Draft Version. Final version published in the International Conference on Intelligent Robots and Systems (IROS), 2015.

Joint Categorization of Objects and Rooms for Mobile Robots*

J.R. Ruiz-Sarmiento, C. Galindo and J. Gonzalez-Jimenez

***Abstract*— In general, the problems of objects' and rooms' categorizations for robotic applications have been addressed separately. The current trend is, however, towards a joint modelling of both issues in order to leverage their mutual contextual relations: *object* $\rightarrow$ *room* (e.g. the detection of a microwave indicates that the room is likely to be a kitchen), and *room* $\rightarrow$ *object* (e.g. if the robot is in a bathroom, it is probable to find a toilet). *Probabilistic Graphical Models* (PGMs) are typically employed to conveniently cope with such relations, relying on inference processes to hypothesize about objects' and rooms' categories. In this work we present a *Conditional Random Field* (CRF) model, a particular type of PGM, to jointly categorize objects and rooms from RGBD images exploiting *object-object* and *object-room* relations. The learning phase of the proposed CRF uses *Human Knowledge* (HK) to eliminate the necessity of gathering real training data. Concretely, HK is acquired through elicitation and codified into an ontology, which is exploited to effortless generate an arbitrary number of representative synthetic samples for training. The performance of the proposed CRF model has been assessed using the NYU2 dataset, achieving a success of $\sim 70\%$ categorizing both, objects and rooms.**

I. INTRODUCTION

A robot performing in human environments has to manage a rich representation of its surroundings for the execution of tasks like navigation, fetch-and-carry, surveillance, etc. Such a world representation has to support the semantics of the human concepts and their relations. That is, the robot must be able to *understand* human knowledge, e.g. "A kitchen is a room where you can find an oven", permitting the human to express his/her orders using natural, and probably incomplete, sentences, e.g. "Please check the oven". The spatial awareness needed by the robot to accomplish this task must account for the existing close relations between objects and their typical locations. Thus, in this context, the robot should solve i) the so-called *room categorization* problem, i.e. to infer the type of space where it is, and ii) the *object categorization* problem, i.e. to classify the perceived objects.

Recent publications (e.g. [1], [2]) have shown that the joint modelling of the object and room categorization problems can outperform other methods that address them separately [3]–[6]. Holistic approaches exploit the fact that objects are located in rooms according to their functionality, so the presence of an object of a certain type is a hint for the room categorization [7]–[9]. Likewise, the category of a room is a good indicator of the object categories that can be found inside [10]. Besides, objects are not placed randomly, but following configurations that make sense from a human perspective [11], [12]. Thereby, the exploitation of these object-object and object-room contextual clues provides categorization methods with useful information.

A recurrently resorted framework to model contextual information is the so-called *Probabilistic Graphical Models* (PGMs) [13]. PGMs permit a categorization system to conveniently model a room, the objects inside, and their contextual relations. Such a representation handles the uncertainty latent in the robot sensing system, and supports the execution of probabilistic inference algorithms (e.g. ICM [14] or LBP [15]). However, a significant drawback of these models is that they require a learning phase where the training dataset must be large and comprehensive enough to properly capture the variability of the domain at hand.

In this work we present a *Conditional Random Field* model (CRF) [13], a particular type of PGM, which enables the joint categorization of objects and rooms by exploiting their contextual relations. A distinctive feature of our approach is the utilization of *Human Knowledge* (HK) during the training phase, removing, thus, the arduous task of gathering real datasets. Concretely, we rely on the acquisition of HK about objects' and rooms' categories through elicitation and its codification into an *ontology* [16]. The advantage of using HK for training CRFs has been proven in [17].

Our approach has been tested with home RGBD scenes from the NYU2 dataset [18] (see figure 1-left). This dataset is employed as a testbed by state-of-the-art methods given its size and challenging features. For example, it is utilized in [1], also employing a CRF, and achieving a success of $\sim 60.5\%$ and $\sim 58.7\%$ recognizing objects and rooms respectively. Although a fair comparison is not possible since the authors consider a different set of object categories and room types, it permits us to qualitatively confirm the promising performance of our approach, which yields a success of $\sim 70\%$ for categorizing both objects and rooms.

II. CONDITIONAL RANDOM FIELDS. APPLICATION TO THE JOINT CATEGORIZATION OF OBJECTS AND ROOMS

The joint room and object categorization problem can be stated as the assignation of classes to both a given area of the robot workspace and the objects within, taking into account their observed geometric/appearance features and contextual relations. The following definitions are required in order to set the problem from a probabilistic stance:

- Let $\boldsymbol{o} = [o_1, .., o_n]$ be a vector of n observed objects, each one characterized through a number of features: size, height, orientation, etc.

*Work funded by the Spanish grant program *FPU-MICINN 2010* and the Spanish projects *TAROTH* (DPI2011-25483) and *PROMOVE* (DPI2014-55826-R), both co-founded by *Fondo Europeo de Desarrollo Regional*.

All authors are with the Department of System Engineering and Automation, University of Málaga, Campus de Teatinos, 29071, Málaga, Spain. Corresponding author J.R. Ruiz-Sarmiento, email: jotaraul@uma.es

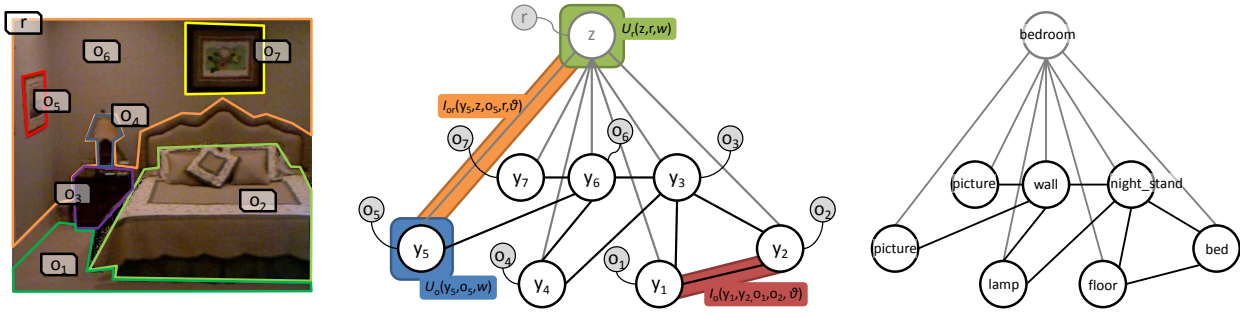

Fig. 1. Left, a coloured point cloud of a room (r) with a number of segmented objects (from o_1 to o_7), extracted from the NYU2 dataset. Middle, the graph structure of a CRF modelling the objects in that room, the room itself, and their contextual relations. Each random variable y_i is associated to an observed object o_i, while z is related to r. The coloured parts indicate the scope of: an object unary factor – blue, a room unary factor – green, an object-object pairwise factor – red, and an object-room pairwise factor – orange. Right, result of a probabilistic inference process over the CRF.

- Let r be the observed room described by a set of features: size, color, etc.
- Define $L_o = \{l_{o_1}, .., l_{o_k}\}$ as the set of the k considered object categories (e.g. bed, oven, towel, etc.)
- Define $L_r = \{l_{r_1}, .., l_{r_j}\}$ to be the set of the j considered room categories (e.g. kitchen, bedroom, etc.).
- Define $\boldsymbol{y} = [y_1, .., y_n]$ to be a vector of discrete random variables assigning a category from L_o to each object in $\boldsymbol{o}$.
- Let $z \mid z \in L_r$ be a discrete random variable assigning a room category from L_r to r.

Thereby, the joint categorization process, modelled through a Conditional Random Field (CRF), consists of maximizing the probability distribution $P(\boldsymbol{y}, z \mid \boldsymbol{o}, r)$, i.e, to find the most probable room's and objects' categories given their characterized observations. CRFs exploit the concept of independence to break this distribution down into smaller pieces, since its high dimensionality prevents an exhaustive definition. A CRF is represented as a graph $H = (V, E)$, where V is a set of nodes representing random variables, and E a set of edges linking dependant/related nodes. In the addressed problem, a node represents a random variable, i.e. y_i or z, while an edge can set two types of dependencies: (a) between two close objects in the room, or (b) between an object and the room containing it. In figure 1, an example of a relation of type (a) is the one between the *night stand* (o_3) and the *lamp* (o_4), while all the relations between the objects (from o_1 to o_7) and the room (r) are examples of relations of type (b). Thus, the categorization of an object affects the categorization of nearby objects, but not those placed far away, while the categorization of a room and its constituent objects has a mutual influence.

According to the Hammersley-Clifford theorem [13], the distribution $P(\boldsymbol{y}, z|\boldsymbol{o}, r)$ can be factorized over H as a product of factors, being a factor a function that represents a probability distribution over a part of H. In this work we have considered four factor types: two unary factors applicable to nodes (object and room unary factors), and two pairwise factors associated to edges (object-object and object-room pairwise factors).

For convenience, the factorization of $P(\boldsymbol{y}, z|\boldsymbol{o}, r)$ over the graph H is expressed by means of log-linear models as:

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

where $Z(\cdot)$ is the normalizing partition function so $\sum_{\xi(\boldsymbol{y},z)} p(\boldsymbol{y}, z|\boldsymbol{o}, r, \boldsymbol{\omega}, \boldsymbol{\theta}) = 1$, being $\xi(\boldsymbol{y}, z)$ an assignation to the variables in $\boldsymbol{y}$ and z, and $\epsilon(\cdot)$ the so-called energy function, which in this work is defined as:

$$\epsilon(\boldsymbol{y}, z, \boldsymbol{o}, r, \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{i \in V_o} U_o(y_i, o_i, \boldsymbol{\omega}) + U_r(z, r, \boldsymbol{\omega}) + \sum_{(i,j) \in E_o} I_o(y_i, y_j, o_i, o_j, \boldsymbol{\theta}) + \sum_{(i,j) \in E_{or}} I_{or}(y_i, z, o_i, r, \boldsymbol{\theta}) \quad (2)$$

being V_o the subset of V containing the nodes associated to variables from $\boldsymbol{y}$, E_o the subset of E entailing the edges that link nodes in V_o, and $E_{or} = E - E_o$, i.e. the edges connecting nodes representing objects with a room node. $U_o(\cdot)$, $U_r(\cdot)$, $I_o(\cdot)$ and $I_{or}(\cdot)$ define the employed factors (see figure 1).

Object unary factor ($U_o(\cdot)$). This factor encodes the likelihood of assigning objects categories from L_o to the random variable y_i, given the features extracted from the object o_i, e.g. height, size, elongation, etc. It is defined as a linear classification model as follows:

$$U_o(y_i, o_i, \boldsymbol{\omega}) = \sum_{l \in L_o} \delta(y_i = l) \boldsymbol{\omega}_l f_o(o_i) \quad (3)$$

where $f_o(o_i)$ is a function that computes the features' vector $\boldsymbol{f}_{o_i}$, $\boldsymbol{\omega}_l = [\omega_{1,l}, .., \omega_{|\boldsymbol{f}_{o_i}|,l}]$ is a vector of weights for each class $l \in L_o$ obtained during the training phase, and $\delta(y_i = l)$ is the Kronecker delta function that takes value 1 when $y_i = l$ and 0 otherwise. The features used to characterize an object are: orientation, planarity, and size of its bounding box, area of its two principal directions, height from the floor, and color hue variation.

Room unary factor ($U_r(\cdot)$). The factor represented by the following linear model:

$$U_r(z, r, \boldsymbol{\omega}) = \sum_{l \in L_r} \delta(z = l) \boldsymbol{\omega}_l f_r(r) \quad (4)$$

encodes the likelihood of the random variable z to belong to the different room types from L_r given the features extracted from the observation r, e.g. size, number of objects, color hue variation, etc. In this case, $f_r(r)$ is the function that computes such a vector of features $\boldsymbol{f}_r$, being the vector of weights $\boldsymbol{\omega}_l = [\omega_{1,l}, .., \omega_{|\boldsymbol{f}_r|,l}]$ associated to the classes in L_r. The features used are: size of the room bounding box, number of objects within the room, and variation of color hue.

Object-object pairwise factor ($I_o(\cdot)$). Nodes related with objects that appear close in the scene are linked by an edge in the CRF. Thus, the object-object pairwise factor is in charge of stating the compatibility of a pair of categories assigned to these nodes. Again, a linear classification model is employed:

$$I_o(y_i, y_j, o_i, o_j, \boldsymbol{\theta}) = \sum_{l_1 \in L_o} \sum_{l_2 \in L_o} \delta(y_i = l_1)\delta(y_j = l_2)\boldsymbol{\theta}_{l_1,l_2} g_o(o_i, o_j) \quad (5)$$

where $g_o(o_i, o_j)$ computes a vector of features $\boldsymbol{f}_{o_i o_j}$ to characterize the relation between objects o_i and o_j, and $\boldsymbol{\theta}_{l_1,l_2} = [\theta_{1,l_1,l_2}, .., \theta_{|\boldsymbol{f}_{o_i o_j}|,l_1,l_2}]$ is a vector of weights, learnt during the training phase, for each pair of classes in L_o. The features characterizing object-object relations are: difference between principal directions, vertical distance of centroids, volume ratio, connectivity and object-object compatibility.

Object-room pairwise factor ($I_{or}(\cdot)$). This encodes the compatibility of finding an object of a certain category into a room of type l_{r_i}, as well as the compatibility of being in a room of a certain category having perceived an object of type l_{o_i}. Its linear classification model is defined as:

$$I_{or}(y_i, z, o_i, r, \boldsymbol{\theta}) = \sum_{l_1 \in L_o} \sum_{l_2 \in L_r} \delta(y_i = l_1)\delta(z = l_2)\boldsymbol{\theta}_{l_1,l_2} g_{or}(o_i, r) \quad (6)$$

being $g_{or}(o_i, o_j)$ a function that yields a fixed value $f_{o_i r}$. Therefore, the learnt vector of weights $\boldsymbol{\theta}_{l_1,l_2}$ for each pair of classes in $(l_1, l_2) \mid (l_1 \in L_o, l_2 \in L_r)$ states the object-room compatibility.

Training and Inference over the CRF. The training of the CRF model, i.e. the learning of the vectors of weights $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$, is performed by means of the optimization of the so-called pseudo-likelihood function, a tractable, alternative objective function to the computationally high-demanding likelihood one [13]. To fed this learning process we employ representative synthetic samples of the domain, which are generated as explained in the next section.

Once trained, the CRF is used to categorize rooms and objects through probabilistic inference. We resort to the Iterated Conditional Modes (ICM) algorithm [14], an efficient, approximated inference method that performs by maximizing local conditional probabilities. Figure 1-right shows the results yielded by this method over the CRF of the figure 1-middle.

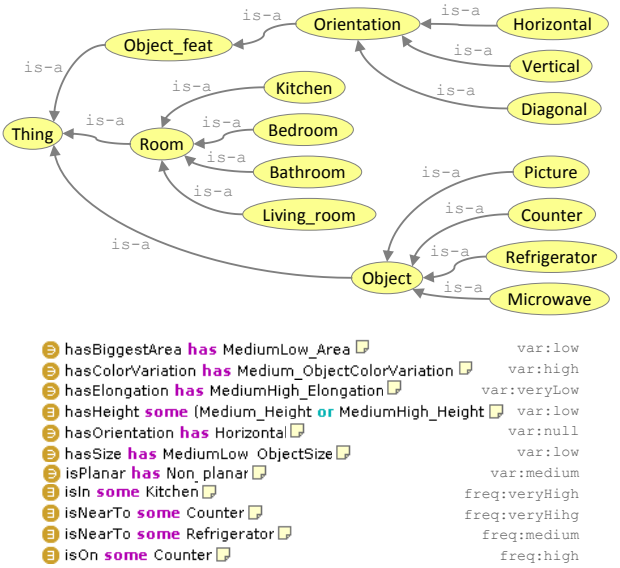

Fig. 2. Top, excerpt of the used ontology. Bottom, definition of the concept `Microwave`.

III. From Human Knowledge to Training Data

The proposed CRF model for the categorization of rooms and objects is tuned following a top-down design. First, knowledge of the domain at hand is collected through human elicitation. This information is codified into an ontology by means of the definition of concepts, e.g. `Kitchen`, and relations, e.g. `Microwave isIn Kitchen` (see section III-A), and then exploited for the generation of an arbitrary number of representative, synthetic training samples (see section III-B). The generated data feed an optimization process that iteratively tunes the CRF parameters defined in section II.

On the other hand, the categorization process performs in a bottom-up fashion. Given a RGBD observation of a room, its constituent objects are segmented and characterized through a set of features (e.g. their size, height, etc.). The RGBD observation itself is also processed in order to characterize the room according to its geometry and appearance. Then, a number of object-object and object-room relations are computed according to the objects' features and locations. Finally, a probabilistic inference process over the trained CRF yields their most probable categories employing: i) objects' features, ii) room's features, and iii) contextual relations.

A. Codification of Human Knowledge

In this work we rely on human knowledge (HK) encoded in an ontology. An ontology is an explicit specification of a conceptualization related to a domain, which entails *concepts*, *relations*, and *individuals*. In the case of a home domain, examples of concepts are `Kitchen`, or `Microwave`, a relation can be stated as `Microwave isIn Kitchen`, and `kitchen-1` or `microwave-3` identify individuals, i.e. instantiations of concepts. The use of HK encoded in

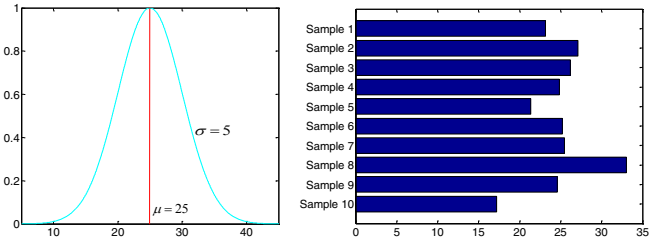

Fig. 3. *Left, unnormalized Gaussian distribution* for the size of a kitchen (in m^3) built according to its definition into the ontology. Right, samples drawn from that distribution to characterize the size of 10 kitchens.

ontologies for mobile robotics exhibits significant advantages for a variety of applications, as reported in [19], [20].

Figure 2-top depicts an excerpt of the ontology used during the conducted experiments, showing some concepts and relations[1]. Figure 2-bottom shows the definition of the `Microwave` concept setting their usual features (geometry and appearance), as well as their contextual relations. It states, for example, that microwaves usually share a medium size, and are placed near counters, within kitchens. This information is collected from humans through an elicitation process, and it is straightforwardly codified into the ontology given its capability to naturally encode notions from natural language. Nevertheless, some human concepts need to be transformed into crispy values as required in our system. For that, the *hasValue* property is added in the ontology to quantify human concepts, like for instance `Vertical`, `Horizontal`, or `Diagonal`. These concepts allow an easy codification of object properties such as `Floor hasOrientation Horizontal` or `Picture hasOrientation Vertical`. The *hasValue* property assigns a crispy value to these concepts (in degrees) that is also gathered through elicitation, e.g. `Vertical hasValue 90`, `Horizontal hasValue 0`, and `Diagonal hasValue 45`.

In order to cope with the inherent variability of the considered domain, our approach annotates properties and relations with an element from the set R_A ={*null, veryLow, low, medium, high, veryHigh*}. For example in the definition of the microwave concept (see figure 2-bottom) the size feature has been annotated with a `veryLow` variability indicating that most of microwaves exhibit similar dimensions. Similarly, these annotations are also used to express the frequency of the object-object and object-room relations. For example, the annotation `Microwave isNear Counter freq:high` sets that microwaves are usually found close to a counter, while the definition `Microwave isIn Kitchen freq:veryHigh` expresses that it is highly probable to find a microwave in a kitchen.

B. Generation of training data

Once the HK about the home domain has been encoded, we use it for the generation of synthetic training data. The presented process can be repeated to generate an arbitrary number of samples, and no human participation is longer required [17]. For clarity sake, it is explained the process for the generation of a synthetic sample reifying a kitchen, but the methodology is the same for any room category:

1) **Room characterization**. The first step is the computation of the room features which, in the used ontology, includes its size (m^3) and color hue variation. For that, a *Gaussian distribution* $\mathcal{N}(\mu, \sigma)$ is considered for each feature, where the mean μ corresponds to the crispy value of the property, while the standard deviation σ symbolizes the annotated variability. For example, given a definition of kitchens where they show a `Medium` size, being `Medium_RoomSize hasValue 25`, and an annotation of *medium* variability[2], the Gaussian distribution results $\mathcal{N}(25, 5)$ (see figure 3-left). The function $f_{sr}(l_r)$ draws a sample from this distribution to get the size of a particular room (see figure 3-right), where l_r represents the kitchen category in this case, and repeats this process with the remaining room features. This function replaces $f_r(r)$ in equation 4 during the CRF training.
2) **Inclusion of objects in the room**. The inclusion of objects in the synthetic room is decided according to the `isIn` property. Only objects that contains the property `isIn value Kitchen` in their definitions are possible candidates. The inclusion of candidates depends on a probability distribution based on their frequency annotations. For example, the `Microwave` category is defined as `isIn value Kitchen freq:high`, which is translated to $P(Microwave_{appearing}) = 0.8$ and $P(Microwave_{notAppearing}) = 0.2$. Samples drawn from these distributions yield the final set of included objects, as it is illustrated in table I-top.
3) **Object characterization**. This step is similar to 1), but considering the properties defined over the objects included in the second step. A number of Gaussian

TABLE I

TOP, EXAMPLE OF OBJECTS INCLUDED IN A ROOM OF TYPE KITCHEN. BOTTOM, OBJECTS RELATED WITH AN INCLUDED MICROWAVE.

Concept	frequency	$P(appearing)$	sample
`Bottle`	*medium*	0.5	not appearing
`Cabinet`	*veryHigh*	0.95	appearing
`Chair`	*medium*	0.5	not appearing
`Counter`	*veryHigh*	0.95	appearing
`Dishwasher`	*high*	0.8	not appearing
`Floor`	*always*	1	appearing
`Microwave`	*high*	0.8	appearing
`Picture`	*low*	0.2	not appearing
`Refrigerator`	*veryHigh*	0.95	appearing
`Stove`	*veryHigh*	0.95	appearing
`Table`	*medium*	0.5	not appearing
Concept	**frequency**	$P(related)$	**sample**
`Cabinet`	*high*	0.8	near
`Counter`	*veryHigh*	0.95	near
`Floor`	*veryLow*	0.05	not near
`Refrigerator`	*medium*	0.5	not near
`Stove`	*medium*	0.5	near

[1]This ontology and other resources are available online at: http://mapir.isa.uma.es/work/objects-rooms-categorization.

[2]To get the standard deviation (σ) of a feature, the variabilities are considered to be a percentage of the crispy values of the properties that they are annotating within the ontology. In this case, being the crispy value 25, and corresponding *medium* to its 20%, the standard deviation is 5.

distributions $N(\mu, \sigma)$ are built according to the different objects' geometric/appearance properties and their annotations, while the function $f_{so_i}(l_{o_i})$ draws samples from them to characterize each included object o_i. This function is used instead of $f_o(o_i)$ for learning the model (recall equation 3).

4) **Object-object context creation**. The contextual relations between objects are established by the `isNear` properties and their annotations. In a similar way to the inclusion of objects, the likelihood of these relations is modelled by a probability distribution according to how frequently two objects appear close to each other in a Kitchen. For example, following the definition of the concept `Microwave`, they are often found near a counter, though it is more uncommon to find them near a table. As an illustrative example, table I-bottom shows the relations established for a microwave and the rest of objects included in a kitchen (in table I-top).
5) **Object-object context characterization**. Different features can be computed to add valuable contextual information to the relations between two objects, e.g. difference of size, difference of height, perpendicularity, etc. These features can be easily computed from the object features extracted in the third step.
   In addition to these context features, two boolean properties are added: `isOn` and `isUnder`, which state if an object is placed on/under another one.
   The function in charge of compiling and yielding this information is $g_{so}(f_{o_i}, f_{o_j})$, being $f_{o_i} = [f_{so_i}(l_{o_i}), l_{o_i}]$, which replaces $g_o(o_i, o_j)$ in equation 5.
6) **Object-room context characterization**. The relation between the room and its objects is characterized by a fixed value, as it is the training process of the CRF which learns automatically the likelihood of finding an object of a certain type into a kitchen. The function $g_{sor}(l_{o_i}, r)$ provides this value, and plays the role of $g_{or}(o_i, r)$ during training (recall equation 6).

In summary, the above six steps yield the objects, room and contextual features needed to feed the unary and pairwise factors during the training of the CRF (equations 3-6). Figure 4 shows an example of a synthetic room represented in the form of a graphical model. It depicts the objects' and room's types, the functions in charge of characterizing them, and their contextual relations.

IV. Evaluation Results

In order to evaluate our approach, a number of CRFs have been tuned using synthetic samples (see section III-B). These CRFs differ in the type of features and factors employed, aiming to contrast the performance achieved by different configurations.

We have resorted to the NYU2 dataset as a testbed, which is widely employed in the literature given its number of scenes and their diverse nature. Concretely, we have extracted 208 RGBD scenes resembling rooms perceived by a robot visiting a home environment, equally divided into four categories: *bathroom*, *bedroom*, *kitchen* and *living-room*. These rooms are compound of a total of 1692 objects belonging to 26 different categories provided by the dataset, including *bottle*, *sink*, *toilet*, *towel*, *sofa*, *bed*, *microwave*, etc.

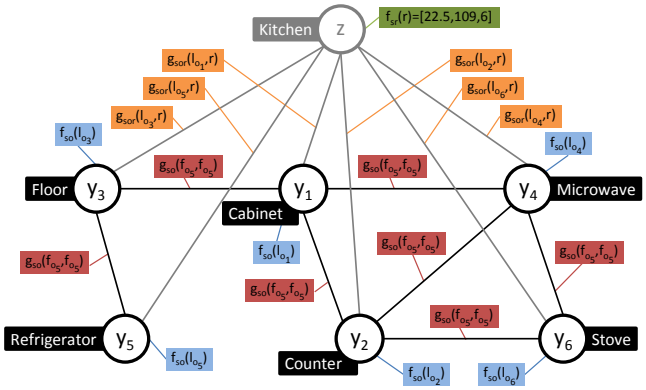

Fig. 4. Example of a CRF resultant from the generation of a synthetic room. The room type is a kitchen, with a total of 6 objects included (see table I). The resultant room's features are $f_{sr}(r) = [22.5, 109, 6]$, which correspond to its size, color hue variation, and number of objects.

TABLE II
Method evaluation results.

	Our approach		Trained with real data	
Configuration	**Object**	**Room**	**Object**	**Room**
Appearance	17.86	27.88	17.79	27.66
Geometry	62.50	46.63	43.85	41.91
App.+geo.	63.87	50.96	47.70	47.22
App.+geo.+obj-obj	66.29	50.96	48.88	47.22
App.+geo.+obj-room	67.48	61.22	49.61	58.09
All combined	**69.61**	**69.71**	**56.08**	**62.65**

In our experiments, the CRFs were trained with a dataset compound of 400 synthetic rooms, and their performance were measured by categorizing objects and rooms from the 208 NYU2 scenes. The implementation uses the Undirected Probabilistic Graphical Models library (UPGMpp) [21].

Table II (left part) shows the results obtained for the different CRF configurations employing our model. Note how the integration of additional features and contextual relations progressively increases the performance. The first group of configurations only considers unary factors, the second one includes object-object or object-room pairwise factors, while the last integrates all of them. A closer look at the data reveals how the integration of object-object contextual relations boosts the performance in categorizing objects a 2.5% w.r.t. a configuration relying only on object local features (appearance and geometry), while the categorization of rooms increases a 10.2% if the object-room relations are considered. The combination of both contextual relations augments these figures to 5.7% and 18.7% respectively, which highlights the benefits of a joint categorization of objects and rooms. Examples of rooms and objects categorized by this last configuration are depicted in figure 5-top. Figure 5-bottom-right reports the rooms' confusion matrix for the last configuration, where rows represent the ground truth information and columns the categorization results.

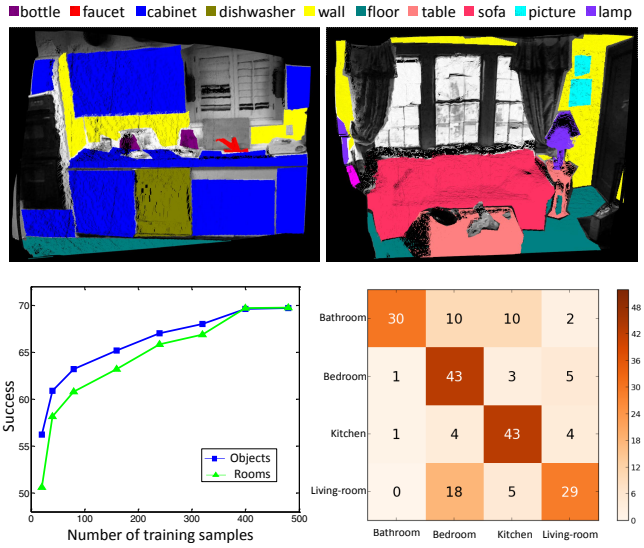

Fig. 5. Top, examples of a kitchen and a living-room correctly categorized as yielded by the method. Bottom-left, categorization success w.r.t. the number of samples used for training. Bottom-right, rooms' confusion matrix.

In order to validate the use of synthetic samples for training, the CRFs have been also trained and tested with the 208 NYU2 scenes following a 5-fold cross validation methodology. The results shown in the right part of table II reveal that despite the positive effect of using contextual relations, these CRFs exhibit a lower performance.

Notice that the proposed training methodology based on HK permits a robot to effortless generate the training dataset, which size largely influences on the results. Figure 5-bottom-left shows the categorization success yielded by a CRF trained with synthetic datasets of different sizes. It can be observed how the addition of more, representative training data boost the performance, from a 60.55% and 51.50% of success for object and room categorization respectively – 40 samples, up to 69.75% and 66.4% – 480 samples. This increment attenuates when the number of training samples approaches 500, which suggests that a success upper-limit can be reached despite the utilization of more samples. Notice that each training sample is compound of a room and its constituent objects so, for example, in the case of training with 480 rooms the number of objects is $\sim$4,900.

V. CONCLUSIONS AND FUTURE WORK

This work has presented a *Conditional Random Field* (CRF) model to jointly categorize objects and rooms into the workspace of a robot. A key feature of this model is that we rely on *Human Knowledge* to replace real training data with prototypal, synthetic samples of the domain codified in an ontology, which removes the tedious and time-consuming task of gathering a real dataset. Additionally, the utilization of an ontology enables the execution of high-level robotic tasks. The approach has been validated against home scenes from the NYU2 dataset, reaching a categorization success of $\sim 70\%$ for both objects and rooms. It is worth to mention that the applicability of the approach is not limited to robots working at home environments, but it is suitable to perform in other domains which properties and semantics can be defined by human elicitation, e.g. office facilities or hospitals.

From here, we plan to endow the system with the capability to identify new categories of rooms and objects. A first step towards this could be the utilization of a logical reasoner over the yielded categorization results in order to check their coherence w.r.t. the set of defined objects and rooms within the ontology.

REFERENCES

[1] Lin, D., Fidler, S., and Urtasun, R. (2013). Holistic Scene Understanding for 3D Object Detection with RGBD cameras. In *International Conference on Computer Vision (ICCV)*.

[2] Rogers, J.G., and Christensen, H.I. (2012). A Conditional Random Field Model for Place and Object Classification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1766-1772.

[3] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645.

[4] Oliva, A., Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. In *Int. J. of Computer Vision*, vol. 42, pp. 145-175.

[5] Quattoni, A., and Torralba, A. (2009). Recognizing Indoor Scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[6] Pronobis, A., Martínez Mozos, O., Caputo, B., and Jensfelt, P. (2010). Multi-modal Semantic Place Classification. In *International Journal of Robotics Research*, vol. 29, n. 2-3, pp. 298-320.

[7] Viswanathan, P., Southey, T., Little, J., and Mackworth, A. (2011). Place Classification Using Visual Object Categorization and Global Information. In *Canadian Conf. on Computer Robot Vision*, pp.1 -7.

[8] Pronobis, A., and Jensfelt, P. (2011). Hierarchical Multi-Modal Place Categorization. In *Proceedings of the 5th European Conference on Mobile Robots (ECMR'11)*.

[9] Espinace, P., Kollar, T., Soto, A., and Roy, N. (2010). Indoor scene recognition through object detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1406-1413.

[10] Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 273-280.

[11] Ruiz-Sarmiento, J.R., Galindo, C., and Gonzalez-Jimenez, J. (2014). Mobile Robot Object Recognition through the Synergy of Probabilistic Graphical Models and Semantic Knowledge. In *European Conf. of Artificial Intelligence*, CogRob workshop.

[12] Anand, A., Koppula, H.S., Joachims, T., and Saxena, A. (2013). Contextually guided semantic labeling and search for three-dimensional point clouds. In *Int. J. Robotic Res.*, vol. 32, n. 1, pp. 19-34, 2013.

[13] Koller, D., and Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. In *MIT Press*.

[14] Besag, J. (1986). On the statistical analysis of dirty pictures. In *Journal of Royal Statistical Society*, Series B (Methodological), pp. 259-302.

[15] Greig, D., Porteous, B., and Seheult, A. (1989). Exact maximum a posteriori estimation for binary images. In *Journal of the Royal Statistical Society*, Series B.

[16] Uschold, M., and Gruninger, M. (1996). Ontologies: principles, methods and applications. In *The Knowledge Engineering Review*, 11.

[17] Ruiz-Sarmiento, J.R., Galindo, C., Gonzalez-Jimenez, J. (2015). Exploiting Semantic Knowledge for Robot Object Recognition. In *Knowledge-Based Systems*.

[18] Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Proc. of the 12th European conference on Computer Vision (ECCV)*, Vol. V.

[19] Galindo, C., Fernández-Madrigal, J-A, and González, J. (2008). Multihierarchical interactive task planning: application to mobile robotics. In *IEEE Trans. on Systems, Man, and Cybernetics*, Part B: Cybernetics.

[20] Galindo, C., Fernández-Madrigal, J-A, González, J., and Saffiotti, A. (2007). Using semantic information for improving efficiency of robot task planning. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*.

[21] Ruiz-Sarmiento, J.R., Galindo, C., Gonzalez-Jimenez, J. (2015). UPGMpp: a Software Library for Contextual Object Recognition. In *3rd. Workshop on Recognition and Action for Scene Understanding*.