%	19.07%	25.25%	33.41%	16.39%
Geometric + Appearance	**73.64%**	**53.30%**	**51.62%**	**71.06%**	**70.00%**	**72.38%**	**74.53%**	**43.04%**
Rooms	Macro p./r.	Micro p.	Micro r.	Macro p./r.	Macro p./r.	Macro p./r.	Macro p./r.	Macro p./r.
Geometric (RGB-D)	25.53%	22.92%	18.33%	32.60%	25.00%	7.40%	22.50%	21.40%
Geometric (Metric map)	27.66%	16.25%	17.38%	40.20%	32.10%	**43.80%**	45.30%	29.80%
Geometric (All)	46.81%	36.64%	37.94%	**41.70%**	28.30%	37.90%	**52.50%**	36.10%
Appearance	44.68%	38.43%	35.73%	37.80%	32.60%	22.10%	42.40%	28.90%
Geo. (All) + Appearance	**57.45%**	**50.09%**	**48.12%**	37.40%	**38.20%**	37.90%	37.40%	**44.00 %**

Figure 7: Robotic platform used to collect the Robot@Home dataset.

element (see Sec. 5.2.1), and (ii) when also considering the contextual relations among elements (see Sec. 5.2.2). To conclude this evaluation, we also describe some sample mapping scenarios in Sec. 5.3, aiming to illustrate the benefits of the proposed *MvSmap*.

5.1. Testbed

The Robot@Home dataset provides 83 sequences containing 87,000+ observations, divided into RGB-D images and 2D laser scans, which survey rooms of 8 different types summing up ~1,900 object instances. From this repository we have extracted 47 sequences captured in the most common room types in home environments, namely: bathrooms, bedrooms, corridors, kitchens, living-rooms and master-rooms. These sequences contain ~1,000 instances of objects that belong to one of the 30 object types considered in this work, *e.g.* bottle, cabinet, sink, toilet, book, bed, pillow, cushion, microwave, bowl. etc.

The observations within the sequences come from a rig of 4 RGB-D cameras and a 2D laser scanner mounted on a mobile robot (see Fig. 7). However, to match this sensory configuration with one more common in robotic platforms, we have only considered information from the 2D laser scanner and the RGB-D camera looking ahead.

5.2. Probabilistic symbol grounding evaluation

In this section we discuss the outcome of a number of experiments that evaluate different configurations for the probabilistic symbol grounding process. To obtain the performance measurements (micro/macro precision/recall, see App. A), a *MvSmap* has been built for each sequence, and MAP queries are executed over the resultant CRFs (recall Sec. 4.4). Concretely, a leave-one-out cross-validation technique is followed, where a sequence is selected for testing and the remaining ones for training. This process is repeated 47 times, changing the sequence used for testing, and the final performance is obtained averaging the results yielded by those repetitions.

5.2.1. Individual grounding of object and room symbols

The aim of this section is to evaluate the performance of our proposal without exploring contextual relations, *i.e.* only considering the geometric/appearance features characterizing the symbols. This *individual grounding* is the traditional approach in semantic mapping, and permits us to set a baseline for measuring the real enhancement of the joint grounding in the next section. Thereby, only the nodes in the CRFs have been considered, characterized by the *object* and *room* features in Tab. 1.

The first three columns in Tab. 2 report the results for grounding object and room symbols according to the described configuration. For objects, we can see how the used geometric features are more discriminative than the appearance ones, but their complementary nature makes that the CRFs resorting to their combination achieves the highest results (73.64%). The same happens when grounding rooms, where the winning option, reaching a performance of 57.45%, combines geometric and appearance features from the RGB-D observations, as well as geometric features from the part of the metric map corresponding to the room.

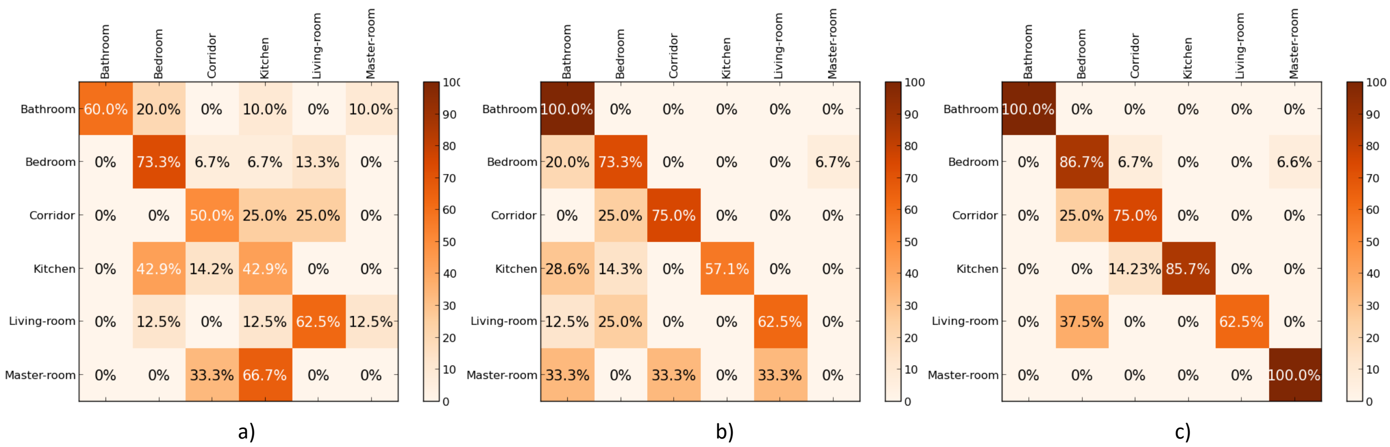

Figure 8: Confusion matrices relating the ground truth information about rooms (rows) with the concept to which they are grounded (columns). a) Confusion matrix for a CRF only employing nodes, b) including object-room relations, and c) considering all the contextual relations.

Table 3: Performance for grounding symbols of CRFs exploiting contextual information. Rows index the type of contextual relations modeled by the CRFs. App. A describes the used metrics.

Objects	Macro p./r.	Micro p.	Micro r.
Object-Object	78.70%	65.58%	53.34%
Object-Room	78.69%	59.38%	53.09%
Object-Object + Object-Room	**81.58%**	**70.71%**	**60.94%**
Rooms	Macro p./r.	Micro p.	Micro r.
Object-Room	80.85%	65.08%	61.33%
Object-Object + Object-Room	**91.49%**	**85.25 %**	**84.98%**

Table 4: Example of the outcome of a grounding process where the contextual relations modeled in a CRF help to disambiguate wrong individual groundings. The first column states the symbols' names, the second one their ground truth category, while the third and fourth columns report the two categories that received the highest beliefs (in parentheses) after a Marginal inference query. The MAP assignment is highlighted in bold.

Symbol	Ground truth	Beliefs	
`obj-3`	Microwave	**Microwave (0.38)**	Nightstand (0.29)
`obj-5`	Counter	Table (0.39)	**Counter (0.30)**
`obj-9`	Counter	**Counter (0.26)**	Table (0.12)
`room-1`	Kitchen	Bedroom (0.49)	**Kitchen (0.22)**

To complete this baseline, they have been also evaluated some of the most popular classifiers also resorting to individual object/room features. In order to make this comparison as fair as possible the same features employed for the CRFs have been used, as well as the same leave-one-out cross-validation approach. Concretely, we have resorted to the implementation in the `scikit-learn` library [88] of the following widely-used methods[4]: Supported Vector Machines, Naive Bayes, Decision Trees, Random Forests, and Nearest Neighbors. The yielded results are reported in the last five columns of Tab. 2, where it is shown how the CRF achieve a similar or even higher success than those classifiers. In fact, the more serious competitor is the one based on Random Forests, which achieves a $\sim$ 1% higher success when categorizing objects, but a $\sim$ 5% lower one when dealing with rooms.

5.2.2. *Joint object-room symbol grounding*

This section explores how the progressive inclusion of different types of contextual relations to the CRFs affects the performance of the grounding method. Tab. 3 gives the figures obtained from this analysis. Taking a closer look at it, we can see how the inclusion of contextual relations among objects increases the success of grounding them by $\sim$ 5%. By only considering relations among objects and rooms, the performance of grounding objects is increased almost the same percentage, while the success of rooms considerably grows from 57.45% up to 80.91%. Finally, with the inclusion of all the contextual relations, the reached grounding success is of 81.58% and 91.49% for objects and rooms respectively. Comparing these numbers with the baseline performance obtained in the previous section also employing CRFs, they achieve a notorious increment in the performance of $\sim$ 8% for objects and $\sim$ 34% for rooms. This approach also clearly outperforms the success reported by the other methods in Tab. 2.

Fig. 8 depicts the confusion matrices obtained while grounding room symbols for each of the aforementioned configurations. In these matrices, the rows index the room ground truth, while the columns index the grounded concept. We can notice how the performance reported in these matrices improves progressively (the values in their diagonals grow) with the inclusion of contextual relations.

To further illustrate the benefits of the conducted joint symbol grounding, Tab. 4 shows the results of the grounding of a number of symbols from a kitchen sequence. The third and fourth columns of this table report the concepts with the two highest beliefs for each symbol, retrieved by a Marginal infer-

[4]Further information about these classifiers can be found in the library webpage: http://scikit-learn.org/

Table 5: Example of grounding results yielded by the proposed method for the symbols within a simple kitchen scenario. The first and the second columns give the symbols' names and their ground truth respectively, while the remaining columns report the five categories with the highest beliefs (in parentheses) as yielded by a Marginal inference query. The MAP assignment is highlighted in bold.

Symbol	Ground truth	Beliefs					
`obj-1`	Microwave	**Nightstand (0.46)**	Microwave (0.42)	Wall (0.06)	Bed (0.04)	Counter (0.04)	Floor(0.1)
`obj-2`	Counter	**Counter (0.70)**	Bed (0.24)	Floor (0.04)	Wall (0.01)	Nightstand (0.01)	Microwave (0.0)
`obj-3`	Wall	**Wall (0.99)**	Counter (0.1)	Nightstand (0.0)	Floor (0.0)	Microwave (0.0)	Bed (0.0)
`obj-4`	Wall	**Wall (0.99)**	Bed (0.01)	Microwave (0.0)	Nightstand (0.0)	Floor (0.0)	Counter (0.0)
`obj-5`	Floor	**Floor (0.99)**	Bed (0.01)	Wall (0.0)	Counter (0.0)	Nightstand (0.0)	Microwave (0.0)
`room-1`	Kitchen	**Bedroom (0.51)**	Kitchen (0.22)	Bathroom (0.19)	Living-room (0.06)	Master-roomr (0.01)	Corridor (0.01)

ence query over the CRF built from such sequence. A traditional grounding approach would only consider the concepts in the third row, while our holistic stance is able to provide the results highlighted in bold (through a MAP query), which match the symbols' ground truth.

5.3. Sample mapping scenarios

In this section we exemplify the building of $\mathcal{M}v\mathcal{S}maps$ for two scenarios exhibiting different complexity. We start by describing a simple scenario where the possible object categories are: floor, wall, counter, bed, nightstand, and microwave. The possible room categories are the same as in the previous section. This is an extension in a real setting of the toy example described in Sec. 3. The chosen sequence of observations from Robot@Home corresponds to a kitchen containing 5 objects of these categories: a counter, a microwave, two walls and the floor. Thus, the $\mathcal{M}v\mathcal{S}map$ built for that scenario consist of (recall Sec. 3.4):

- An occupancy grid map of the explored room.
- 6 anchors representing the spatial elements (5 objects and a room).
- 6 symbols in the symbolic level.
- An ontology of the home domain.
- $6^5 \times 6^1 = 46,656$ possible universes, which supposes a multiverse size of $\sim 456kB$.

Tab. 5 shows the grounding results yielded by the execution of MAP and Marginal queries over the CRF representation of such map. We can see how the MAP assignment fails at grounding the symbols `obj-1` and `room-1`, but the right groundings of such symbols also receive a high belief value. As a consequence of this, their respective universes could also exhibit high probabilities, hence the importance of their consideration. Notice that the size of the multiverse could be further reduced by applying the previously proposed strategy. For example, considering an ambiguity factor of $\alpha = 0.2$, the number of possible universes is 12, being the size (in memory) of the multiverse of only 132 bytes.

We also describe a more complex scenario considering the room and object categories introduced in Sec. 5.1. In this case, we discuss the progressive building of the $\mathcal{M}v\mathcal{S}map$ at 4 different time instants during the robot exploration of a bedroom. Fig. 9 depicts the evolution of the groundings of the spatial elements perceived by the robot during such exploration, where the big and small coloured boxes represent the groundings with the two highest beliefs. In this case, the groundings provided by MAP queries match with those showing the highest beliefs.

We can see how until the time instant t_1 the robot surveyed 8 objects, being so confident about the category of 5 of them. This supposes a total of 9 anchors and 9 symbolic representations (8 objects plus a room). The most ambiguous result is for an object placed on the bed, which is in fact a towel. This ambiguity is due to the features exhibited by the object, its position, and its unusual location in a bedroom. In its turn, the belief about the room being grounded to the `Bedroom` concept is high, 0.76, as a result of the surveyed spatial elements and their relations. Until time t_2 the room is further explored, appearing three new objects: a chair, a table and a wall, hence adding 3 new anchors and their respective symbols to the $\mathcal{M}v\mathcal{S}map$. The surveyed table is the only one showing an ambiguous grounding because of its features and few contextual relations. However, in the observations gathered until the time instant t_3, two new objects are perceived on top of the table, a book and a bottle, increasing the belief value about its grounding to the `Table` concept. With these new objects and relations the uncertainty about the category of the room also decreases. Finally, considering all the information gathered until the time instant t_4, where a pillow has been observed on top of the bed, the belief about the room category increases up to 0.99. Notice how the detection of such pillow also decreases the uncertainty about the grounding of the bed. The *modus operandi* of *traditional* semantic maps is to consider the towel on the bed as a book, which can lead to, for example, the failure of a robot ordered to bring all the towels in the house to the bathroom. This can be tackled through the utilization of $\mathcal{M}v\mathcal{S}maps$ and the clarification of uncertain groundings.

Thereby, the $\mathcal{M}v\mathcal{S}map$ built in this scenario is compounded of 15 anchors (14 objects plus a room), 15 symbols at the symbolic level, and a total of $30^{14} \times 6^1 \simeq 2.8 \times 10^{21}$ universes. This supposes a multiverse with an intractable size, however, applying the previous strategy where only uncertain results generate new universes, the size of the multiverse is considerably reduced to 40 universes and 760 bytes.

6. Potential Applications of Multiversal Semantic Maps

The main purpose of the proposed $\mathcal{M}v\mathcal{S}map$ is to provide a mobile robot with a probabilistic, rich representation of its environment, empowering the efficient and coherent execution

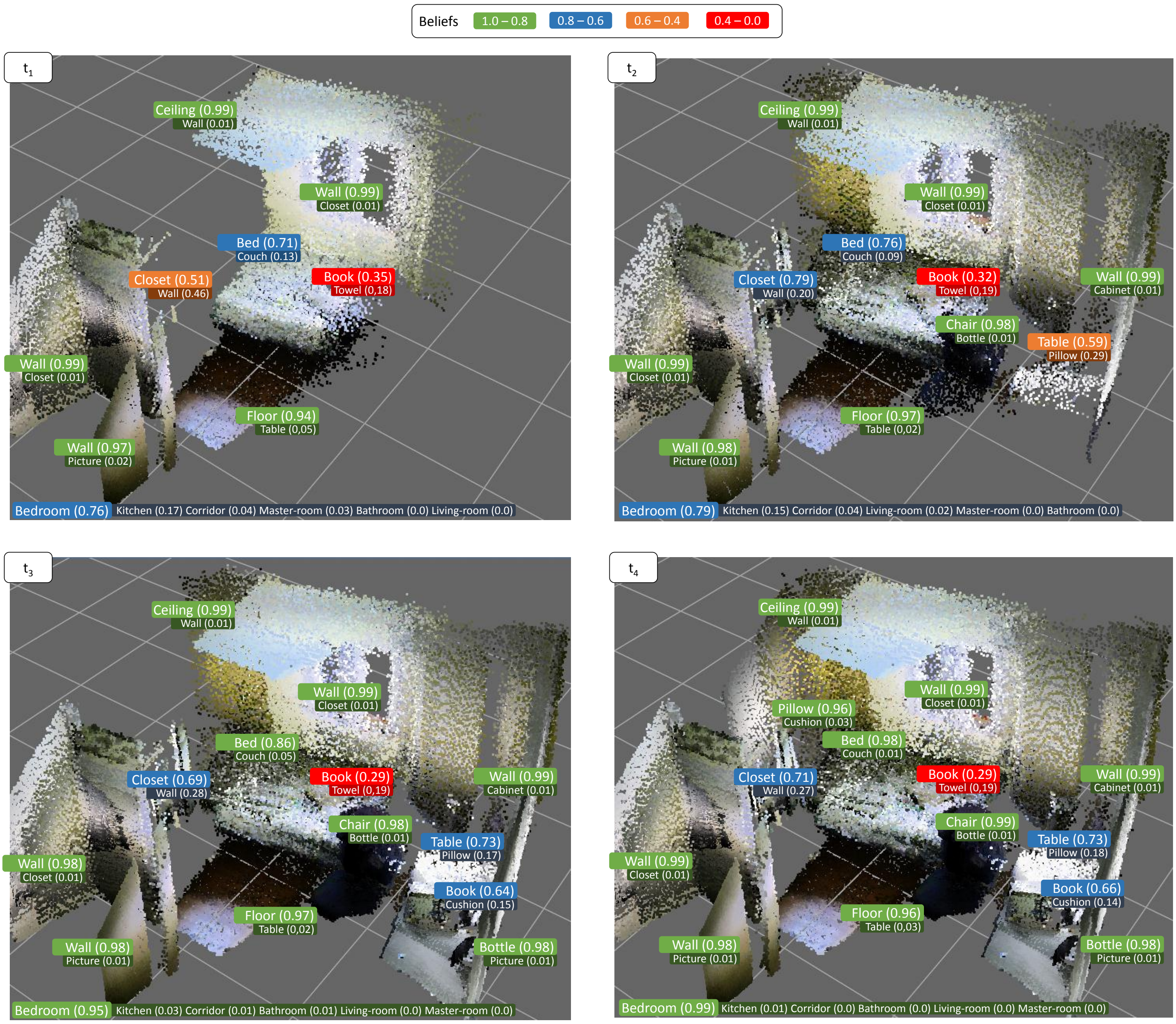

Figure 9: Grounding results and their belief values for the symbols of spatial elements perceived during the robot exploration of a bedroom. The registered point clouds in each image are shown for illustrative purposes.

of high-level tasks. For that, the *MvSmap* accommodates the uncertainty about the grounded concepts as universes, which can be seen as different interpretations of the workspace. Notice that *MvSmaps* can be exploited for traditional semantic map applications (*e.g.* task planning, planning with incomplete information, navigation, human-robot interaction, localization, etc.) by considering only a universe, albeit its potential to measure the (un)certainty of the robot's understanding can be exploited for an intelligent, more efficient robotic operation.

A clear example of this can be envisioned while planning an object search task. Let's suppose an scenario where the robot is commanded to bring the slippers to the user. If the slippers have not been detected before, the robot could infer (according to its semantic knowledge) that their most probable location is a bedroom. Fortunately, a room, corresponding to the farthest one from the robot location, has been already grounded as being a bedroom with a belief of 0.42, and 0.41 of being a kitchen. Another room, close to the robot location, has been grounded to the `Kitchen` concept with a belief of 0.47, and to the `Bedroom` one with 0.45. The utilization of only the most probable universe would lead to the exploration of the farthest room, with a 42% of being the correct place, while the consideration of both interpretations would produce the more logical plan of taking a look at the closer one first. Moreover, the Conditional Random Field employed in this work is able to provide a more fine-grained and coherent prediction than just employing semantic knowledge: it permits to hypothesize about the exact location of an object or a room, and to retrieve the likelihood of such location through an inference method [48, 16]. By repeating this process in different locations, the robot can operate according to a list of possible object locations ordered by their likelihood.

Another typical application of semantic maps resorting to logical reasoning engines is the classification of rooms according to the objects therein [25]. For example, if an object is grounded as a refrigerator, and kitchens are defined in the Knowledge Base as rooms containing a refrigerator, a logical reasoner can infer that the room is a kitchen. Again, this reasoning relying on crispy information can provoke undesirable results if the symbol grounding process fails at categorizing the object, which can be avoided employing *MvSmaps*.

Galindo and Saffiotti [18], envisages an application of semantic maps where they encode information about how things should be, also called norms, allowing the robot to infer deviations from these norms and act accordingly. The typical norm example is that ”towels must be in bathrooms”, so if a towel is detected, for example, on the floor of the living room, a plan is generated to bring it to the bathroom. This approach works with crispy information, *e.g.* an object is a towel or not. Instead, the consideration of a *MvSmap* would permit the robot to behave more coherently, for example gathering additional information if the belief of an object symbol being grounded to `Towel` is 0.55 while to `Carpet` is 0.45. In this example, a crispy approach could end up with a carpet in our bathroom, or a towel in our living room. The scenarios illustrated in this section compound a – non exhaustive – set of applications where *MvSmaps* clearly enhance the performance of traditional semantic maps.

7. Conclusions and Future Work

In this work we have presented a solution for tackling the symbol grounding problem in semantic maps from a probabilistic stance, which has been integrated into a novel environment representation coined Multiversal Semantic Map (*MvSmap*). Our approach employs Conditional Random Fields (CRFs) for performing symbol grounding, which permits the exploitation of contextual relations among object and room symbols, also dealing with the uncertainty inherent to the grounding process. The uncertainties concerning the grounded symbols, yielded by probabilistic inference methods over those CRFs, allow the robot to consider diverse interpretations of the spatial elements in the workspace. These interpretations are called universes, which are encoded as instances of the codified ontology with symbols grounded to different concepts, and annotated with their probability of being the right one. Thereby, the proposed *MvSmap* represents the robot environment through a hierarchy of spatial elements, as well as a hierarchy of concepts, in the form of an ontology, which is instantiated according to the considered universes. This paper also describes the processes involved in the building of *MvSmaps* for a given workspace. We have also proposed an strategy for tackling the exponential growing of the multiverse size in complex environments, and analyzed some of the applications where *MvSmaps* can be used to enhance the performance of traditional semantic maps.

The suitability of the proposed probabilistic symbol grounding has been assessed with the challenging Robot@Home dataset. The reported success without considering contextual relations were of $\sim 73.5\%$ and $\sim 57.5\%$ while grounding object and room symbols respectively, while including them these figures increased up to $\sim 81.5\%$ and 91.5%. It has been also shown the building of *MvSmaps* according to the information gathered by a mobile robot in two scenarios with different complexity.

Typically, the semantic knowledge encoded in a semantic map is considered as written in stone, *i.e.* it is defined at the laboratory and does not change during the robot operation. We are studying how to modify this knowledge according to the peculiarities of a given domain, also in combination with a CRF [24]. We think that this line of research is interesting since it would permit the robot, for example, to consider new object or room types not previously introduced, or to modify the properties and relations of those already defined. Additionally, we plan to progressively exploit the presented *MvSmaps* for the applications analyzed in this paper and/or other of interest.

Acknowledgements

This work is supported by the research projects TEP2012-530 and DPI2014-55826-R, funded by the Andalusia Regional Government and the Spanish Government, respectively, both financed by European Regional Development's funds (FEDER).

Appendix A. Performance metrics

The *precision* metric for a given type of object/room l_i reports the percentage of elements recognized as belonging to l_i that really belong to that type. Let $recognized(l_i)$ be the set of objects/rooms recognized as belonging to the type l_i, $gt(l_i)$ the set of elements of that type in the ground-truth, and $|\cdot|$ the cardinality of a set, then the *precision* of the classifier for the type l_i is defined as:

$$precision(l_i) = \frac{|recognized(l_i) \bigcap gt(l_i)|}{|recognized(l_i)|} \quad \text{(A.1)}$$

In its turn, the *recall* for a class l_i expresses the percentage of the spatial elements that belonging to l_i in the ground-truth are recognized as members of that type:

$$recall(l_i) = \frac{|recognized(l_i) \bigcap gt(l_i)|}{|gt(l_i)|}. \quad \text{(A.2)}$$

Precision and *recall* are metrics associated to a single type. To report more general results, we are interested in the performance of the proposed methods for all the considered types. This can be measured by adding the so-called macro/micro concepts. *Macro precision/recall* represents the average value of the precision/recall for a number of types, defined in the following way:

$$macro\_precision = \frac{\sum_{i \in L} precision(l_i)}{|L|} \quad \text{(A.3)}$$

$$macro\_recall = \frac{\sum_{i \in L} recall(l_i)}{|L|} \quad (A.4)$$

being L the set of considered objects/rooms. Finally, *micro precision/recall* represents the percentage of elements in the dataset that are correctly recognized with independence of their belonging type, that is:

$$micro\_precision(l_i) = \frac{\sum_{i \in L} |recognized(l_i) \bigcap gt(l_i)|}{\sum_{i \in L} |recognized(l_i)|} \quad (A.5)$$

$$micro\_recall(l_i) = \frac{\sum_{i \in L} |recognized(l_i) \bigcap gt(l_i)|}{\sum_{i \in L} |gt(l_i)|} \quad (A.6)$$

Since we assume that the spatial elements belong to a unique class, then $\sum_{i \in L} |gt(l_i)| = \sum_{i \in L} |recognized(l_i)|$, and consequently the computation of both micro precision/recall metrics gives the same value.

[1] A. Elfes, Sonar-based real-world mapping and navigation, IEEE Journal on Robotics and Automation 3 (3) (1987) 249–265.

[2] S. Thrun, Learning occupancy grid maps with forward sensor models, Autonomous Robots 15 (2) (2003) 111–127.

[3] E. Remolina, B. Kuipers, Towards a general theory of topological maps, Artificial Intelligence 152 (1) (2004) 47–104.

[4] A. Ranganathan, E. Menegatti, F. Dellaert, Bayesian inference in the space of topological maps, IEEE Transactions on Robotics 22 (1) (2006) 92–107.

[5] S. Thrun, Learning metric-topological maps for indoor mobile robot navigation, Artificial Intelligence 99 (1) (1998) 21 – 71.

[6] J. Blanco, J. Gonzlez, J.-A. Fernndez-Madrigal, Subjective local maps for hybrid metric-topological {SLAM}, Robotics and Autonomous Systems 57 (1) (2009) 64 – 74.

[7] S. Thrun, W. Burgard, D. Fox, Probabilistic Robotics, Intelligent robotics and autonomous agents, MIT Press, 2005.

[8] A. Ranganathan, F. Dellaert, Semantic modeling of places using objects, in: Robotics: Science and Systems Conference III (RSS), MIT Press, 2007.

[9] S. Ekvall, D. Kragic, P. Jensfelt, Object detection and mapping for service robot tasks, Robotica 25 (2) (2007) 175–187.

[10] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, D. G. Lowe, Curious george: An attentive semantic robot, Robots and Autonomous Systems 56 (6) (2008) 503–511.

[11] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, M. Beetz, Towards 3d point cloud based object maps for household environments, Robotics and Autonomous Systems 56 (11) (2008) 927 – 941, semantic Knowledge in Robotics.

[12] J. McCormac, A. Handa, A. Davison, S. Leutenegger, Semanticfusion: Dense 3d semantic mapping with convolutional neural networks, arXiv preprint arXiv:1609.05130.

[13] S. Harnad, The symbol grounding problem, Phys. D 42 (1-3) (1990) 335–346.

[14] S. Coradeschi, A. Saffiotti, An introduction to the anchoring problem, Robotics and Autonomous Systems 43 (2-3) (2003) 85–96.

[15] S. Coradeschi, A. Loutfi, B. Wrede, A short review of symbol grounding in robotic and intelligent systems, KI - Künstliche Intelligenz 27 (2) (2013) 129–136.

[16] A. Pronobis, P. Jensfelt, Large-scale semantic mapping and reasoning with heterogeneous modalities, in: Robotics and Automation (ICRA), 2012 IEEE International Conference on, 2012, pp. 3515–3522.

[17] B. Mutlu, N. Roy, S. Šabanović, Cognitive Human–Robot Interaction, Springer International Publishing, Cham, 2016, pp. 1907–1934.

[18] C. Galindo, A. Saffiotti, Inferring robot goals from violations of semantic knowledge, Robotics and Autonomous Systems 61 (10) (2013) 1131–1143.

[19] C. Galindo, J. Fernandez-Madrigal, J. Gonzalez, A. Saffiotti, Robot task planning using semantic maps, Robotics and Autonomous Systems 56 (11) (2008) 955–966.

[20] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. Kruijff, W. Burgard, Conceptual spatial representations for indoor mobile robots, Robotics and Autonomous Systems 56 (6) (2008) 493 – 502, from Sensors to Human Spatial Concepts.

[21] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[22] J. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Mobile robot object recognition through the synergy of probabilistic graphical models and semantic knowledge, in: European Conf. on Artificial Intelligence. Workshop on Cognitive Robotics, 2014.

[23] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Learning approaches for probabilistic graphical models. application to scene object recognition., Submitted.

[24] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Probability and common-sense: Tandem towards robust robotic object recognition in ambient assisted living, 10th International Conference on Ubiquitous Computing and Ambient łIntelligence (UCAmI 2016).

[25] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. A. Fernandez-Madrigal, J. Gonzalez, Multi-hierarchical semantic maps for mobile robotics, in: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, pp. 2278–2283.

[26] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, The Knowledge Engineering Review 11 (1996) 93–136.

[27] L. Karlsson, Conditional progressive planning under uncertainty, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 431–436.

[28] M. P. R. S. Al-Moadhen, Ahmed Abdulhadi, R. Qiu, Robot task planning in deterministic and probabilistic conditions using semantic knowledge base, International Journal of Knowledge and Systems Science (IJKSS) 7 (1) (2016) 56–77.

[29] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Robot@home, a robotic dataset for semantic mapping of home environments, Submitted.

[30] R. Marfil, L. J. Manso, J. P. Bandera, A. Romero-Garcés, A. Bandera, P. Bustos, L. V. Calderita, J. C. González, Á. García-Olaya, R. Fuentetaja, et al., Percepts symbols or action symbols? generalizing how all modules interact within a software architecture for cognitive robotics, in: 17th Workshop of Physical Agents (WAF), 2016, pp. 9–16.

[31] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Vol. 1, 2001, pp. 511–518.

[32] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[33] J. Knopp, M. Prasad, G. Willems, R. Timofte, L. Van Gool, Hough transform and 3d surf for robust three dimensional classification, in: Proceedings of the 11th European Conference on Computer Vision: Part VI, ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 589–602.

[34] M. Pontil, A. Verri, Support vector machines for 3d object recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (6) (1998) 637–646.

[35] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2161–2168.

[36] W. L. Hoo, C. H. Lim, C. S. Chan, Keybook: Unbias object recognition using keywords, Expert Systems with Applications 42 (8) (2015) 3991 – 3999.

[37] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, in: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), 2006, pp. 13–13.

[38] O. Mozos, C. Stachniss, W. Burgard, Supervised learning of places from range data using adaboost, in: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on, 2005, pp. 1730–1735.

[39] A. Oliva, A. Torralba, Building the gist of a scene: The role of global image features in recognition, Progress in brain research 155 (2006) 23–36.

[40] H. Andreasson, A. Treptow, T. Duckett, Localization for mobile robots using panoramic vision, local features and particle filter, in: Proceedings of the 2005 IEEE International Conference on Robotics and Automation,

2005, pp. 3348–3353.

[41] A. C. Murillo, J. J. Guerrero, C. Sagues, Surf features for efficient robot localization with omnidirectional images, in: Proceedings 2007 IEEE International Conference on Robotics and Automation, 2007, pp. 3901–3907.

[42] C. Weiss, H. Tamimi, A. Masselli, A. Zell, A hybrid approach for vision-based outdoor robot localization using global and local image features, in: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007, pp. 1047–1052.

[43] A. Pronobis, B. Caputo, Confidence-based cue integration for visual place recognition, in: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007, pp. 2394–2401.

[44] A. Oliva, A. Torralba, The role of context in object recognition, Trends in Cognitive Sciences 11 (12) (2007) 520–527.

[45] S. Divvala, D. Hoiem, J. Hays, A. Efros, M. Hebert, An empirical study of context in object detection, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1271–1278.

[46] C. Galleguillos, S. Belongie, Context based object categorization: A critical survey, Computer Vision and Image Understanding 114 (6) (2010) 712–722.

[47] J. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, UPGMpp: a Software Library for Contextual Object Recognition, in: 3rd. Workshop on Recognition and Action for Scene Understanding, 2015.

[48] A. Anand, H. S. Koppula, T. Joachims, A. Saxena, Contextually guided semantic labeling and search for three-dimensional point clouds, In the International Journal of Robotics Research 32 (1) (2013) 19–34.

[49] J. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, P. Torr, Mesh based semantic modelling for indoor and outdoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), 2013, pp. 2067–2074.

[50] X. Xiong, D. Huber, Using context to create semantic 3d models of indoor environments, in: In Proceedings of the British Machine Vision Conference (BMVC 2010), 2010, pp. 45.1–11.

[51] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Exploiting semantic knowledge for robot object recognition, Knowledge-Based Systems 86 (2015) 131–142.

[52] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Scene object recognition for mobile robots through semantic knowledge and probabilistic graphical models, Expert Systems with Applications 42 (22) (2015) 8805–8816.

[53] J. G. Rogers, H. I. Christensen, A conditional random field model for place and object classification, in: Robotics and Automation (ICRA), 2012 IEEE International Conference on, 2012, pp. 1766–1772.

[54] D. Lin, S. Fidler, R. Urtasun, Holistic scene understanding for 3d object detection with rgbd cameras, IEEE International Conference on Computer Vision 0 (2013) 1417–1424.

[55] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Joint categorization of objects and rooms for mobile robots, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[56] K. P. Murphy, Y. Weiss, M. I. Jordan, Loopy belief propagation for approximate inference: An empirical study, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, 1999, pp. 467–475.

[57] J. S. Yedidia, W. T. Freeman, Y. Weiss, Generalized Belief Propagation, in: Advances Neural Information Processing Systems, Vol. 13, 2001, pp. 689–695.

[58] A. Nüchter, J. Hertzberg, Towards semantic maps for mobile robots, Robots and Autonomous Systems 56 (11) (2008) 915–926.

[59] E. Prestes, J. L. Carbonera, S. R. Fiorini, V. A. M. Jorge, M. Abel, R. Madhavan, A. Locoro, P. Goncalves, M. E. Barreto, M. Habib, A. Chibani, S. Grard, Y. Amirat, C. Schlenoff, Towards a core ontology for robotics and automation, Robotics and Autonomous Systems 61 (11) (2013) 1193 – 1204, ubiquitous Robotics.

[60] M. Tenorth, L. Kunze, D. Jain, M. Beetz, Knowrob-map - knowledge-linked semantic object maps, in: 2010 10th IEEE-RAS International Conference on Humanoid Robots, 2010, pp. 430–435.

[61] D. Pangercic, B. Pitzer, M. Tenorth, M. Beetz, Semantic object maps for robotic housework - representation, acquisition and use, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 4644–4651.

[62] L. Riazuelo, M. Tenorth, D. D. Marco, M. Salas, D. Glvez-Lpez, L. Msenlechner, L. Kunze, M. Beetz, J. D. Tards, L. Montano, J. M. M. Montiel, Roboearth semantic mapping: A cloud enabled knowledge-based approach, IEEE Transactions on Automation Science and Engineering 12 (2) (2015) 432–443.

[63] J. O. Reinaldo, R. S. Maia, A. A. Souza, Adaptive navigation for mobile robots with object recognition and ontologies, in: 2015 Brazilian Conference on Intelligent Systems (BRACIS), 2015, pp. 210–215.

[64] M. Günther, T. Wiemann, S. Albrecht, J. Hertzberg, Building semantic object maps from sparse and noisy 3d data, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013), 2013, pp. 2228–2233.

[65] E. Bastianelli, D. D. Bloisi, R. Capobianco, F. Cossu, G. Gemignani, L. Iocchi, D. Nardi, On-line semantic mapping, in: Advanced Robotics (ICAR), 2013 16th International Conference on, 2013, pp. 1–6.

[66] G. Gemignani, D. Nardi, D. D. Bloisi, R. Capobianco, L. Iocchi, Interactive semantic mapping: Experimental evaluation, in: A. M. Hsieh, O. Khatib, V. Kumar (Eds.), Experimental Robotics: The 14th International Symposium on Experimental Robotics, Vol. 109 of Springer Tracts in Advanced Robotics, Springer International Publishing, 2016, pp. 339–355.

[67] I. Kostavelis, A. Gasteratos, Semantic mapping for mobile robotics tasks: A survey, Robotics and Autonomous Systems 66 (2015) 86–103.

[68] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, New York, NY, USA, 2007.

[69] K. Zhou, M. Zillich, H. Zender, M. Vincze, Web mining driven object locality knowledge acquisition for efficient robot behavior, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), 2012, pp. 3962–3969.

[70] R. Speer, C. Havasi, Conceptnet 5: a large semantic network for relational knowledge, in: The Peoples Web Meets NLP. Theory and Applications of Natural Language, Springer, 2013, pp. 161–176.

[71] R. Gupta, M. J. Kochenderfer, Common sense data acquisition for indoor mobile robots, in: Proceedings of the 19th National Conference on Artifical Intelligence, AAAI'04, AAAI Press, 2004, pp. 605–610.

[72] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: IEEE International Conference on Computer Vision (ICCV 2005), Vol. 2, 2005, pp. 1816–1823 Vol. 2.

[73] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[74] S. Thrun, Robotic mapping: A survey, in: Exploring Artificial Intelligence in the New Millennium, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003, pp. 1–35.

[75] F. Lu, E. Milios, Globally consistent range scan alignment for environment mapping, Autonomous Robots 4 (4) (1997) 333–349.

[76] J. J. Leonard, H. F. Durrant-Whyte, Mobile robot localization by tracking geometric beacons, IEEE Transactions on Robotics and Automation 7 (3) (1991) 376–382.

[77] C. Galindo, J. Fernandez-Madrigal, J. Gonzalez, A. Saffiotti, Using semantic information for improving efficiency of robot task planning, in: IEEE International Conference on Robotics and Automation (ICRA), Workshop on Semantic Information in Robotics, Rome, Italy, 2007.

[78] A. Sloman, J. Chappell, The altricial-precocial spectrum for robots, in: International Joint Conference on Artificial Intelligence, Vol. 19, Lawrence Erlbaum Associates LTD, 2005, p. 1187.

[79] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, Pellet: A practical owl-dl reasoner, Web Semantics: Science, Services and Agents on the World Wide Web 5 (2) (2007) 51–53.

[80] D. Tsarkov, I. Horrocks, FaCT++ Description Logic Reasoner: System Description, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 292–297.

[81] V. Haarslev, K. Hidde, R. Möller, M. Wessel, The racerpro knowledge representation and reasoning system, Semantic Web Journal 3 (3) (2012) 267–277.

[82] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, J. J. Leonard, Simultaneous localization and mapping: Present, future, and the robust-perception age, arXiv preprint arXiv:1606.05830.

[83] P. J. Besl, N. D. McKay, A method for registration of 3-d shapes, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (1992) 239–256.

[84] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289.

[85] J. Hammersley, P. Clifford, Markov fields on finite graphs and lattices, unpublished manuscript (1971).

[86] M. J. Wainwright, M. I. Jordan, Graphical models, exponential families, and variational inference, Found. Trends Mach. Learn. 1 (1-2) (2008) 1–305.

[87] Y. Weiss, W. T. Freeman, On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs, IEEE Trans. Inf. Theor. 47 (2) (2006) 736–744.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.