Searching for Objects in Human Living Environments based on Relevant Inferred and Mined Priors*

Alejandra C. Hernandez[1] , Maximilian Durner[2], Clara Gomez[1], Iris Grixa[2],
Oskars Teikmanis[2], Zoltan-Csaba Marton[2] and Ramon Barber[1]

***Abstract*— Service robots performing tasks in human environments constantly face changes due to the dynamic of the environments. Such robots need to reason about their surrounding for a better understanding of it. Besides, it is important to demonstrate capabilities that potential users would find useful, thus validating the development of such systems. One of these capabilities is to help a person to find what she or he is looking for. This mundane task of searching for an object is highly relevant in showing the non-expert user that a robot can understand the world. In this paper, we propose an efficient search strategy to find target objects that have not been seen before, based on the reasoning about in which scenes and with which objects they co-occur. Our method consists of an inference process based on a Conditional Random Field (CRF), that fuses the information about other previously detected objects, the semantic floor map, and the object-object/-room relations, to build a prediction map with the most promising locations for an unseen object. To validate our work, comparative experiments in simulated environments have been performed, demonstrating the efficiency of our proposed search strategy.**

I. INTRODUCTION

Robots operating alongside humans are permanently confronted with changing environments. Besides modeling the robot's surrounding, a major challenge is that the locations of task-relevant objects is changing within a dynamic environment. Despite the efforts to equip robots with complex perception systems, both modeling and updating the robot's world representation remain open research questions. In this work, we focus on the ability of a mobile robot to find possible locations of movable objects in human living environments. We are building on the environment modeling from our previous work [1], and using the benchmark dataset presented there with a new object-centric focus. Given a query to search for a target object, several simple strategies can be applied, e.g. random or room-by-room exploration. Such brute-force strategies however do not consider multiple information channels that we as humans would use, which makes them inefficient and highly influenced by the robot's starting position, as well as the dimensions and complexity of the environment.

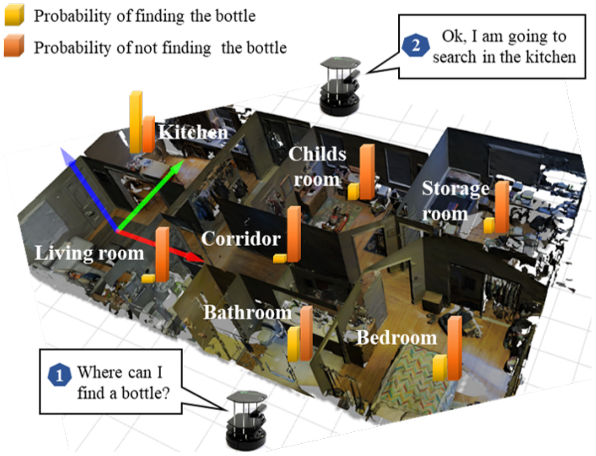
🟨 Probability of finding the bottle
🟧 Probability of not finding the bottle
2 Ok, I am going to search in the kitchen
Kitchen
Childs room
Storage room
Living room
Corridor
Bathroom
Bedroom
1 Where can I find a bottle?

Fig. 1. Given the semantic annotated floor plan and the acquired knowledge about other objects within the apartment, the system generates a map containing the probability to find an unseen target object in different locations. Then it prioritizes the high-probability locations during the search.

Thus, we propose a multi-cue search method for objects in human living environments, that can appropriately combine different types of prior knowledge, such as semantic information and object-object/-room relations. Through this a probabilistic understanding about the location of unseen target objects is obtained, resulting in a more efficient search strategy, as illustrated in Fig. 1. The core of our method is a CRF [2] and its ability to encode known relations between different observations. We fuse the information about the locations where related objects were detected previously, a semantic floor map, as well as co-occurrence statistics of objects and room types, and determine the probability of finding the object in a specific location. Thanks to the rapid advancements in the field, very large images (in our case floor-maps) can now be processed with complex CRF models. While in [1] we used structured prediction [3], here we opted for a fully connected model, for which efficient inference is available as well in our use-case [4]. In this paper we focus on the process of choosing the most promising room where an unseen target object can be located. The in-room exploration is beyond the scope of this work.

The main contributions of this work are: first, a search method based on a probabilistic graphical model that can efficiently determine a set of most likely locations of a queried unseen target object, by exploiting co-occurrence statistics mined from online datasets. Second, we conduct a qualitative and quantitative evaluation on the Bosch Se-

*This work has received funding from HEROITEA: Heterogeneous Intelligent Multi-Robot Team for Assistance of Elderly People (RTI2018-095599-B-C21), funded by Spanish Ministerio de Economía y Competitividad, and the RoboCity2030-DIH-CM project (S2018/NMT-4331, RoboCity2030-Madrid Robotics Digital Innovation Hub). The authors thank Oliver Porges for his support.

[1]Robotics Lab, Department of Systems Engineering and Automation, Carlos III University of Madrid, Spain. {alejhern,clgomezb, rbarber}@ing.uc3m.es

[2]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany {name.lname}@dlr.de

978-1-6654-1213-1/21/$31.00 ©2021 IEEE

mantic Interpretation Challenge dataset [1] obtaining better performance with respect to a baseline method. In addition, we integrate the search method with a topological navigation system [5] for performing the optimized search, in order to show the benefits and feasibility of implementing our approach in mobile robots.

II. RELATED WORK

The interaction with objects has become crucial in order to build a semantic representation of the space [6]. Research focused on object search strategies appeared in the literature since the 70s. Early works like [7], [8] propose to extend the search for one or several intermediate objects that commonly have a relation with the actual target object. In [9] an indirect search method considering distance and directional relations between objects is proposed. More recently, Loncomilla et al. [10] proposed a Bayesian based method for searching through secondary objects. Even though this strategy can reduce the computational costs as well as the search area, sometimes the spatial relationship between both, target and intermediate object, is not strong or does not exist. Furthermore, the difficulty of accurate object detection is still valid.

Other approaches address the problem of searching for objects through direct strategies. Veiga et al. [11] use information about objects and their relations to obtain uncertainty estimates about the object location in a domestic environment. In [12] the search strategy is based on the Markov-chain Monte Carlo method and includes the knowledge about which objects tend to be close to one another. Kunze et al. [13] incorporate ontological concepts and relations for reasoning about the locations of object classes. Experiments show increasing efficiency in the object search process by minimizing the movement cost and the number of processed images. Aydemir et al. [14] combine semantic information in a partially observable Markov decision process. The dependencies between objects and scene categories are modeled through a probabilistic chain graph model. In [15] a model based solely on temporal relationships between objects and search locations is proposed. Joho et al. [16] pose the task of looking for objects as an exploration problem and propose a reactive search technique which determines where to explore next based on local information about objects obtained by radio-frequency identification sensors. The approaches described above consider object and scene relations separately and most of them have been evaluated in small environments.

With the rise of deep learning techniques, some works apply more complex methods to find objects. In [17] and [18] the search problem is approached through a deep reinforcement learning model that learns action policies to reach the target object. In [17] the target object has been previously seen and the semantic segmentation and the depth information are computed in each robot observation. Druon et al. [19] propose a visual navigation method based on the context information of previously detected objects to calculate the similarity to the target object. The main drawback of these methods is that they are time-consuming and demanding and sometimes fail due to the agent getting stuck in the same place repeating the same actions. Confusions between objects with similar semantics can also appear.

In our work, we propose a direct search strategy that appropriately combine different types of prior information. We consider the relations between objects and rooms, as well as their influence on the task of finding new unseen objects.

III. GENERAL OVERVIEW

Our approach is based on the assumptions that objects mainly occur in specific environments (e.g. a pan in a kitchen) and that objects often co-occur with other objects. Fig. 2 shows an overview of the proposed method. As it can be seen, the core element is the fully connected CRF [2], which fuses the information about some previously detected objects, a semantic floor map, as well as object-object/room relations to calculate the probability of finding the target object within the corresponding grid-cell.

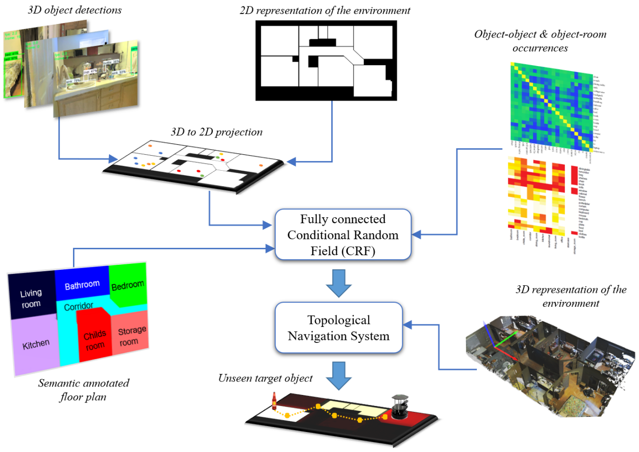
3D object detections
2D representation of the environment
Object-object & object-room occurrences
3D to 2D projection
Fully connected Conditional Random Field (CRF)
Living room
Bathroom
Bedroom
Corridor
Kitchen
Childs room
Storage room
Semantic annotated floor plan
Topological Navigation System
3D representation of the environment
Unseen target object

Fig. 2. Our proposed object search method. A 2D representation of the environment, 3D object detections and the semantic information of the environment are combined properly through a fully connected CRF to obtain the most probable room where a target object is located.

Initially, the method assumes as input information: first, a 2D representation of the environment which is generated previously by the robot as shown in [20]. Second, we feed a room-wise semantic annotated floor plan, as proposed in [1], in combination with object-object and object-room co-occurrences into the CRF. The co-occurrences have been built based on the NYU-Depth V2 dataset [21] and the COCO dataset [22]. Third, the 3D information of some previously detected objects is considered to feed our CRF. A Faster R-CNN object detector [23] is trained on a large amount of different object classes included in COCO dataset. Thereby, given the robot's location and the corresponding depth image of the detector's RGB input image, the detected objects are projected from 3D (by taking the median depth in the bounding box) into the 2D floor-map of the environment. Finally, the resulting object-annotated floor-map is further input of the CRF. Then, during the inference process the CRF outputs a heat map visualizing the most probable locations (rooms) for the unseen target object. The method prioritizes the high-probability locations during the search. Next, the information about the most probable room is incorporated into our previously developed topological navigation system [5] to generate an optimal search plan through the environment to reach the desired room.

IV. BUILDING OBJECT AND ROOM ASSOCIATIONS

Associating objects of daily use with certain categories of places facilitates the search for an object in a specific room context among other task [24], [25]. In the same way, some objects tend to be near or far from others. We include both concepts by exploiting object-room and object-object occurrences. While for the object-room relations the statistics presented in [1] based on NYU-Depth V2 dataset are applied, the object-object co-occurrences are built based on the COCO dataset [22]. The object categories that appear among our detections have been identified, and the probabilities that they occur close to other objects from this subset are computed. Working with co-occurrences generally implies the use of similarity measures to normalize the data. The associations between objects have been computed through the Jaccard Similarity Index which enables to compare the similarity, dissimilarity, and distance of members for two sets [26]. Fig. 3 shows the object-object co-occurrences of the object categories selected for this work.

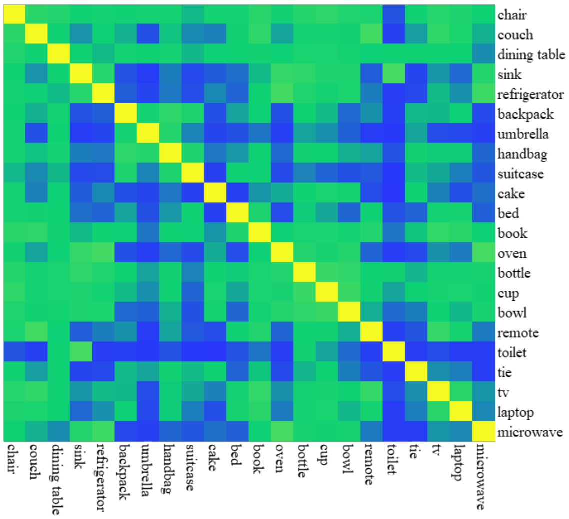

Fig. 3. Object-object co-occurrences of the object categories detected in our dataset. Blue color represents the lowest value and yellow color the highest value. All remaining values get a color based on the probability.

The Jaccard coefficient $J(A_k, A_l) \in [0, 1]$ is the ratio between the intersection and the union of the two sets A_k and A_l:

$$J(A_k, A_l) = \frac{|A_k \cap A_l|}{|A_k \cup A_l|} = \frac{|A_k \cap A_l|}{|A_k| + |A_l| - |A_k \cap A_l|}, \quad (1)$$

where A_k and A_l depict subsets of the set of training images $\mathcal{I}$ in which object k respectively l is detected. The higher the value of $J(A_k, A_l)$, the greater is the probability of the two objects k and l occurring close to each other. Through this, helpful semantic cues are obtained that are subsequently incorporated as inputs into our search method.

V. MODELING THE METHOD TO SEARCH FOR OBJECTS WITH CRF

In this section, we present a detailed explanation of how to fuse the different input data to obtain a robust estimate of the location of an unseen target object. Graphical models provide a natural way to represent the dependence of some variables with others which makes them suitable for this use case. A CRF [27] is a discriminative undirected probabilistic graphical model that considers known relationships (contexts) between observations to construct consistent predictions. Here, the model generates a pixel-wise prediction in the floor-map about the probability of the target object's location.

Consider a set of random variables $X = \{X_1, \ldots, X_N\}$ where X_i corresponds to pixel x_i of the 2D geometric floor-map $G \in \mathbb{R}^{mxn}$, obtained by [20]. The Gibbs energy function that characterizes a fully connected CRF to obtain the final probability of finding a target object in pixel x_i is:

$$E(X|G) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \quad (2)$$

where i and j range from 1 to N. The pixel-wise unary potential is depicted as $\psi_u(x_i)$, while the pairwise potential $\psi_p(x_i, x_j)$ is modeled as mixtures of Gaussian kernels.

The *unary potential* is calculated independently for each pixel in G by fusing the object detector output and the semantic room annotation. Given a target object o_t, the softmax output for a previously detected object o_d in pixel x_i is defined as $p(o_d)$. Hence the unary potential is:

$$\psi_u(x_i) = \begin{cases} p(o_d) * J(o_t, o_d) & if\ p(o_d) > 0 \\ J(o_t, \xi_s) & otherwise \end{cases} \quad (3)$$

where $J(o_t, o_d)$ is the co-occurrence probability between the objects o_t and o_d, and $J(o_t, \xi_s)$ is the co-occurrence probability of finding the target object o_t in the room ξ_s.

The *pairwise potentials* represent the relationships between all pair of pixels, that in our model have the form:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} \omega^{(m)} k^{(m)}(f_i, f_j) \quad (4)$$

where, similar to [4], $\mu(x_i, x_j)$ is a label compatibility function and $\omega^{(m)}$ a linear combination weight. The term $k^{(m)}(f_i, f_j)$ defines a Gaussian kernel where f_i and f_j are feature vectors for pixels x_i and x_j in an arbitrary feature space. We implement two kernels which are defined as:

$$\underbrace{\omega^{(1)} exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2})}_{\text{appearance kernel}} + \underbrace{\omega^{(2)} exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2})}_{\text{smoothness kernel}} \quad (5)$$

These kernels have been defined in terms of the color vectors I_i and I_j and the position vectors p_i and p_j. The parameters θ_α, θ_β and θ_γ control the weighting within a kernel and have to be set experimentally. The appearance kernel is a color-dependent term where the features are composed of the pixel location and the RGB pixel values. The second kernel, smoothness kernel, removes local isolated outliers within the map. The scalars $\omega^{(1)}$ and $\omega^{(2)}$ define the weights which must also be adjusted. In this way, the inference process is applied in the whole environment to merge all the potentials and thus obtain a 2D floor-map with the final probabilities in each pixel for the object sought.

VI. EXPERIMENTAL EVALUATION

A. Experimental Setup

To evaluate the performance of our proposed search model all the experiments have been conducted using the Bosch Semantic Interpretation Challenge dataset [1]. This dataset consists of 10 apartments from real homes that contains for each a 3D mesh, rendered mesh views from various viewpoints and a 2D projected ground truth floor plan with annotated room types: bedroom, bathroom, living room, dining room, storage room, kitchen, office, laundromat, child room, and corridors. As target objects we selected seven objects, which normally occur in a typical house: chair, bed, bottle, cup, bowl, tv and book. Since this dataset does not contain object annotations and to our knowledge no other datasets for the task of object search exists, the Ground Truth (GT) object locations are generated by applying a Faster R-CNN object detector and map onto the 2D floor plan.

Besides, we can not place new objects into the apartments of the dataset and scan it again, we simulate this process by removing an object from the previous detections, and then, ask to start searching for it. With this, our search method receives a modified map with some missing objects guaranteeing that the target object was not seen before. This way we have an unbiased GT location for the object, and all the information is equivalent to it being placed there after the initial scan. In our experiments we use the library presented in [4] to implement our CRF based model. Also, our method has been integrated with a topological navigation system [5] to obtain the best path to reach the most probable room.

B. Parameters Adjustment

In our experiments the parameters: $\omega^{(1)}$, $\omega^{(2)}$, θ_α, θ_β and θ_γ of our CRF (5) are set experimentally. Due to a good train-test split is not possible, as the dataset is very small, to adjust these parameters we have performed a grid-search study. We define $\omega^{(1)}$ as a parameter between 0 and 1 and $\omega^{(2)} = 1 - \omega^{(1)}$ and the parameter $\theta = \theta_\alpha = \theta_\beta = \theta_\gamma$. For each value of $\omega^{(1)}$ and θ, we evaluate our CRF based method in all the apartments and target objects. Fig. 4 shows a heat map of testing our CRF based method with different parameter values. The darker the color the better are the results predicting the most probable room. For a wide range of values for $\omega^{(1)}$ and θ the method obtains good results.

This study shows that our CRF based method is robust to the choice of the relative weight of the pairwise potentials, with a wide range of values performing well on all apartments. Where $\omega^{(1)} \in [0.31, 0.41]$ would describe the range of values that works best for our method. The standard deviations have a very limited effect, because we are ranking complete rooms, thus the variance is not too sensitive. This suggests that the parameters can be intuitively selected with good generalization to unseen situations.

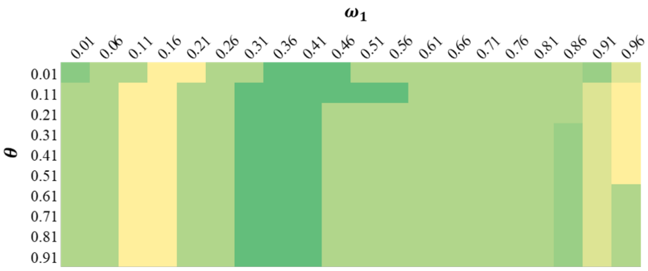

Fig. 4. Heat map of the parameters used by our CRF based method. The heat map values correspond to the mean order of the GT room.

C. Quantitative CRF results

Fig. 5 shows an example of the proposed object search method. Three different objects (bottle, cup and bowl), which were not seen in advance, are searched in three apartments (1, 2 and 5) of the Bosch dataset. First, the 3D detections of some previously detected objects are projected in the 2D floor plan (a). Based on the scene labeled maps (b), an inpainted process is applied to each room in order to eliminate the walls (c). Since walls do not contain valuable information, their unary potential is 0 which negatively affects the final outcome. Due to the fully connected characteristic of the CRF the 0 potential would radiate into the room which would result in higher probabilities in the center of the room than closer to the walls. Then, the unary potentials are calculated (d), fusing the object-object/-room probabilities and the information about the previously detected objects. Through this, an initial probability map is obtained for the target object in each room. After that, the pairwise potentials are computed, resulting in a final heat map for the target object location (e). The GT room used for comparisons can be seen in (f). Lighter colors on the heat map represent more likely locations for the target object. The darker areas represent lower probabilities to find the target object there.

To evaluate the influence of the potentials in our method, Fig. 6 shows the results of the ranking of the GT room for each target object in each apartment, by applying only unary potentials and incorporating the pairwise potentials into the search process. The y-axis represents the frequency with which the GT room is identified as the most likely room to find the target object. The x-axis corresponds to the ranking of the GT room. In addition, Table I shows the results of evaluating our method using only unary potentials and when incorporating our CRF based method in the 10 apartments of the Bosch Semantic dataset.

TABLE I

PERCENTAGES OF HITS OF THE GROUND TRUTH ROOM FOR THE 10 APARTMENTS USED DURING THE EXPERIMENTS.

Apt.	1	2	3	4	5	6	7	8	9	10	Avg.
Unary	0.33	0.17	0.17	0.0	0.17	0.20	0.50	0.40	0.43	0.14	0.25
CRF	0.67	0.67	0.50	0.67	0.67	0.80	0.83	0.60	0.43	0.57	**0.65**

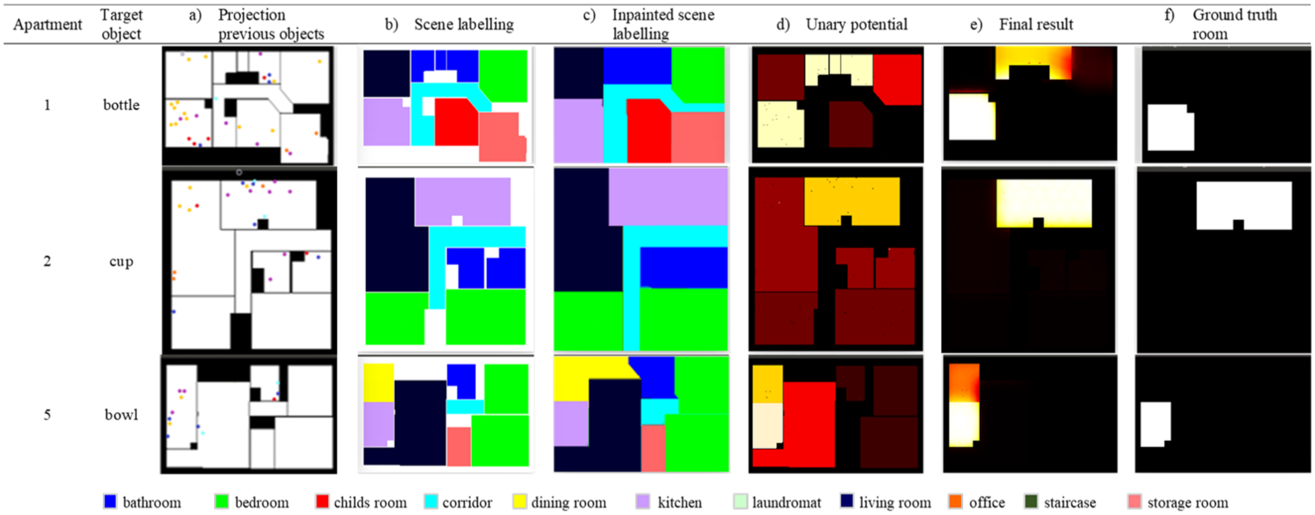

Fig. 5. The proposed search method applied to three apartments searching for three unseen objects. a) shows the projection of some previously detected objects, b) corresponds with the scene labeling of the apartment, c) is the inpainted scene, d) the unary potential heat map, e) the final CRF heat map and (f) the ground truth created for comparisons.

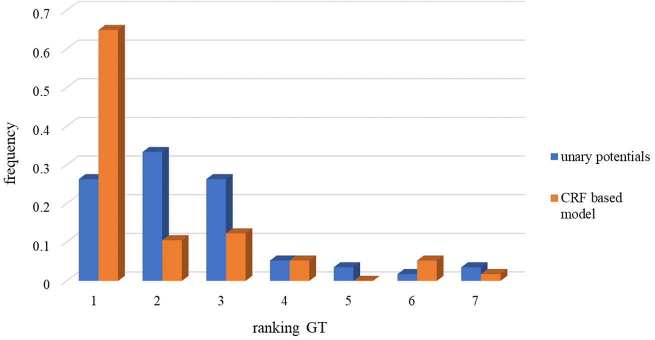

Fig. 6. Histograms of the ranking of the GT room for each target object in each apartment. The GT room occupies the first position many more times when using our CRF approach compared to applying only unary potentials.

As it can be seen, our CRF approach generates better results, with the highest percentage of times that the GT room is detected in the first place. 65% of the times the most likely room corresponds to the GT room when our CRF based search method is applied. In addition, 75% of the times the GT room is classified as the first or second option to look for the object. On the other hand, 25% of the times, the GT room is classified as the first option to find the objects when the model only considers the unary potential information. Furthermore, the average of the final probability in the most likely room is 89.23% considering only unary potentials in (5), which increases to 97.16% when pairwise potentials are included in our CRF based method.

D. Integration with a Topological Navigation System

In order to show the applicability and to evaluate the efficiency of our proposed method, we integrate it with the topological navigation system presented in [5]. The test has been carried out in the apartment 1 of the Bosch dataset using the Gazebo simulator. We have selected a differential drive robot ($20cm$x$30cm$), equipped with a Hokuyo URG-04LX-UG01 laser to map the environment. Three target objects were selected: a bottle, a bed and a chair. Initially, the robot is placed on the starting point to begin executing the search task. The most likely room information generated by our search method feeds the navigation system that calculates the optimal path to reach it. The path planning process is based on the Dijkstra algorithm. The decision about where to go is based on the maximum utility, that is, the robot chooses the room with the highest probability given by our CRF based approach. In the case of ambiguities, meaning several rooms with the same probability, the model considers the minimum path. Table II compares the search task by applying only unary potentials and incorporating pairwise potentials to find the target object. The distance traveled, the time spent on the search task and the number of steps through which the robot has to pass until it reaches the goal are calculated.

TABLE II
RESULTS OF THE SEARCH FOR OBJECTS THROUGH THE TOPOLOGICAL NAVIGATION SYSTEM.

	Unary			CRF		
Target object	distance (m)	time (s)	steps	distance (m)	time (s)	steps
bottle	14.81	12.21	2	3.39	3.17	1
bed	27.55	15.64	3	12.20	8.06	1
chair	23.53	13.63	3	10.01	8.20	1

The results show that our CRF based method requires less distance and invests less time going to the room where the target object is located compared to using only unary potential information. Regarding the number of steps, in most cases using the estimates of the CRF method, the room with the highest probability corresponds to the correct room. Hence, the robot tends to make only one step to reach the goal. In other cases, when the correct room is not the most likely, the planner directs to the first room and then re-plans to the next most likely room and so on until it reaches the GT room. The results from the experiments carried out in this work demonstrate the effectiveness, robustness and validity of our approach to the task of searching for an object when information of the environment is appropriately considered.

E. Comparison with a Baseline Method

To obtain fair comparison results, the methods have to be evaluated in the same conditions, target objects and object poses. To the best of the authors' knowledge, there is no dataset available for comparison of object search methods in human living environments. To overcome this issue, we have designed a baseline method to compare with our CRF model using the Bosch Semantic dataset. While in the CRF, cues from several methods and information channels are merged, one could also think of a Convolutional Neural Network (CNN) learning implicitly a mapping from visual cues in an image to the probability of finding the target object within this camera view. Based on this idea, we implement a baseline method that consists of four main steps (Fig. 7).

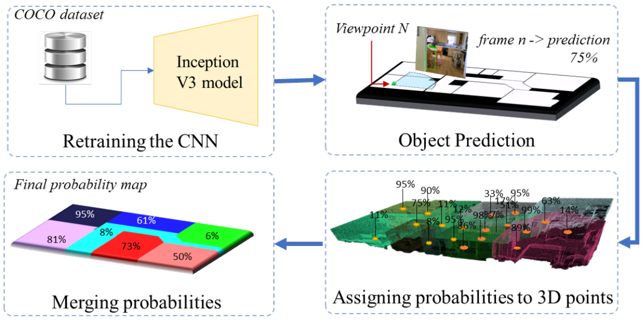

Fig. 7. Baseline method. Using transfer learning a CNN is retrained to predict the most probable location of a target object in each of the 10 apartments of the Bosh dataset.

First, the fully connected layer of a pre-trained Inception V3 model is retrained on the COCO dataset. The dataset contains images with and without the target objects, leading to a binary classification problem. This way, the CNN implicitly learns the object-object/-room relations. Second, the CNN is applied to image frames within apartments of the Bosch dataset. As a result, the probability of the target object being in each image is obtained. Third, the predicted output for each frame is associated with the 3D point cloud according to the viewpoint. At the end, each seen 3D point of the cloud has a probability of the target object. Then, the probability of finding the target object in each room is obtained by merging the probabilities of the points that belong to each room. To do that, we apply a majority voting to obtain the maximum probability found in each room. Through this, the highest probability of each room associated with the target object is obtained.

Fig. 8 shows the results of comparing the baseline method with our CRF based approach in the estimation of the most likely room where the unseen target objects can be located. This method has been executed on the same dataset and looking for the same target objects as described in section VI-A. As it can be seen, our CRF approach outperforms the baseline method, with the highest percentage of times that the GT room is detected in the first place. Table III shows the results of evaluating both approaches in the 10 apartments of the dataset. With the baseline method, only 42% of the times the most probable room corresponds to the GT room compared to 65% when the CRF based model is applied. Moreover, 75% of the times the GT room is classified as the first or second option applying our CRF based method compared to the 55% when the baseline method is applied.

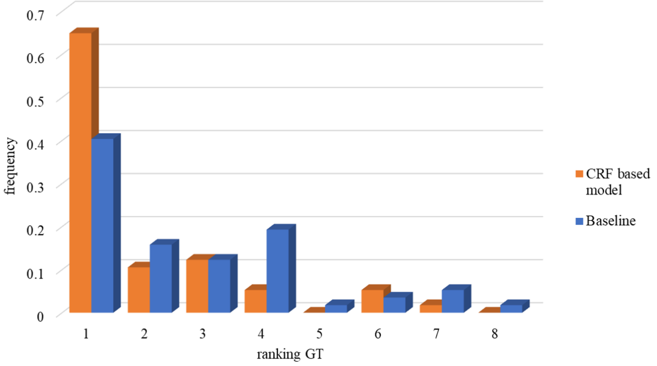

Fig. 8. Comparison of the ranking of the GT room for each target object in each apartment using our CRF approach and the baseline method.

TABLE III

PERCENTAGES OF HITS OF THE GROUND TRUTH ROOM FOR THE 10 APARTMENTS APPLYING OUR CRF METHOD AND THE BASELINE.

Apt.	1	2	3	4	5	6	7	8	9	10	Avg.
CRF	0.67	0.67	0.50	0.67	0.67	0.80	0.83	0.60	0.43	0.57	**0.65**
Baseline	0.0	0.50	0.33	0.67	0.33	0.80	0.67	0.40	0.43	0.14	0.42

In addition, the average of the final probability in the most likely room is 94.83% applying the baseline method and 97.16% building the estimates with our CRF based search method. Applying the baseline approach, in some cases the model predicts objects that are not present in the real detections. This is due to the method's characteristic training the network on objects related to context where ambiguities might occur.

VII. CONCLUSIONS

In this work we proposed an efficient search strategy to find target objects that have not been seen before in human living environments. Our CRF based method considers additional cues which influence the robots understanding of its environment, namely the object and context relations as well as semantic information. The presented experiments demonstrate the usefulness and efficiency of our method to estimate the most probable room where a target object can be located. In addition, our method has been tested in the whole Bosch Semantic dataset, that is built from 10 real apartments, which demonstrates the flexibility to apply it in different environmental conditions.

As future work, we plan to conduct tests in real-world environments and enhance the cost function of the topological navigation model to optimize the calculation of the best path. Likewise, we would like to study the influence of the walls in the calculation of our search model based on CRF and the incorporation of another type of semantic information.

References

[1] M. Brucker, M. Durner, R. Ambruş, Z. C. Márton, A. Wendt, P. Jensfelt, K. O. Arras, and R. Triebel, "Semantic labeling of indoor environments from 3d rgb maps," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1871–1878.

[2] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[3] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Efficient Structured Prediction with Latent Variables for General Graphical Models," in *Proc. ICML*, 2012.

[4] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[5] C. Gomez, A. Hernandez, J. Crespo, and R. Barber, "Uncertainty-based localization in a topological robot navigation system," in *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2017, pp. 67–72.

[6] S. Vasudevan and R. Siegwart, "Bayesian space conceptualization and place classification for semantic maps in mobile robotics," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 522–537, 2008.

[7] T. D. Garvey, "Perceptual strategies for purposive vision." 1976.

[8] L. E. Wixson and D. H. Ballard, "Using intermediate objects to improve the efficiency of visual search," *International Journal of Computer Vision*, vol. 12, no. 2-3, pp. 209–230, 1994.

[9] L. Kunze, K. K. Doreswamy, and N. Hawes, "Using qualitative spatial relations for indirect object search," in *2014 IEEE International Conf.on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 163–168.

[10] P. Loncomilla, J. Ruiz-del Solar, and M. Saavedra, "A bayesian based methodology for indirect object search," *Journal of Intelligent & Robotic Systems*, vol. 90, no. 1-2, pp. 45–63, 2018.

[11] T. S. Veiga, P. Miraldo, R. Ventura, and P. U. Lima, "Efficient object search for mobile robots in dynamic environments: Semantic map as an input for the decision maker," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2745–2750.

[12] X. Nie, L. L. Wong, and L. P. Kaelbling, "Searching for physical objects in partially known environments," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5403–5410.

[13] L. Kunze, M. Beetz, M. Saito, H. Azuma, K. Okada, and M. Inaba, "Searching objects in large-scale indoor environments: A decision-theoretic approach," in *2012 IEEE International Conference on Robotics and Automation*. Citeseer, 2012, pp. 4385–4390.

[14] A. Aydemir, A. Pronobis, M. Göbelbecker, and P. Jensfelt, "Active visual object search in unknown environments using uncertain semantics," *Transactions on Robotics*, vol. 29, no. 4, pp. 986–1002, 2013.

[15] R. Toris and S. Chernova, "Temporal persistence modeling for object search," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 3215–3222.

[16] D. Joho, M. Senk, and W. Burgard, "Learning search heuristics for finding objects in structured environments," *Robotics and Autonomous Systems*, vol. 59, no. 5, pp. 319–328, 2011.

[17] X. Ye, Z. Lin, J.-Y. Lee, J. Zhang, S. Zheng, and Y. Yang, "Gaple: Generalizable approaching policy learning for robotic object searching in indoor environment," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4003–4010, 2019.

[18] X. Ye, Z. Lin, H. Li, S. Zheng, and Y. Yang, "Active object perceiver: Recognition-guided policy learning for object searching on mobile robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6857–6863.

[19] R. Druon, Y. Yoshiyasu, A. Kanezaki, and A. Watt, "Visual object search by learning spatial context," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1279–1286, 2020.

[20] R. Ambruş, S. Claici, and A. Wendt, "Automatic room segmentation from unstructured 3-d data of indoor environments," vol. 2, no. 2, pp. 749–756, April 2017.

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[22] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[23] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[24] M. Hanheide, C. Gretton, R. Dearden, N. Hawes, J. Wyatt, A. Pronobis, A. Aydemir, M. Göbelbecker, and H. Zender, "Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, ser. IJCAI'11. AAAI Press, 2011, p. 2442–2449.

[25] J. Young, V. Basile, M. Suchi, L. Kunze, N. Hawes, M. Vincze, and B. Caputo, "Making sense of indoor spaces using semantic web mining and situated robot perception," in *The Semantic Web: ESWC 2017 Satellite Events*, E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, and O. Hartig, Eds. Cham: Springer International Publishing, 2017, pp. 299–313.

[26] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 6, 2013, pp. 380–384.

[27] C. Sutton, A. McCallum *et al.*, "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.