Accepted Authors' version. Final version published in Knowledge-Based Systems, DOI: https://doi.org/10.1016/j.knosys.2019.01.005

Ontology-Based Conditional Random Fields for Object Recognition

Jose-Raul Ruiz-Sarmiento[a,*], Cipriano Galindo[a], Javier Monroy[a], Francisco-Angel Moreno[a], Javier Gonzalez-Jimenez[a]

[a]*Machine Perception and Intelligent Robotics Group, System Engineering and Automation Department, and Biomedical Research Institute of Málaga (IBIMA), University of Málaga, Campus de Teatinos, 29071, Málaga, Spain.*

Abstract

Object recognition is a cornerstone task in autonomous and/or assistance systems like robots, autonomous vehicles, or those assisting to visually impaired, aiming to achieve a certain level of understanding of their surroundings. Probabilistic models, such as *Conditional Random Fields* (CRFs), have been successfully applied to this end given their ability to exploit contextual and situation information, *e.g.* a bowl is typically found in a cabinet and not in a night-stand. In this work we propose to evolve CRFs into *Ontology-based Conditional Random Fields* (*ob*CRFs), which define a multi-level structure where each level assigns a category with different granularity to the same set of objects. For example, a level could assign to an object the category *appliance* or *furniture*, while the next one could categorize it into the *tv*, *microwave*, *cabinet*, or *table* types. In this way, *general* categorizations can guide the classification into more *specialized* ones (and vice versa), improving recognition success, and mitigating the CRFs limitations when modeling a high number of object categories (shared, in general, by *Machine Learning* techniques). To set the categories in each level we propose to mimic the hierarchical structure of ontologies, where categories are naturally codified following a subsumption ordering. This leads us to the second advantage of *ob*CRFs : the multi-labeling of objects provides a richer understanding of the scene, which can be leveraged for accomplishing high-level tasks (*e.g.* object search or scheduling). Our approach has been tested with scenes from two state-of-the-art datasets: *Robot@Home* and *Cornell-RGBD*, outperforming the results provided by standard CRFs.

Keywords: object recognition, conditional random fields, ontologies, probabilistic graphical models

1. INTRODUCTION

For a proper operation, autonomous and/or assistance systems providing services in human environments like houses, offices, stores, etc., need to recognize the objects in their surroundings. Examples of these systems could be mobile robots, autonomous vehicles, or vision-based wearables for impaired people, to name a few. Recent general purpose object recognition systems based on intensity (RGB) images rely on *Convolutional Neuronal Networks* (CNNs) [1, 2], like Faster R-CNN [3], SSD [4], or R-FCN [5], while *Probabilistic Graphical Models* [6, 7], as Markov or Conditional Random Fields [8, 9], are commonly used to process RGB-D (RGB plus depth) or point cloud representations of the environment. There are also works that explore the combination of both techniques [2, 10, 11], like the DeepLab v2 [12] model. These approaches achieve a high recognition success in datasets with a moderated number of objects' categories, like the PASCAL VOC 2012 Classification dataset [13] with 20 categories from both indoors and outdoors (*e.g.* person, bird, airplane, bottle, etc.).

Nevertheless, the recognition of objects by an autonomous system in human-like environments has some peculiarities to take into account, including i) a typically large number of object categories relevant to the system operation, or ii) the availability of additional information like the physical system location or the relations among perceived objects [14]. To illustrate the first point, we can highlight the Robot@Home dataset [15], a collection of data from homes containing objects from 157 different categories, the Cornell-RGBD repository [16], with information from offices and homes containing 129 labelled types, or the NYUv2 dataset [17], which spans over a variety of facilities and includes 894 object categories, figures that are considerably higher than those employed by general purpose systems in the literature [2]. Although there are CNNs that have been designed to work with large datasets and a high number of categories, their recognition successes notably decrease (*e.g.* the CNN in [3] achieves a mean averaged precision of 21.9% with the +80 categories of the COCO dataset [18] for a 75.9% with PASCAL) hence affecting the proper system operation. However, these cutting-edge models are not mutually exclusive with the ideas proposed here since, as we will discuss, they could be incorporated into modern CNNs.

Regarding the additional information available for recognition, like the autonomous system location in the environment or the spatial relations among the perceived objects, it can be of great utility for disambiguation to incorporate contextual awareness in the recognition process: *e.g.* a toothbrush inside a white cylindrical object helps to categorize it as a toothbrush holder and not as a mug. In this way, *Conditional Random*

*Corresponding author

Email addresses: jotaraul@uma.es (Jose-Raul Ruiz-Sarmiento), cipriano@ctima.uma.es (Cipriano Galindo), jgmonroy@uma.es (Javier Monroy), famoreno@uma.es (Francisco-Angel Moreno), javiergonzalez@uma.es (Javier Gonzalez-Jimenez)

Accepted Authors' version. Final version published in Knowledge-Based Systems, DOI: https://doi.org/10.1016/j.knosys.2019.01.005

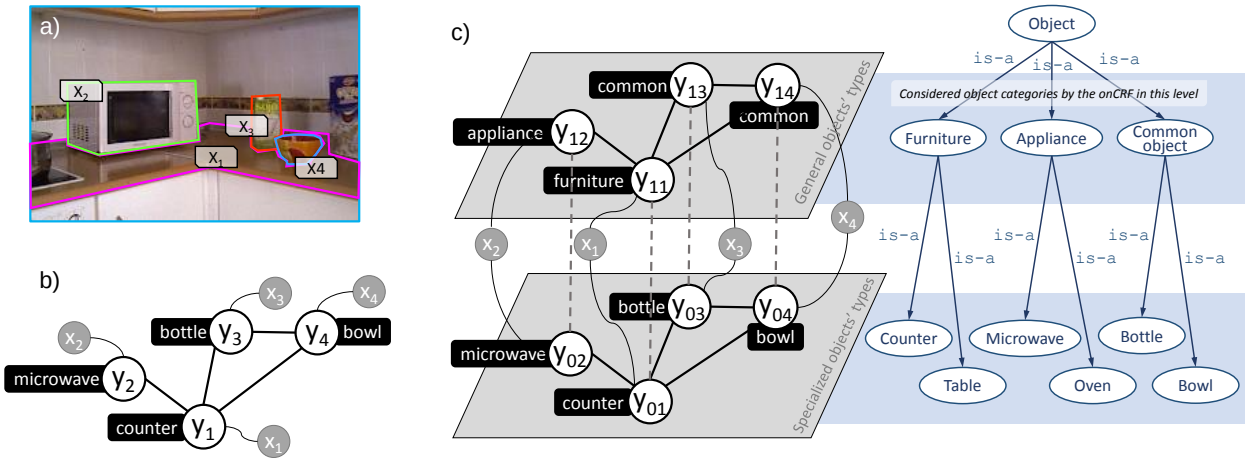

Figure 1: Example illustrating the ideas behind an *ob*CRF: a) scene capturing part of a kitchen with the observed objects labeled $(x_1, \ldots, x_4)$. b) Standard CRF graph built from that scene, including as many nodes $(y_1, \ldots, y_4)$ as objects. c) *ob*CRF graph of the scene in a) with two levels, indicating on the right the object categories considered at each level from an ontology.

Fields (CRFs) [7], a particular type of Probabilistic Graphical model, have stood out as a powerful tool for context-based object recognition [19, 10, 11, 8]. CRFs codify each object category through a number of parameters associated to the features used to describe them (size, shape, color, etc.), as well as to the features characterizing their spatial contextual relations (proximity, difference in height, perpendicularity, etc.) [7]. The higher the number of object categories, the higher the number of needed parameters, resulting in more complex models that require a more challenging design and training phases. Furthermore, a higher number of categories usually entails a harder discrimination among them, *i.e.* objects belonging to different types can show similar features. These factors, likewise working with general purpose *Machine Learning* models, directly affect the CRFs recognition success [20, 8].

To deal with these peculiarities, in this work we propose a new approach called *Ontology-based Conditional Random Fields* (*ob*CRFs), whose modeling is inspired by the subsumption ordering of *ontologies* [21]. Ontologies are by nature hierarchical representations of the elements of a domain, codified by a set of concepts structured according to a subsumption ordering. Such ordering ranges from general to specialized concepts (see Fig. 1), *e.g.* the concept `Thing` could subsume the `Object` one, `Object` subsumes `Appliance`, and `Appliance` subsumes the `Microwave` concept. Intuitively, it is easier to categorize objects in general types sharing some properties (*e.g.* structural surface, furniture, appliance, etc.) than in more specialized ones (*e.g.* wall, window, door, chair, couch, etc.). *ob*CRFs leverage this for building a graph representation with multiple levels, where each level assigns a category with different granularity to each object in the scene. For example, the bottom level could assign the category `Microwave` to an object, the next one `Appliance`, and so on.

One of the advantages of this model is that the categorization into more general levels guides the recognition of specialized ones –and vice versa– in a principled way. This provides a high system performance, also mitigating the decrease in the recognition success when recognizing objects from a growing number of categories. To set the categories considered in each level we employ the hierarchy of concepts defined in an ontology. It is important to remark that, usually, general concepts within ontologies group together specialized ones that share some property, like their functionality, despite their physical attributes. However, in the context of human-like settings like a home or an office, general concepts referring to object categories can be defined in such a way that they also share physical properties that are useful for recognition: *e.g.* structural surfaces are typically large, flat planes. Another advantage of this approach is that the multiple categorizations of each object provide rich information exploitable for high-level tasks, *e.g.* object search [22, 23], scheduling [24, 25], etc.

To evaluate *ob*CRFs we have employed the aforementioned Robot@Home and Cornell-RGBD datasets. Robot@Home is a challenging repository, collected by a mobile robot, which contains data from 83 sequences of RGB-D images surveying home environments, including objects from 157 different categories. Cornell-RGBD contains scenes from office and home environments, reconstructed from 550 RGB-D observations, with 2,5k objects therein belonging to 129 categories. A battery of tests has been carried out considering sets of object types with different size and measuring the performance of standard CRFs and *ob*CRFs , obtaining promising results. The efficiency of *ob*CRFs has been also assessed analyzing their training/inference execution times.

To summarize, our proposal consists of:

- A new CRF-based model with multiple levels that assigns to each scene object a number of categories with different granularity, giving consistency to them.
- A set of categories considered at each level borrowed from an ontology (or any taxonomy) ranging from specialized to general types.
- A multiple categorization for each object that can be exploited for carrying out high-level tasks.
- A detailed evaluation of the proposed method with

two state-of-the-art datasets: Robot@Home and Cornell-RGBD, reporting a high performance.

After a review of related works in Sec. 2, Sec. 3 introduces Conditional Random Fields (CRFs) and their application to scene object recognition. Then, Sec. 4 describes ontologies and their subsumption ordering, while Sec. 5 formally defines the ontology-based CRFs (*ob*CRFs), and gives some directions for their jointly utilization with Convolutional Neuronal Networks. Finally, Sec. 6 validates our proposal resorting to the Robot@Home and Cornell-RGBD datasets, and Sec. 7 concludes the paper by discussing the work done and future steps.

2. RELATED WORK

This section puts our paper in the context of other related works in the literature. For that, Sec. 2.1 reviews relevant works addressing the object recognition problem, including general purpose ones, as well as techniques based on Convolutional Neuronal Networks and Probabilistic Graphical models. Then, Sec. 2.2 shifts the review to recognition models employing *multi-level* Conditional Random Fields in different ways, clearly stating how our proposal differs from them.

2.1. Object Recognition

General purpose object recognition methods have reached a reasonable success relying on the local appearance and/or geometry of objects for their categorization. Examples of these methods are the veteran and widely-used cascade classifier by Viola and Jones [26], or those capturing the appearance of objects through features like Scale-Invariant Feature Transform (SIFT) [27] or Speeded Up Robust Features (SURF) [28], and their exploitation by means of some Machine Learning classifier like Bag-of-Words (BoWs) [29] or Support Vector Machines (SVMs) [30] based ones. A comprehensive review of methods following this pipeline can be found in the work by Zhang et al. [31]. Another well-known approach is such of Deformable Part Models (DPM) [32] and its variants, which are especially suitable for representing and detecting highly variable object classes. The recent trend, promoted by the developments in high-performance GPUs and the availability of public, large datasets (like Pascal VOC [13], COCO [18], or ImageNet [33]), is towards approaches relying on Convolutional Neuronal Networks [1], like the aforementioned Faster Region-based Convolutional Neural Network (Faster R-CNN) [3], Single Shot Detector (SSD) [4], and Region-based Fully Convolutional Network (R-FCN) [5], or the also popular models You Only Look Once v2 (YOLOv2) [34] and Neural Architecture Search Net (NASNet) [35].

Despite their virtues, these methods can provide ambiguous results while recognizing objects which features fit well with different categories. For example, they could experience problems while recognizing a white cylindrical object as a mug or a toothbrush holder. Contextual information can help to disambiguate this: if a toothbrush is recognized inside it, the toothbrush holder option should be the right one [36, 37, 38, 39, 40]. This is why there is a tendency towards recognition methods that exploit information of such nature, being the framework of Probabilistic Graphical Models (PGMs) [7] widely used to this end.

One of the most popular works resorting to PGMs is the one by Anand *et al.* [41], where a model isomorphic to a Markov Random Field (MRF) was used to recognize objects within office and home scenes. The same tool was employed in Xiaofeng *et al.* [42] for the recognition of objects in more general indoor environments. More recently, Conditional Random Fields (CRFs), the discriminative counterpart of MRFs, are increasingly being used with this aim, as evidenced by the works by Husain *et al.* [43] where contextual relations are codified and exploited in both indoor and outdoor scenes, Wolf *et al.* [19], which employs parallelization techniques for the fast segmentation and classification of 3D point clouds using CRFs, or Xiong and Huber [44] where they classify objects into coarse categories like clutter, wall, floor or ceiling. Ruiz-Sarmiento *et al.* also employed CRFs in combination with Semantic Knowledge to improve their performance in different ways [24, 25, 45].

It is also worth mentioning the recent arrival of studies towards the combination of CRFs with different types of Neuronal Networks for semantic scene segmentation (assign to each pixel in an image a category) [2, 10, 11], given the groundbreaking performance of the latter. A known issue of these parametric approaches is the performance decrease when modeling problems in complex domains with a high number of object types [20, 8]. The proposed model palliates this problem with the utilization of *multi-level* Conditional Random Fields mimicking the subsumption ordering of ontologies.

2.2. Multi-level Conditional Random Fields

The concept of *multi-level* or *multi-layer* CRF was previously used in the literature, but with a different meaning from the one given here. For example, Kosov *et al.* [46] introduced a two-layer CRF for the classification of partially occluded objects, where one level modeled the class of the occluded object and the second one that for the occluding one. The same number of levels was used for simultaneous classification of land cover and land use in images, given their interrelation [47]. A two-layer CRF is also provided in the work by Sulimowicz *et al.* [48] to address the semantic segmentation problem.

The term *hierarchical CRF* was employed in the paper by Huang *et al.* [49], referring to a two-stage CRF model where the first stage obtains initial pixel labels for a given image, while a CRF trained with similar images refines the labeling in a second stage. The same denomination is used in the work by Reynolds and Murphy [50]. In that case, a segmentation algorithm returns a set of super pixels at different scales that form a tree-structured model, and a classifier assigns a category to them resorting to PGM inference for giving consistency. A similar idea was explored by Yang and Föstner [51], where each level operates over a different clustering of an image yielded by a multi-scale mean shift segmentation algorithm, so that the hierarchy relates coarse segments with finer ones. Wu *et al.* [23] explored the concept of *hierarchical semantic label* for the semantic segmentation problem, assigning categories from a se-

mantic hierarchy to an image with different clustering. Unlike these approaches, in this work we adopt the term multi-level referring to a structure where each level assigns a category with a different granularity to the same set of objects, ranging from general to specialized ones. This approach achieves an enhanced performance w.r.t. standard CRFs, and produces rich information towards scene understanding and a proper execution of high-level tasks.

3. BACKGROUND ON CONDITIONAL RANDOM FIELDS FOR SCENE OBJECT RECOGNITION

The problem of scene object recognition is such of assigning a type or label, *e.g.* table, chair, laptop, bottle, picture, etc., from a known set $\mathcal{L} = \{l_1, \ldots, l_k\}$ to a number of portions of a scenario. Let us introduce the following definitions in order to state the problem from a probabilistic stance:

- Let $\boldsymbol{x} = [x_1, \ldots, x_n]$ be the vector containing the observations of the n objects in the scene, where each one is characterized through a vector of m features $\mathbf{f}_{x_i\mathbf{u}} = [f_{x_i u_1}, \ldots, f_{x_i u_m}]$ (*e.g.* size, orientation, color, shape, etc.).
- Define $\boldsymbol{y} = [y_1, \cdots, y_n]$ as a vector of random variables modeling the types of the objects in $\boldsymbol{x}$ and taking values from the set $\mathcal{L}$.

CRFs model the recognition problem through the definition of the probability distribution $P(\boldsymbol{y}|\boldsymbol{x})$, which yields the probability of each possible assignment to the random variables in $\boldsymbol{y}$ conditioned on the object observations in $\boldsymbol{x}$ [7]. The goal of the recognition process is to find the assignment to $\boldsymbol{y}$ with the highest probability. Since the full definition of $P(\boldsymbol{y}|\boldsymbol{x})$ is unfeasible due to its high dimensionality, CRFs rely on independence assumptions among the random variables in order to break it down into smaller pieces. Thus, a graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is built, where $\mathcal{V}$ is a set of nodes representing the random variables in $\boldsymbol{y}$, and $\mathcal{E}$ contains the edges linking related variables/nodes[1] (see Fig. 1-a and b). This graph permits the efficient codification and exploitation of the contextual relations among the scene objects.

In this way, the probability $P(\boldsymbol{y}|\boldsymbol{x})$ is codified in the graph $\mathcal{G}$ employing the notion of factors. A factor can be interpreted as a function defining a piece of $P(\boldsymbol{y}|\boldsymbol{x})$ over a part of the graph. In this work, we have employed unary factors $\mathcal{U}(\cdot)$, defined over nodes, and pairwise factors $\mathcal{I}(\cdot)$, codified over edges. Both of them are expressed by means of log-linear models as follows:

$$\mathcal{U}(y_i, x_i, \boldsymbol{\omega}) = \sum_{l \in L} \delta(y_i = l)\boldsymbol{\omega}_l \boldsymbol{f}_{\boldsymbol{x}_i u} \quad (1)$$

$$\mathcal{I}(y_i, y_j, x_i, x_j, \boldsymbol{\theta}) = \sum_{l_1 \in L} \sum_{l_2 \in L} \delta(y_i = l_1)\delta(y_j = l_2)\boldsymbol{\theta}_{l_1,l_2} \boldsymbol{f}_{\boldsymbol{x}_i \boldsymbol{x}_j p} \quad (2)$$

where $\delta(y_i = l)$ is the Kronecker delta function, $\boldsymbol{f}_{x_i\mathbf{u}}$ is the aforementioned vector of object features, $\boldsymbol{f}_{x_i x_j \mathbf{p}}$ is the vector of pairwise features characterizing the relation between the objects x_i and x_j, and $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ are vectors of weights learned during the CRF training. According to the Hammersley-Clifford theorem [7], the probability $P(\boldsymbol{y}|\boldsymbol{x})$ can be factorized over the graph $\mathcal{G}$ by means of these factors:

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta})} e^{-\epsilon(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta})} \quad (3)$$

$$\epsilon(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{i \in V} U(y_i, x_i, \boldsymbol{\omega}) + \sum_{(i,j) \in E} I(y_i, y_j, x_i, x_j, \boldsymbol{\theta}) \quad (4)$$

where $Z(\cdot)$ is known as the partition function, so $\sum_{\xi(\boldsymbol{y})} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta}) = 1$, being $\xi(\boldsymbol{y})$ a possible assignment to the variables in $\boldsymbol{y}$, and $\epsilon(\cdot)$ is the so-called energy function.

As already mentioned, once the graph $\mathcal{G}$ for a given scene has been built, the object recognition results are provided through the finding of the most probable assignment $\hat{\boldsymbol{y}}$ to the variables in $\boldsymbol{y}$, that is:

$$\hat{\boldsymbol{y}} = arg \max_{y} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{\theta}) \quad (5)$$

To perform such inference we have applied the Loopy Belief Propagation (LBP) algorithm [52], an approximate method with a good trade-off between recognition success and computational times [8].

4. ONTOLOGIES AND THEIR STRUCTURE

An ontology $\mathcal{O}$ is a formal representation of the knowledge within a domain of discourse codified through a set of predicates $\mathcal{O} = \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ [21]. This representation usually has the form of a hierarchy of concepts arranged according to a subsumption ordering. For example, in the home domain, concepts stating objects' categories could be `Microwave`, `Vase`, or `Pillow`. Thereby, this structure results in more general concepts at the higher levels of the hierarchy, and in subsumed, more specific ones at the lower levels. Such structure forms a taxonomy ordered by `is-a` predicates, like the `is-a(Microwave,Appliance)` one which codifies that a microwave is an appliance. Another useful predicate is such of `instance-of`, which permits us to link factual data from the workspace (like a perceived object) with the defined concepts. For example, if it is perceived an unknown object identified as `obj-3`, the predicate `instance-of(obj-3,Microwave)` says that it has been classified as a microwave. Although this predicate is not necessary for the current work, it could be useful for formally codifying the results of the proposed object recognition method and their posterior exploitation in high-level tasks [24, 22, 23].

The `is-a` and `instance-of` predicates are common to most ontologies, but additional ones could be needed to fully define the concepts of a domain. The reader is referred to [53] to see examples of predicates used to define the typical visual and geometric features of objects commonly found at homes (*e.g.* `has-size` or `has-height`), with the goal of carrying out object recognition, or to [54] to identify the predicates explored for having a mobile robot operating and interacting with people in a mall.

[1] In the case of the object recognition problem, two nodes are related if their associated observations are close to each other in the scene.

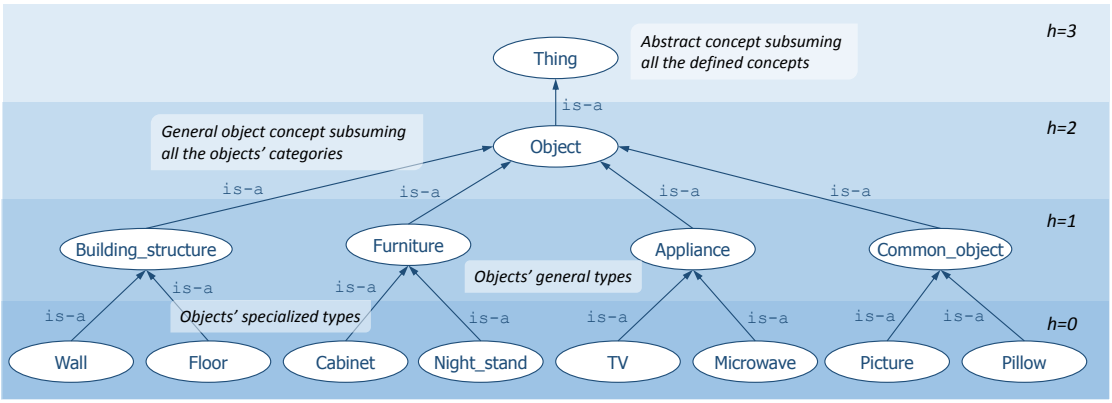

Figure 2: Simplified version of the ontology used in this work, which has two levels describing objects with a different granularity.

As firstly described in [21], the formal definition of an ontology of a certain domain consists of two steps. On the one hand, the domain of interest has to be analyzed to *capture* the relevant concepts, predicates and relations. For this step, different techniques can be used like web-mining, knowledge acquisition systems, or the one followed in this paper: expert elicitation [24]. On the other hand, the resultant ontology is *coded* to explicitly represent the conceptualization of the domain by means of a formal language. One of the most extended languages to accomplish this is the Web Ontology Language (OWL [55]), which is resorted in this work through the Protégè tool [56]. Once defined, the ontology can be employed for different purposes. For example, a query language like SPARQL [57] can be used to retrieve codified knowledge, or a logical reasoner as Pellet [58] or FaCT++ [59] to infer new information [60] or classify instances into concepts [61].

To illustrate an already-built ontology, Fig. 2 shows a simplified version of the one used in this work, which defines a set of concepts typically found in the home domain. The root concept is `Thing`, an abstract concept that subsumes all the concepts within the ontology. It has a child, `Object`[2], which in its turn subsumes the `Building_structure`, `Furniture`, `Appliance` and `Common_object` concepts that set general objects' types. In turn, these concepts subsume more specific types, hence leading to a finer classification of objects. Notice that, according to the subsumption ordering stated by the `is-a` predicate, a TV, for example, `is-a` appliance, an appliance `is-an` object, and an object `is-a` thing. The next section explains how this structure is exploited for the *ob*CRF modeling.

5. ONTOLOGY-BASED CONDITIONAL RANDOM FIELDS

An Ontology-based Conditional Random Field (*ob*CRF) enhances standard CRFs (recall Sec. 3) with additional nodes and relations according to a multi-level structure, where the categories considered in each level mimic the subsumption ordering of ontologies. To give an intuition about how this has been done, let's consider the notion of *height* h of the ontology in Fig. 2, so the root node has a height $h = 3$, while the leaf concepts defining specialized types have $h = 0$ (see right part of Fig. 2). Thus, the first (bottom) level in the *ob*CRF structure classifies the objects in the scene into the types defined by concepts with height 0, while the second level considers those with height 1 to classify the same objects. Notice that the number of layers of an *ob*CRF is not limited to 2, so more levels could be added if needed.

5.1. Model definition

In this way, a general *ob*CRF model is formally defined by the following items at each level s:

- The same vector of objects' observations $\boldsymbol{x}_s = [x_{s1}, \ldots, x_{sn}]$ used in standard CRFs.
- A set $\mathcal{L}_s = \{l_{s1}, \ldots, l_{sk}\}$ stating the objects' types considered in such level, which corresponds to those with $h = s$ in the ontology (*e.g.* if $s = 0$, and according to Fig. 2, then these types are cabinet, night-stand, tv, etc.).
- A vector of random variables $\boldsymbol{y}_s = [y_{s1}, \ldots, y_{sn}]$ that assigns a value from $\mathcal{L}_s$ to the objects' observations in $\boldsymbol{x}_s$.

Thus, in the graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of an *ob*CRF with n_l levels, the set of nodes is composed of the objects' nodes modeling the variables at each level, that is $\mathcal{V} = \{\mathcal{V}_s = \boldsymbol{y}_s \mid s \in [0, \ldots, n_l - 1]\}$, while the set of edges contains both the edges between the nodes at each level, and the edges between those at each level and the level above, *i.e.* $\mathcal{E} = \{\mathcal{E}_{ss} \mid s \in [0, \ldots, n_l - 1]\} \cup \{\mathcal{E}_{sr} \mid s \in [0, \ldots, n_l - 2], r = s + 1\}$. As an example, Figure 1-c shows the *ob*CRF built for the scene in Fig. 1-a, including these nodes and edges. Notice that, since two levels are considered in that example, each object in such scene is now represented by two nodes, categorizing it as belonging to two related types with different granularity (*e.g.* appliance and microwave). The idea behind this is that it is easier to classify objects into more general types sharing some features, and that this can help when classifying them into more specific ones. As these classifications are jointly performed within the same *ob*CRF, the result is a principled model

[2] In a more complete ontology, siblings' concepts of `Object` could have been defined, for example the concept `Room`.

Accepted Authors' version. Final version published in Knowledge-Based Systems, DOI: https://doi.org/10.1016/j.knosys.2019.01.005

with an increase in the performance w.r.t. traditional CRFs, also becoming more robust against the negative effects behind the addition of more types (see Sec. 6).

The proposed elements of $\mathcal{G}$ in an *ob*CRF entail the utilization of n_l types of unary factors, one $\mathcal{U}_s(\cdot)$ per level, plus $n_l + (n_l - 1)$ pairwise factors: one $\mathcal{I}_{ss}(\cdot)$ per level and those $\mathcal{I}_{sr}(\cdot)$ connecting nodes in consecutive levels. Therefore, the definitions of $\mathcal{U}_s(\cdot)$ and $\mathcal{I}_{ss}(\cdot)$ $\forall s \in [0, \ldots, n_l - 1]$ are analogous to those in Eq.1 and Eq.2 respectively, while we introduce the new $\mathcal{I}_{sr}(\cdot)$ factors, that is:

$$\mathcal{U}_s(y_{si}, x_{si}, \boldsymbol{\omega}) = \sum_{l \in L_s} \delta(y_{si} = l) \boldsymbol{\omega}_l \boldsymbol{f}_{\boldsymbol{x}_{si} u} \tag{6}$$

$$\mathcal{I}_{ss}(y_{si}, y_{sj}, x_{si}, x_{sj}, \boldsymbol{\theta}) = \sum_{l_1 \in L_s} \sum_{l_2 \in L_s} \delta(y_{si} = l_1) \delta(y_{sj} = l_2) \boldsymbol{\theta}_{l_1, l_2} \boldsymbol{f}_{\boldsymbol{x}_{si} \boldsymbol{x}_{sj} p} \tag{7}$$

$$\mathcal{I}_{sr}(y_{si}, y_{rj}, x_{si}, x_{rj}, \boldsymbol{\theta}) = \sum_{l_1 \in L_s} \sum_{l_2 \in L_r} \delta(y_{si} = l_1) \delta(y_{rj} = l_2) \boldsymbol{\theta}_{l_1, l_2} \boldsymbol{f}_{\boldsymbol{x}_{si} \boldsymbol{x}_{rj} p} \tag{8}$$

In this work we have employed the same vector of features describing the objects' observations in the unary factors $\mathcal{U}_s(\cdot)$, so $\boldsymbol{f}_{x_{si} u} = \boldsymbol{f}_{x_{ri} u}$, $\forall i \in [1, \ldots, n]$ and $r = s + 1$, as well as the same set to describe the contextual relations among nodes at the different levels ($\boldsymbol{f}_{x_{si} x_{sj} p}$). In turn, the feature used to describe the relations between different types of nodes in $\mathcal{I}_{sr}$ is a *bias* one, *i.e.* a feature that always takes the same value, and its associated weights in $\boldsymbol{\theta}$ (learned during the *ob*CRF tuning) are in charge of stating the compatibility among them. These compatibilities codify the hierarchical relations encoded in the ontology, for example, that a microwave is an appliance and not a building structure.

Finally, the energy function $\epsilon(\cdot)$ (recall Eq.4) of an *ob*CRF has the form (the vectors of weights $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ have been omitted for clarity):

$$\epsilon(\boldsymbol{y}, \boldsymbol{x}) = \sum_{\mathcal{V}_s \in \mathcal{V}} \sum_{i \in \mathcal{V}_s} \mathcal{U}_s(y_{si}, x_{si}) + \sum_{\mathcal{E}_{ss} \in \mathcal{E}} \sum_{(i,j) \in \mathcal{E}_{ss}} \mathcal{I}_{ss}(y_{si}, y_{sj}, x_{si}, x_{sj}) + \sum_{\mathcal{E}_{sr} \in \mathcal{E}} \sum_{(i,j) \in \mathcal{E}_{sr}} \mathcal{I}_{sr}(y_{si}, y_{rj}, x_{si}, x_{rj}) \tag{9}$$

Like in the CRF case, we have resorted to the LBP algorithm to perform inference over the resultant *ob*CRF , achieving short processing times (see Sec. 6).

5.2. Cooperation ways for obCRFs and CNNs

The ideas behind *ob*CRFs could work in collaboration with, or even be ported to, the framework that is the winning horse in the last object recognition competitions: Convolutional Neuronal Networks (CNNs) [1]. The most intuitive way to combine them is to employ the output of CNNs, like Faster R-CNN [35], NASNet [35], or Mask R-CNN [62], directly as unary factors in the *ob*CRF (recall Eq.6), that is, $\mathcal{U}(\cdot) = [c_1, \ldots, c_k]$ where each c_i expresses the score for an object belonging to the category l_i as reported by the used CNN [63]. To mimic the *ob*CRF structure, they must be used as many CNNs as levels in the *ob*CRF (each one trained with the object types to be considered in its respective level), or a CNN able to provide multiple categories to the same object (multi-labeling).

This approach could be extended by integrating the CNN scores in the unary factors instead of replacing them, which permits the CRF to model complementary higher-level features not computed by CNNs while keeping its complexity low. This promising idea was explored in [64, 65] in the context of object recognition by a mobile robot, but using an off-the-shelf model instead of a CNN.

Another relevant problem tightly related to object recognition is such of semantic segmentation, where the goal is to assign to each pixel in an image a category from $\mathcal{L}$ [2]. CNNs and CRFs are actively collaborating for addressing this issue, as it is shown by the two-phase methods that carry out a CRF refinement of the categories over the CNN results (*e.g.* LRR [66] or DeepLab v2 [12]), or those that integrate the CRF inference directly into the CNN structure (*e.g.* Higher Order CRF [63] or Deep Gaussian CRF [67]). The *ob*CRF concept could be also ported to these networks by considering pixels with multiple categories of different granularity and using CRFs to give consistency to them. These approaches open exciting research lines for the future, although their evaluation is out of the scope of this paper, which focuses on the validation of the *ob*CRF concept in itself.

6. EVALUATION

The performance of *ob*CRFs has been tested within the Robot@Home and Cornell-RGBD datasets, which are briefly introduced in Sec. 6.1. Sec. 6.2 and Sec. 6.3 report the yielded results in both cases and compare them with the outcome of standard CRFs, while in Sec. 6.4 an analysis of the computational time required by these models is carried out.

6.1. Testbed

Robot@Home [15]. This dataset is a large repository of raw and processed information from different domestic settings. It was collected by a mobile robot endowed with a 4 RGB-D cameras-rig and a laser scanner, and processed with the Object Labeling Toolkit (OLT) [68]. The top row of Fig. 3 shows some scenes from this dataset. Among all the provided data, in this work we are interested in the RGB-D observations, so we have employed those captured by one of the equipped cameras. The dataset provides information from 47 rooms, containing more than 1,900 instances of objects belonging to 157 types, from which we have selected 36 (specialized) types with enough instances for addressing the CRFs training phase with guarantees. These types are further grouped into 5 general types, so that the built *ob*CRFs have two levels.

Cornell-RGBD [16]. This data repository contains three-dimensional reconstructions from 24 office and 28 home

Accepted Authors' version. Final version published in Knowledge-Based Systems, DOI: https://doi.org/10.1016/j.knosys.2019.01.005

Table 1: Recognition results yielded by CRF and *ob*CRF models with different configurations within the Robot@Home dataset. The symbols in these configurations mean: nodes for specialized types ($\mathcal{V}_s$), edges between the previous nodes ($\mathcal{E}_{ss}$), nodes for general types ($\mathcal{V}_{gg}$), edges between nodes of the previous type ($\mathcal{E}_{gg}$), edges between the two types of nodes ($\mathcal{E}_{sg}$).

		Specialized types			General types		
Model	Configuration	micro p./r.	macro p.	macro r.	micro p./r.	macro p.	macro r.
CRF	$\mathcal{V}_s$	62.63%	38.32%	35.71%	–	–	–
	$\mathcal{V}_s + \mathcal{E}_{ss}$	71.64%	54.72%	46.77%	–	–	–
	$\mathcal{V}_g$	–	–	–	72.79%	63.89%	62.09%
	$\mathcal{V}_g + \mathcal{E}_{gg}$	–	–	–	77.33%	68.01%	64.70%
*ob*CRF	$\mathcal{V}_s + \mathcal{V}_g + \mathcal{E}_{sg}$	67.22%	43.23%	43.78%	78.17%	70.58%	69.30%
	$\mathcal{V}_s + \mathcal{E}_{ss} + \mathcal{V}_{gg} + \mathcal{E}_{sg}$	69.31%	42.55%	44.51%	82.61%	72.28%	73.52%
	$\mathcal{V}_s+\mathcal{E}_{ss}+\mathcal{V}_{gg}+\mathcal{E}_{sg}+\mathcal{E}_{gg}$	**76.99%**	**64.37%**	**53.92%**	**85.14%**	**77.93%**	**76.26%**

scenes. The bottom row in Fig. 3 provides some examples of these reconstructions. For consistency with the previous dataset, we will work here with those from home environments, which sum, in total, 129 object categories with 1,387 instances. However, there are categories with a few (or even just one) instances that must be discarded. Indeed, in the work presenting Cornell-RGBD, where the authors proposed a MRF-based recognition model, this number was reduced to 17 types. Here we will expand them up to 21 specialized types and 5 general ones. In fact, as in the case of Robot@Home, this is the maximum number of specialized categories with enough instances in the dataset for a reliable tuning of the models.

Metrics and Software. To provide the recognition results, a leave-one-out cross-validation process has been followed. For example, in the case of the Robot@Home dataset, the data from one room were used for testing, while those from the remaining 46 fed the training process. This procedure was repeated 47 times by changing the room used for testing, and the obtained results from each one were averaged. These results are expressed by means of the micro/macro precision/recall metrics [25].

The CRFs and *ob*CRFs in this work were trained and tested using the open-source Undirected Probabilistic Graphical Models in C++ library (UPGMpp) [69].

6.2. *Results with Robot@Home*

The obtained results for the considered CRF and *ob*CRF models with Robot@Home are shown in Tab. 1. Regarding the CRFs, the first configuration in such table (first row) reports the performance achieved by a graph only with nodes, that is, without exploiting contextual relations, with a success (micro p./r.) of $\sim$ 62.5%. The inclusion of contextual information (second row) considerably increases that figure by a $\sim$ 9%, revealing the benefits of its exploitation. Similarly, the third configuration is a CRF that employs only nodes and classifies objects into general types. This model achieved a success of $\sim$ 72.5%, ten percentage points higher than working with specialized types, while the fourth row shows, again, the effect of including context. Concerning the macro p./r. metrics, the addition of context also increases them, although to different extents.

In turn, the configuration number five (fifth row) is the first one employing *ob*CRFs, and reports the results yielded by a model considering nodes of both types and edges only among nodes of different types. The reached success is of $\sim$ 67% for specialized types, substantially higher than using only that type of nodes, and of $\sim$ 78% for general types. The sixth row shows the numbers reached including more contextual relations and, finally, the last configuration incorporates all the relations proposed in this work. The results yielded, in this case, achieved

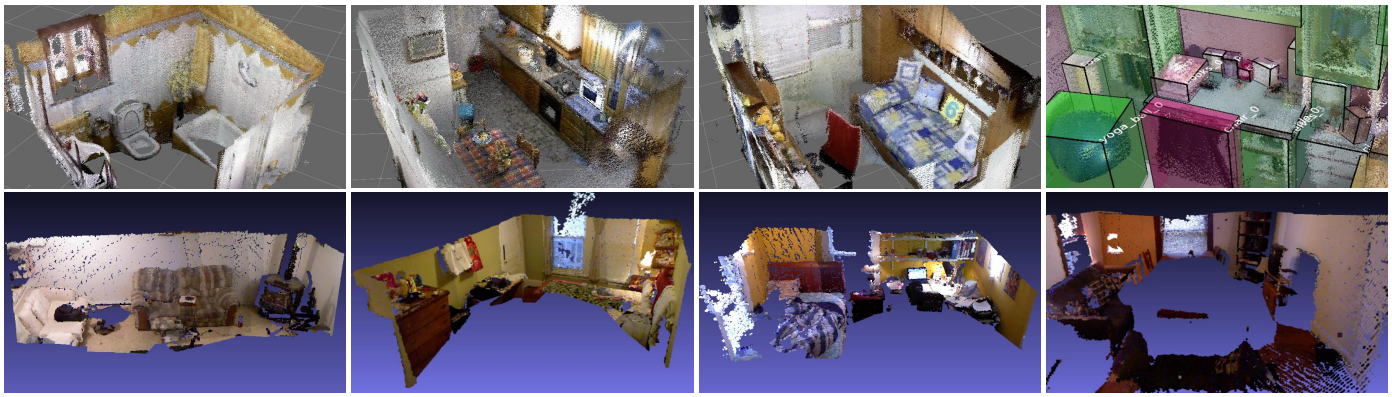

Figure 3: Example scenes from the two considered datasets: Robot@Home (top row) and Cornell-RGBD (bottom row). The last image in the first row shows an excerpt of a scene labeled with ground truth categories.

Accepted Authors' version. Final version published in Knowledge-Based Systems, DOI: https://doi.org/10.1016/j.knosys.2019.01.005

a ~ 77% of success with specialized types and a ~ 85% with general ones. Regarding micro p./r., this supposes more than a ~ 5.5% increase when recognizing specialized types w.r.t. the CRF configuration that exploits contextual relations, an increase of ~ 9.5% considering macro precision, and of ~ 7% for macro recall.

Despite the outstanding results obtained, they are a snapshot of the performance achieved for a certain number of categories. We have conducted an additional experiment in order to assess the effect of the number of object types in the CRFs performance. Its outcome is depicted in Fig. 4, reporting a clear decrease in the CRFs performance with the addition of objects' types, measured with two metrics: their micro p./r., and their F-measure (*i.e.* the harmonic mean of macro precision and recall [25]):

$$F - measure = 2\,\frac{macro\ p.\ macro\ r.}{macro\ p. + macro\ r.} \tag{10}$$

From that figure, we can conclude that the achievement of the *ob*CRF measured with the micro p./r. metric for 36 types (~ 77%) is the same as using CRFs that recognize only ~ 21 specialized types, while the F-measure metric reveals similar performance (~ 58.68%) between our approach for 36 types and CRFs with only ~ 16 types, clearly supporting our proposal.

6.3. Results with Cornell-RGBD

To further support the claimed high performance of *ob*CRFs , and to avoid jumping to conclusions based on the tests on a particular dataset, we have also evaluated our proposal with the Cornell-RGBD repository. The obtained results are reported in Tab. 2, being similar to the ones yielded with Robot@Home. The first four rows show the outcome of standard CRFs: the first one for graphs only containing nodes modeling specialized object types, achieving a success (micro p./r.) of ~ 55.6%, the second row includes contextual relations and increases that figure by ~ 10%, while the third and fourth rows report a success percentage of ~ 63% categorizing objects into general types and ~ 68.5% also considering context, respectively.

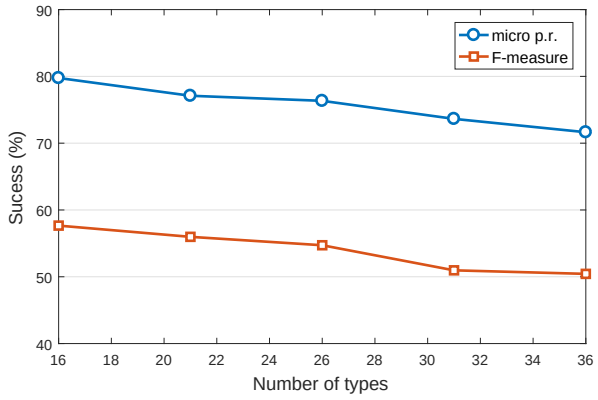

Figure 4: Evolution of the performance achieved by CRFs considering a different number of specialized object types.

The performance yielded by *ob*CRFs is reported in the next three rows. The configuration in the fifth row employs nodes of both, specialized and general categories, and edges only among nodes of different types. It achieves a success percentage of ~ 60% for specialized objects and ~ 69% for general ones, again proving the mutual influence of their categorization. The next row adds the contextual relations between specialized nodes to the model, further increasing the success. Finally, the last row integrates all the possible contextual relations, reaching the best performance: a success percentage of ~ 70% and ~ 76.5% for specialized and general types, respectively.

These figures provide an increment of ~ 4% in the micro p./r. metric classifying specialized types, and of ~ 8% with general ones. Regarding the reported macro precision and recall, while categorizing specialized types, *ob*CRFs achieve an increase of ~ 8% and ~ 7.5% respectively w.r.t. standard CRFs , and of ~ 6.5% and ~ 8% dealing with general ones. As a closing remark, these figures are considerably higher than those reported by CRFs using the initial set of 17 objects (see [8] for further information).

6.4. Analysis of computational time

This section measures the effect that the inclusion of additional nodes and edges by *ob*CRFs has on the training/LBP inference efficiency. Unless otherwise indicated, the provided measurements represent the average execution time from cross-validation. On the one hand, working with Robot@Home, the training process of the *ob*CRF in the last configuration (which has to be carried out only once in the model design phase [8]) took 12 minutes, while the inference process took 4ms. with a maximum execution time of 49ms. The standard CRF employed 3ms for inference, with a maximum of 17ms. On the other hand, the same *ob*CRF configuration considering Cornell-RGBD needed ~ 3 minutes to complete the training, while the inference processes took 4ms. with a maximum time of 20ms. In this case, the standard CRF spent 2ms. on average for performing inference, with a maximum execution time of 7ms. These numbers reveal that *ob*CRFs also keep low inference times. The experiments were run on a computer with an Intel Core i7-3820 at 3.60GHz. microprocessor and a RAM memory of 4x4GB. DDR3 at 1,600MHz. As introduced in the following discussion, further efficiency gains could be achieved by considering nodes only in certain *ob*CRF levels, depending on the goal of the application at hand.

7. CONCLUSIONS

In this work, we have proposed the utilization of Ontology-based Conditional Random Fields (*ob*CRF) for the recognition of objects by autonomous and/or assistance systems, like robots, autonomous vehicles, wearables for impaired people, etc. For that, we have described how CRFs are applied to the object recognition problem, as well as how ontologies structure the information of a domain of discourse, which are the building blocks of the proposed model. *ob*CRFs mimic the subsumption ordering of ontologies, borrowing the idea of making

Table 2: Recognition results yielded by CRF and *ob*CRF models with different configurations within the Cornell-RGBD dataset. See Tab. 1 for a description of the symbols in the *Configuration* column.

		Specialized types			General types		
Model	Configuration	micro p./r.	macro p.	macro r.	micro p./r.	macro p.	macro r.
CRF	$\mathcal{V}_s$	55.64%	31.85%	30.64%	–	–	–
	$\mathcal{V}_s + \mathcal{E}_{ss}$	65.97%	42.42%	39.07%	–	–	–
	$\mathcal{V}_g$	–	–	–	62.93%	53.58%	53.57%
	$\mathcal{V}_g + \mathcal{E}_{gg}$	–	–	–	68.58%	62.99%	59.32%
*ob*CRF	$\mathcal{V}_s + \mathcal{V}_g + \mathcal{E}_{sg}$	60.31%	41.78%	39.10%	69.10%	62.47%	60.41%
	$\mathcal{V}_s + \mathcal{E}_{ss} + \mathcal{V}_{gg} + \mathcal{E}_{sg}$	68.83%	48.96%	45.79%	75.90%	68.03%	65.06%
	$\mathcal{V}_s + \mathcal{E}_{ss} + \mathcal{V}_{gg} + \mathcal{E}_{sg} + \mathcal{E}_{gg}$	**69.91%**	**50.45%**	**46.52%**	**76.75%**	**69.47%**	**67.19%**

use of hierarchical structures were general concepts are defined over more specialized ones. This permits the recognition model to include nodes in the CRF graph representing coarse types of objects sharing some attributes (*e.g.* furniture, building structure, appliance, etc.), which are easier to recognize, and employ this subsumption relation to assist their joint classification into more specialized types (*e.g.* cabinet, closet, floor, wall, oven, etc.). By doing so, the performance of standard CRFs is improved, and the information provided (multiple categorizations with different granularity for each scene object) is richer and exploitable for the execution of high-level tasks.

The suitability of the proposal has been assessed in different ways. First, the recognition success of *ob*CRFs have been compared with those yielded by standard CRFs with two datasets: Robot@Home and Cornell-RGBD, achieving the highest performance when classifying objects into specialized ($\sim$ 71.5% vs. $\sim$ 77% and $\sim$ 66% vs. $\sim$ 70%, respectively) and more general types ($\sim$ 73.3% vs. $\sim$ 85.1% and $\sim$ 69% vs. $\sim$ 77%). Then, we have analyzed the performance of CRFs working with a different number of categories and found that (dealing with the Robot@Home dataset) *ob*CRFs achieved the same results considering 36 specialized types than CRFs with only 16. Finally, the computational time required by *ob*CRFs has been studied, concluding that the addition of additional edges and nodes does not compromise the efficiency of the LBP method to perform inference, keeping short execution times (4ms. on average for both datasets).

In the future, we plan to exploit *ob*CRFs for the efficient execution of high-level robotic tasks, like the search of an object of a certain type by the robot. In this scenario, *ob*CRFs could first perform inference in a top level with general types, and then consider, in a bottom level with specialized ones, the nodes that have a high probability of belonging to the searched type, leading to an efficient and robust operation. Another option is, in scenarios where the searched object is not found, to replace it with other detected object with similar functionality: *e.g.* replace a mug by a glass (information that can be naturally codified into the ontology). The utilization of *ob*CRFs opens interesting possibilities for addressing these issues in a novel and promising way. Additionally, we also consider the utilization of a CNN as a baseline detector and the refinement of its results by an *ob*CRF, also giving spatial and temporal consistency to them.

Acknowledgements

This work is supported by the research projects *WISER* ([DPI2017-84827-R]), funded by the Spanish Government, and financed by European Regional Development's funds (FEDER), and *MoveCare* ([ICT-26-2016b-GA-732158]), funded by the European H2020 program, and by a postdoc contract from the I-PPIT-UMA program financed by the University of Málaga.

[1] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: A survey, IEEE Signal Processing Magazine 35 (1) (2018) 84–100. doi:10.1109/MSP.2017.2749125.

[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. G. Rodríguez, A review on deep learning techniques applied to semantic segmentation, CoRR abs/1704.06857.
URL http://arxiv.org/abs/1704.06857

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137–1149. doi:10.1109/TPAMI.2016.2577031.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, Cham, 2016.
URL http://arxiv.org/abs/1512.02325

[5] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., USA, 2016, pp. 379–387.
URL http://dl.acm.org/citation.cfm?id=3157096.3157139

[6] L. E. Sucar, Probabilistic graphical models, Advances in Computer Vision and Pattern Recognition. London: Springer London. doi 10 (2015) 978–1.

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

[8] J.-R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, A survey on learning approaches for undirected graphical models. Application to scene object recognition, International Journal of Approximate Reasoning 83 (C) (2017) 434–451. doi:10.1016/j.ijar.2016.10.009.
URL http://www.sciencedirect.com/science/article/pii/S0888613X16302043

[9] X. Liu, H. Li, W. Meng, S. Xiang, X. Zhang, 3d point cloud classification based on discrete conditional random field, in: International Conference on Technologies for E-Learning and Digital Entertainment, Springer, 2017, pp. 115–137.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected

CRFs, in: International Conference on Learning Representations (ICLR), 2015.
[11] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr, Conditional random fields as recurrent neural networks, in: The IEEE International Conference on Computer Vision (ICCV), 2015.
[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 834–848.
[13] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.
[14] J.-L. Blanco, J.-A. Fernández-Madrigal, J. González-Jiménez, Towards a unified bayesian approach to hybrid metric-topological slam, IEEE Transactions on Robotics 24 (2) (2008) 259–270.
[15] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Robot@home, a robotic dataset for semantic mapping of home environments, The International Journal of Robotics Research 36 (2) (2017) 131–141. doi:10.1177/0278364917695640.
URL https://doi.org/10.1177/0278364917695640
[16] 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.
[17] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor Segmentation and Support Inference from RGBD Images, in: Proc. of the European Conference on Computer Vision (ECCV 2012), 2012, pp. 746–760.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer International Publishing, Cham, 2014, pp. 740–755.
[19] D. Wolf, J. Prankl, M. Vincze, Fast semantic segmentation of 3d point clouds using a dense crf with learned parameters, in: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 2015.
[20] N. Pinto, D. D. Cox, J. J. DiCarlo, Why is real-world visual object recognition hard?, PLOS Computational Biology 4 (1) (2008) 1–6.
[21] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, The Knowledge Engineering Review 11 (1996) 93–136.
[22] M. A. Gutierrez, L. Manso, P. N. Trujillo, P. Bustos, Planning object informed search for robots in household environments, 2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) (2018) 205–210.
[23] C. Wu, I. Lenz, A. Saxena, Hierarchical semantic labeling for task-relevant rgb-d perception, in: Robotics: Science and Systems (RSS), 2014.
[24] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Building multiversal semantic maps for mobile robot operation, Knowledge-Based Systems 119 (2017) 257–272. doi:10.1016/j.knosys.2016.12.016.
URL https://doi.org/10.1016/j.knosys.2016.12.016
[25] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Exploiting semantic knowledge for robot object recognition, Knowledge-Based Systems 86 (2015) 131–142.
[26] 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.
[27] L. Chang, M. M. Duarte, L. Sucar, E. F. Morales, A bayesian approach for object classification based on clusters of SIFT local features, Expert Systems with Applications 39 (2) (2012) 1679 – 1686.
[28] 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.
[29] W. L. Hoo, C. H. Lim, C. S. Chan, Keybook: Unbias object recognition using keywords, Expert Systems with Applications 42 (8) (2015) 3991 – 3999.
[30] M. Pontil, A. Verri, Support vector machines for 3d object recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (6) (1998) 637–646.
[31] 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.
[32] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE transactions on pattern analysis and machine intelligence 32 (9) (2010) 1627–1645.
[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.
URL http://dx.doi.org/10.1007/s11263-015-0816-y
[34] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 6517–6525.
[35] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning transferable architectures for scalable image recognition, CoRR abs/1707.07012.
[36] C. Galleguillos, S. Belongie, Context based object categorization: A critical survey, Computer Vision and Image Understanding 114 (6) (2010) 712–722. doi:10.1016/j.cviu.2010.02.004.
[37] A. Oliva, A. Torralba, The role of context in object recognition, Trends in Cognitive Sciences 11 (12) (2007) 520–527.
[38] 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.
[39] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al., The limits and potentials of deep learning for robotics, The International Journal of Robotics Research 37 (4-5) (2018) 405–420.
[40] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831–1840.
[41] 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.
[42] X. Ren, L. Bo, D. Fox, Rgb-(d) scene labeling: Features and algorithms, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), 2012, pp. 2759–2766.
[43] F. Husain, L. Dellen, C. Torras, Recognizing point clouds using conditional random fields, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 4257–4262.
[44] 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.
[45] J. R. Ruiz-Sarmiento, C. Galindo, J. Gonzalez-Jimenez, Probability and Common-Sense: Tandem Towards Robust Robotic Object Recognition in Ambient Assisted Living, Vol. 10070 of Lecture Notes in Computer Science, Springer, 2016.
[46] S. G. Kosov, P. Kohli, F. Rottensteiner, C. Heipke, A two-layer conditional random field for the classification of partially occluded objects, CoRR abs/1307.3043.
URL http://arxiv.org/abs/1307.3043
[47] L. Albert, F. Rottensteiner, C. Heipke, A two-layer conditional random field model for simultaneous classification of land cover and land use, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives 40 (2014) 40 (3) (2014) 17–24.
[48] L. Sulimowicz, I. Ahmad, A. Aved, A multi-layer approach to superpixel-based higher-order conditional random field for semantic image segmentation, arXiv preprint arXiv:1804.02032.
[49] Q. Huang, M. Han, B. Wu, S. Ioffe, A hierarchical conditional random field model for labeling and segmenting images of street scenes, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1953–1960.
[50] J. Reynolds, K. Murphy, Figure-ground segmentation using a hierarchical conditional random field, in: Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on, 2007, pp. 175–182. doi:10.1109/CRV.2007.32.
[51] M. Y. Yang, W. Frstner, A hierarchical conditional random field model for labeling and classifying images of man-made scenes, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Work-

shops), 2011, pp. 196–203.

[52] 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.

[53] 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.

[54] R. Capobianco, J. Serafin, J. Dichtl, G. Grisetti, L. Iocchi, D. Nardi, A proposal for semantic map representation and evaluation, in: Mobile Robots (ECMR), 2015 European Conference on, 2015, pp. 1–6.

[55] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, OWL Web Ontology Language reference, W3C Recommendation (2004).

[56] R. Gonçalves, M. Horridge, M. Musen, C. Nyulas, S. Tu, T. Tudorache, Protégè home page, http://protege.stanford.edu/, [Online; accessed 26-June-2015] (2015).

[57] S. Harris, A. Seaborne, E. Prudhommeaux, Sparql 1.1 query language, W3C recommendation 21 (10).

[58] 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.

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

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

[61] 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. doi:10.1109/IROS.2005.1545511.

[62] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask r-cnn, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2980–2988.

[63] A. Arnab, S. Jayasumana, S. Zheng, P. H. Torr, Higher order conditional random fields in deep neural networks, in: European Conference on Computer Vision, Springer, 2016, pp. 524–540.

[64] J. R. Ruiz-Sarmiento, M. Guenther, C. Galindo, J. Gonzalez-Jimenez, J. Hertzberg, Online context-based object recognition for mobile robots, in: 17th International Conference on Autonomous Robot Systems and Competition (ICARSC), 2017.

[65] M. Gnther, J. Ruiz-Sarmiento, C. Galindo, J. Gonzlez-Jimnez, J. Hertzberg, Context-aware 3d object anchoring for mobile robots, Robotics and Autonomous Systems 110 (2018) 12 – 32. doi:https://doi.org/10.1016/j.robot.2018.08.016.
URL http://www.sciencedirect.com/science/article/pii/S0921889017307856

[66] G. Ghiasi, C. C. Fowlkes, Laplacian pyramid reconstruction and refinement for semantic segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 519–534.

[67] S. Chandra, I. Kokkinos, Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs, in: European Conference on Computer Vision, Springer, 2016, pp. 402–418.

[68] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, OLT: A Toolkit for Object Labeling Applied to Robotic RGB-D Datasets, in: European Conference on Mobile Robots, 2015.

[69] J. R. 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.