2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France
[Award version is not citable, please cite the contributed session version]

Semantic Linking Maps for Active Visual Object Search

Zhen Zeng* Adrian Röfer† Odest Chadwicke Jenkins*

***Abstract*— We aim for mobile robots to function in a variety of common human environments. Such robots need to be able to reason about the locations of previously unseen target objects. Landmark objects can help this reasoning by narrowing down the search space significantly. More specifically, we can exploit background knowledge about common spatial relations between landmark and target objects. For example, seeing a table and knowing that cups can often be found on tables aids the discovery of a cup. Such correlations can be expressed as distributions over possible pairing relationships of objects. In this paper, we propose an active visual object search strategy method through our introduction of the Semantic Linking Maps (*SLiM*) model. *SLiM* simultaneously maintains the belief over a target object's location as well as landmark objects' locations, while accounting for probabilistic inter-object spatial relations. Based on *SLiM*, we describe a hybrid search strategy that selects the next best view pose for searching for the target object based on the maintained belief. We demonstrate the efficiency of our *SLiM*-based search strategy through comparative experiments in simulated environments. We further demonstrate the real-world applicability of *SLiM*-based search in scenarios with a Fetch mobile manipulation robot.**

I. Introduction

Being able to efficiently search for objects in an environment is crucial for service robots to autonomously perform tasks [9], [27], [7]. When asked where a target object can be found, humans are able to give hypothetical locations expressed by spatial relations with respect to other objects. For example, a *cup* can be found "on a table" or "near a sink". *Table* and *sink* are considered landmark objects that are informative for searching for the target object *cup*. Robots should be able to reason similarly about objects locations, as shown in Figure 1.

Previous works [10], [13], [26] assume landmark objects are static, in that they mostly remain where they were last observed. This assumption can be invalid for dynamic landmark objects that change their location over time, such as chairs, food carts and toolboxes. Temporal assumptions can mislead the search process if the prior on the landmarks' locations is too strong. Further, there also exists uncertainty in the spatial relations between landmark objects and the target object, and between landmark objects themselves. For example, a *cup* can be "in" or "next to" a *sink*.

Considering the problem of dynamic landmarks, we propose the Semantic Linking Maps (*SLiM*) model to account for uncertainty in the locations of landmark objects during object search. Building on Lorbach et al. [18], we model inter-object spatial relations probabilistically via a factor graph. The marginal belief on inter-object spatial relations inferred from the factor graph is used in *SLiM* to account for probabilistic spatial relations between objects.

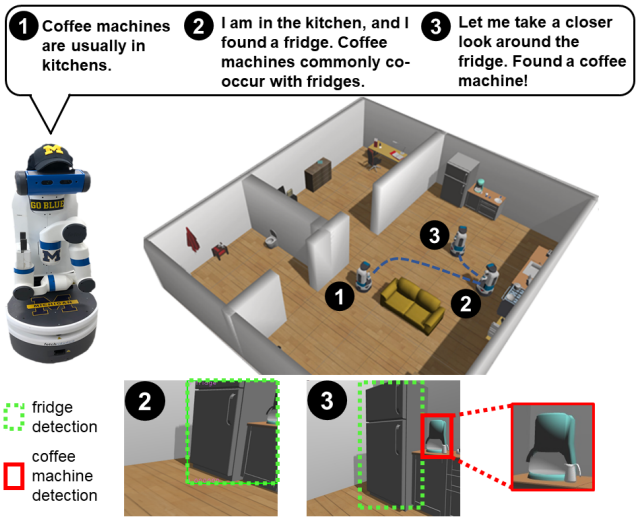

Fig. 1: Robot tasked to find a coffee machine.

Using the maintained belief over target and landmark objects' locations from *SLiM*, we propose a hybrid strategy for active object search. We select the next best view pose, which guides the robot to explore promising regions that may contain the target and/or landmark objects. Previous works [30], [6], [25], [2] have shown the benefit of purposefully looking for landmark objects (*Indirect Search*) before directly looking for the target object (*Direct Search*). The proposed hybrid search strategy draws insights from both indirect and direct search. We demonstrate the efficiency of the proposed hybrid search strategy in our experiments.

In this paper, we describe the Semantic Linking Maps model as a Conditional Random Field (CRF). Our description of *SLiM* as a CRF allows us to simultaneously maintain the belief over target and landmark object locations with probabilistic modeling over inter-object spatial relations. We also describe a hybrid search strategy based on *SLiM* that draws upon ideas from both indirect and direct search representations. This *SLiM*-based search makes use of the maintained belief over objects' locations by selecting the next best view pose based on the current belief. In our experiments, we show that the proposed object search approach is more robust to noisy priors on landmark locations by simultaneously maintaining belief over the locations of target and landmark objects.

*Z. Zeng, O.C. Jenkins are with the Department of Electrical Engineering and Computer Science, Robotics Institute, University of Michigan, USA
†A. Röfer is with the Department of Computer Science, University of Bremen, Germany

II. Related Work

Existing works have studied object search with different assumptions on prior knowledge of the environment. Some assume priors on landmark objects' locations in the environment, and utilize the spatial relations between the target object and landmark objects to prioritize regions to search. Kollar et al. [10] utilize object-object co-occurrences extracted from image tags on Flickr.com to infer target object locations. Kunze et al. [13] expanded the generic notion of co-occurrences to more restrictive spatial relations (e.g. "in front of", "left of"), which provide more confined regions to search, thus improving the search efficiency. Toris et al. [26] proposed to learn a temporal model on inter-object spatial relations to facilitate search. These methods assume the landmark objects to be static, however, we believe accounting for the uncertainty in landmark objects' locations is important for object search.

Existing works have also explored known priors on spatial relations between landmark and target objects. Given exact spatial relations between landmark and target objects, Sjöö et al. [25] used an *indirect object search* strategy [30], [6], where the robot first searches for landmark objects, and then searches for a target object in regions satisfying given spatial relations. On the other hand, given a probabilistic distribution over the spatial relations between objects, Aydemir et al. [2] formulate the object search problem as a Markov Decision Process. In our work, we learn the probabilistic inter-object spatial relations by building on ideas of Lorbach et al. [18], where inter-object relations are being probabilistically modeled via a factor graph.

There are also works that do not assume prior knowledge of the environment. Researchers have explored object search with visual attention mechanisms [23], [24], [19], such as saliency detection. Similar to [10], [13], other research [17], [4], [8] utilizes object-object co-occurrences to guide the search for a target object. Positive and negative detections of landmark objects will result in an updated belief over the target object. We expand object-object co-occurrences to finer-grained spatial relations between objects, i.e., "in", "on", "proximity", "disjoint", which specify more confined regions for object search.

Other literature [29], [12], [28] has also explored object-place relations to facilitate object search. Wang et al. [29] build a belief road map based on object-place co-occurrences for efficient path planning during object search. Kunze et al. [12] bootstraps commonsense knowledge on object-place co-occurrences from the Open Mind Indoor Common Sense (OMICS) dataset. Samadi learned similar knowledge by actively querying the World Wide Web (WWW). Our work also takes object-place co-occurrences into account. Aydemir et al. [1] made use of place-place co-occurrences to infer the type of the room next door, as the robot explores an environment during search. Manipulation-based object search, as in [32], [31], [14], is not within the scope of this paper.

III. Problem Statement

Let $O = \{o^i | i = 1, \cdots, N\}$ be the set of objects of interest, including landmark objects and the target object for search. Given observations $z_{0:T}$ and robot poses $x_{0:T}$, we aim to maintain the belief over object locations $P(O_T | x_{0:T}, z_{0:T})$, while accounting for the probabilistic spatial relations R_{ij} between objects $o^i, o^j \in O$. For this work, we consider the set of spatial relations to be $R_{ij} \in \{In, On, Contain, Support, Proximity, Disjoint\}$. For example, the relation $R_{ij} = In$ indicates that object o_i is inside object o_j. The probabilistic spatial relations between object o^i, o^j is represented by the belief over R_{ij}, denoted as $\mathcal{B}(R_{ij})$.

Based on the maintained belief $P(O_T | x_{0:T}, z_{0:T})$, the robot searches for the target object by selecting the next best view pose ranked by an utility function $U : \tau \mapsto \mathbb{R}$. τ specifies the 6 DOF of camera view pose. The utility function U trades off between navigation cost and the probability of search success. Upon a user request to find a target object, the robot iterates between the belief update of objects' locations and view pose selection, until the target object is found or the maximum search time is reached.

IV. Semantic Linking Maps

For Semantic Linking Maps (*SLiM*), we consider inter-object spatial relations, while maintaining the belief over target and landmark objects' locations. Building on our previous work [33], we probabilistically formalize the object location estimation problem via a Conditional Random Field (CRF). The model is now extended to account for probabilistic inter-object spatial relations, as shown in Figure 2.

The posterior probability of object locations O history is

$$p(O_{0:T} | x_{0:T}, z_{0:T}) = \frac{1}{Z} \prod_{t=0}^{T} \prod_{i=1}^{N} \phi_p(o_t^i, o_{t-1}^i) \phi_m(o_t^i, x_t, z_t) \prod_{i,j} \phi_{c,\mathcal{B}(R_{ij})}(o_t^i, o_t^j) \quad (1)$$

where Z is a normalization constant. Robot pose x_t and observation z_t are known. We assume that the robot stays localized given a metric map of the environment.

$\phi_p(o_t^i, o_{t-1}^i)$ is the *prediction potential* that models the movement of an object over time. We assume objects to remain static or move with temporal coherence (varies across object classes) during the search, i.e.

$$\phi_p(o_t^i, o_{t-1}^i) = e^{-(o_t^i - o_{t-1}^i)^T \Sigma^{-1} (o_t^i - o_{t-1}^i)}$$

$\phi_m(o_t^i, x_t, z_t)$ is the *measurement potential* that accounts for the observation model, and $z_t = \{z_t^i | i = 1, \cdots, N\}$ are (potentially noisy) detections for each object o^i at time t. Because z_t^i and o^j are independent if $j \neq i$, we simplify $\phi_m(o_t^i, x_t, z_t)$ to $\phi_m(o_t^i, x_t, z_t^i)$ s.t.,

$$\phi_m(o_t^i, x_t, z_t^i) = \begin{cases} P_{FN}, & \text{if } o_t^i \in E_t^i \text{ but } z_t^i = \emptyset \\ P_{TN}, & \text{if } o_t^i \notin E_t^i \text{ and } z_t^i = \emptyset \\ P_{TP}, & \text{if } \pi(o_t^i) \in z_t^i \\ P_{FP}, & \text{otherwise} \end{cases} \quad (2)$$

where each P stands for the probability of false negative, true negative, true positive, and false positive detection. E_t^i is the

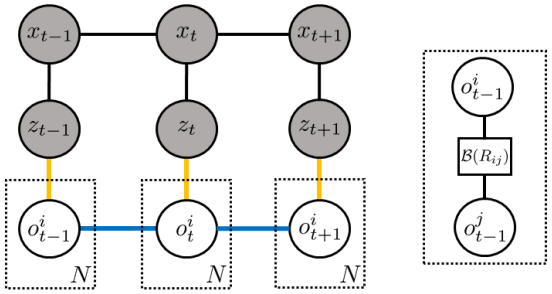

Fig. 2: CRF-based *SLiM* model: (a) Known: $\{x^t\}$ robot poses, $\{z^t\}$ sensor observations; Unknown: $O_t = \{o_t^1, o_t^2, \cdots, o_t^N\}$. (b) Plate notation: at time t, the spatial relations between each object pair o^i, o^j is parameterized by the belief over their spatial relations $\mathcal{B}(R_{ij})$.

effective observation region for o^i given robot pose at time t. Note, E_t^i is larger for larger objects, which can be reliably detected from longer distance compared to small objects. π is the camera projection matrix, and $\pi(o_t^i) \in z_t^i$ denotes that the projected object lies in the detected bounding box in z_t^i.

We model the spatial relations between objects with *context potential* $\phi_{c,\mathcal{B}(R_{ij})}$. Here, we extend ϕ_c from our previous work by paramtererizing it with the belief $\mathcal{B}(R_{ij})$ over the inter-object spatial relation between o^i, o^j,

$$\phi_{c,\mathcal{B}(R_{ij})} = \sum_r \mathcal{B}(R_{ij} = r)\phi_{c,r}(o_t^i, o_t^j, R_{ij} = r) \tag{3}$$

where r can take any value in the set of possible relations $\{$*In, On, Contain, Support, Proximity, Disjoint*$\}$.

For $r \in \{$*In, On, Contain, Support*$\}$, $\phi_{c,r}(o_t^i, o_t^j, R_{ij} = r)$ is equal to 1 if objects o_t^i, o_t^j satisfy the spatial relation given the width, length and height of the object, otherwise 0. For $r =$ *Proximity*, $\phi_{c,r}(o_t^i, o_t^j, R_{ij} =$ *Proximity*$)$ corresponds to a Gaussian distribution that models $o_t^j \sim \mathcal{N}(o_t^i, \Sigma^{ij})$ and Σ^{ij} is determined by the size of objects o^i, o^j. The larger the size of o^i, o^j, the larger the variance in Σ^{ij}. For $r =$ *Disjoint*, $\phi_{c,r}(o_t^i, o_t^j, R_{ij} = Disjoint) = 1 - \sum_{r \neq Disjoint} \phi_{c,r}(o_t^i, o_t^j, R_{ij} = r)$.

A. Inference

We propose a particle filtering inference method for maintaining the belief over object locations, as shown in Algorithm 1. Examples of the belief update over time are available in Figure 3. Instead of estimating the posterior of the complete history of object locations $p(O_{0:T}|x_{0:T}, z_{0:T})$, we recursively estimate the posterior probability of each object $o_t^i \in O_t$, similarly to [33], [15].

To deal with particle decay, we reinvigorate the particles of each o^i by sampling in known room areas, as well as around other objects o^j based on $\mathcal{B}(R_{ij})$. In step 5, $j \in \Gamma(i)$ only if $1 - \mathcal{B}(R_{ij} = Disjoint) > 0.2$. Across our experiments, we use 100 particles for each object. The inference algorithm does not assume single object instance for each object class. The inference algorithm has a complexity of $\mathcal{O}(nKM^2)$, where K is the average cardinality of $\Gamma(i)$. Further works can be done to decrease the complexity down to $\mathcal{O}(nKMC)$ by sampling

Algorithm 1: Inference of objects locations in *SLiM*.

Input: Observation z_t, Robot pose x_t,
Particle set for each object:
$o_{t-1}^i = \{\langle o_{t-1}^{i(k)}, \alpha_{t-1}^{i(k)} \rangle | k = 1, \cdots, M\}, i \in 1:N$

1 Resample M particles $o_{t-1}^{i(k)}$ from o_{t-1}^i with probability proportional to importance weights $\alpha_{t-1}^{i(k)}$;
2 **for** $i = 1, \cdots, n$ **do**
3 **for** $k = 1, \cdots, M$ **do**
4 Sample $o_t^{i(k)} \sim \phi_p(o_t^i, o_{t-1}^{i(k)})$;
5 Assign weight
$\alpha_t^{i(k)} \propto \phi_m(o_t^{i(k)}, x_t, z_t) \prod_{j \in \Gamma(i)} \phi_{c,\mathcal{B}(R_{ij})}(o_t^{i(k)}, o_{t-1}^j)$;
6 where $\phi_{c,\mathcal{B}(R_{ij})}(o_t^{i(k)}, o_{t-1}^j) =$
$\sum_r \sum_{l=1}^{M} \mathcal{B}(R_{ij} = r)\alpha_{t-1}^{j(l)} \phi_{c,r}(o_t^i, o_t^j, R_{ij} = r)$
7 **end**
8 **end**

C representative and divergent particles from the original M particles ($C < M$).

B. Probabilistic Inter-Object Spatial Relations

To get the belief over inter-object spatial relations $\mathcal{B}(R_{ij})$ for each object pair $o^i, o^j \in O$, we use a factor graph by building on preceding work by Lorbach et al [18]. We generalize [18] by relaxing the assumption on known spatial relations between landmark objects.

The factor graph $G: \{\mathbb{V}, \mathbb{F}, \mathbb{E}\}$ consists of variable vertices $\mathbb{V} = \{R_{ij} | \forall_{i \neq j}\ o^i, o^j \in O\}$, factor vertices $\mathbb{F} = \{F_{CS}, F_{LC}\}$ and edges $\mathbb{E}$ which connect factor vertices with variable vertices. Specifically, $F_{CS}: R_{ij} \mapsto \mathbb{R}$ is a unary factor that considers *commonsense knowledge* on spatial relation between objects,

$$F_{CS}(R_{ij}) = \text{Frequency}(R_{ij})$$

Similar to [18], we extract commonsense knowledge on R_{ij} from online image search engine (e.g. Flickr) by counting the frequency of certain spatial relation between objects o^i, o^j. For example, the frequency of $R_{cup,table} =$ *On* is computed as the number of search results of a query "cup on the table" divided by the number of search results of a query "on the table". These extracted frequencies can be noisy. For example, the frequency of "laptop on kitchen" is larger than 0, but it is not a valid expression because it refers to a laptop being on top of the room geometry of a kitchen. We manually encode the $F_{CS}(R_{ij})$ for invalid expressions to 0.

$F_{LC}: (R_{ij}, R_{ik}, R_{jk}) \mapsto \{0, 1\}$ is a triplet factor that considers *logical consistency* between a triplet of objects o^i, o^j, o^k,

$$F_{LC}(R_{ij}, R_{ik}, R_{jk}) = \begin{cases} 1, & \text{if consistent.} \\ 0, & \text{otherwise.} \end{cases}$$

For example, if o^i is in o^j, and o^j is in o^k, then o^i should be in o^k to satisfy logical consistency, i.e., $F_{LC}(R_{ij} = In, R_{ik} = In, R_{jk} = In) = 1$. Previous work [18] assumes the spatial relations between landmark objects to be known,

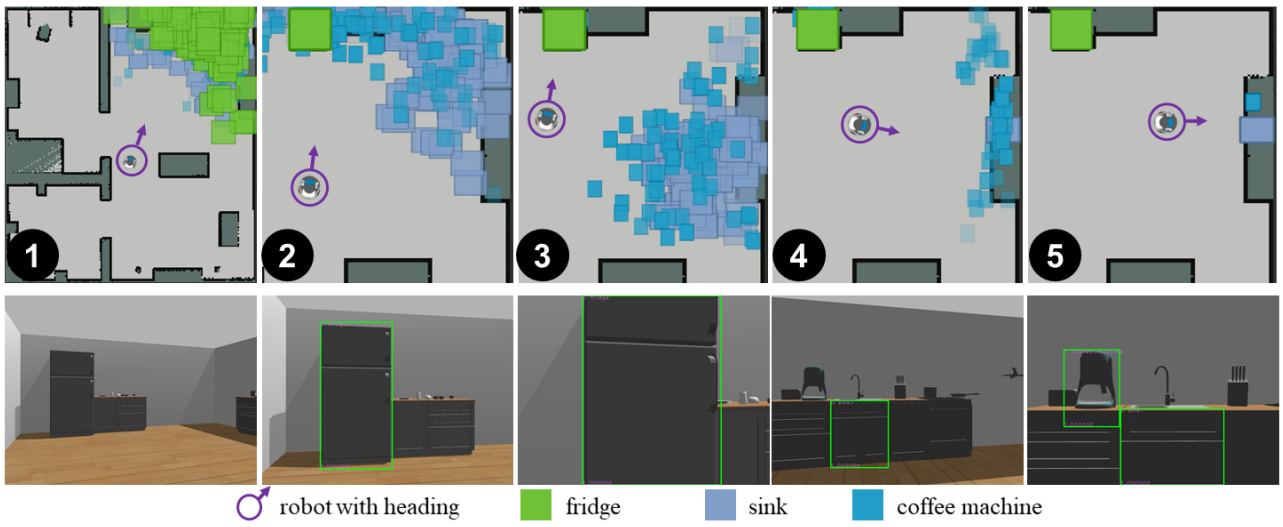

Fig. 3: Examples of belief updates in *SLiM*. given observations. *Upper*: Evolution of particles of *fridge, sink, coffee machine* over time. *Lower*: RGB observation (with object detection) over time. (Best viewed in color).

and only relations $R_{target,j}$ connecting target object o_{target} and landmark object o_j to be unknown. Their pairwise factor enforcing *logical consistency* is a binary function $F_{LC} : (R_{target,j}, R_{target,k}) \mapsto \{0,1\}$. In contrast, our formulation employs a trinary factor F_{LC} considering all possible combinations of (R_{ij}, R_{ik}, R_{jk}) and evaluating their logical consistency.

By applying Belief Propagation [11] on the factor graph formulated as above, we can get the marginal belief over inter-object relations $\mathcal{B}(R_{ij})$ between all object pairs. We use the libDAI [20] library for inference. An example of the probabilistic inter-object spatial relations inferred from the factor graph is as shown in Figure 5, and it is used in our experiments.

V. Search Strategy

Based on the belief over the object locations, we actively search for the target object, by generating promising view poses and select the best one ranked by a utility function. Given the particle set $\langle o_t^{(k)}, \alpha_t^{(k)} \rangle$ of the target object o as being maintained in IV, we fit Gaussian Mixture Models (GMMs) through Expectation Maximization to the particles by auto selecting the number of clusters [5],

$$\langle o_t^{(k)}, \alpha_t^{(k)} \rangle \sim \langle \mathcal{N}(x_n, \Sigma_n), \omega_n \rangle \tag{4}$$

A. View Pose Generation

For each Gaussian component $\mathcal{N}(\mathbf{x}_n, \Sigma_n)$, we generate a set of camera view pose candidates $\{\tau_n^i = (\mathbf{c}_n^i, \psi_n^i)\}$, where $\mathbf{c}_n$ and ψ_n denote the translation and the rotation of the camera respectively.

Initially, we sample the location of the camera $\mathbf{c}_n$ evenly from a circle with a fixed radius around the center $\mathbf{x}_n$ of the Gaussian component, and assign a default value to rotation ψ_n. Note, that these initially sampled view poses can put the robot in collision with the environment, and the camera is not necessarily looking at $\mathbf{x}_n$. Thus, we formulate a view pose optimization problem under constraints as below,

$$\underset{\tau_n}{\operatorname{argmin}}\ 1 - \mathbf{v}_n \cdot \frac{\mathbf{x}_n - \mathbf{c}_n}{\|\mathbf{x}_n - \mathbf{c}_n\|} \ \text{ s.t } \mathbf{x}_n \in E_{\tau_n}, \quad c(\tau_n) > 0 \tag{5}$$

where $\mathbf{v}_n$ is the view direction given τ_n, E_{τ_n} denotes the effective observation region of the target object at camera pose τ_n, and $c : \tau \mapsto \mathbb{R}$ is a function that computes a signed distance of a configuration τ to the collision geometry of the environment.

B. View Pose Selection

We propose two different utility functions to rank the view pose candidates:

*1) **Direct Search** utility:* $\mathbf{U}_{\text{DS}}$ encourages the robot to explore promising areas that could contain the target object while accounting for navigation cost,

$$\mathbf{U}_{\text{DS}}(\tau_k) = \omega_n + \alpha \frac{1}{\arctan(\sigma d_{nav})} \tag{6}$$

where ω_n is the weight of the Gaussian component (as in (4)) that τ_k is generated from, and d_{nav} is the navigation distance from the current robot location to view pose τ_k. Parameter α trades off between the probability of finding the target object and the navigation cost. Parameter σ determines how quickly the $\arctan(\sigma d_{nav})$ plateaus.

With $\mathbf{U}_{\text{DS}}$, the object search id **direct** because we are directly considering promising areas represented by the GMMs for the target object.

*2) **Hybrid Search** utility:* $\mathbf{U}_{\text{HS}}$ encourages the robot to explore promising areas that could contain the target object and/or any landmark object, while accounting for navigation cost

$$\begin{aligned}\mathbf{U}_{\text{HS}}(\tau_k) = \omega_n + \alpha \frac{1}{\arctan(\sigma d_{nav})} \\ + \beta \max_{j,n} \text{CoOccur}(o, o^j) \omega_n^j \mathbf{I}_n^j\end{aligned} \tag{7}$$

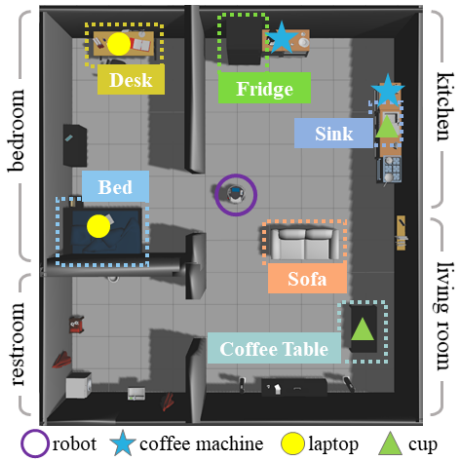

Fig. 4: Simulation experiments setup in Gazebo: an apartment-like environment with four rooms. There are 6 landmark objects and 3 target objects: *coffee machine, laptop, cup*. Each target object has two equally possible locations.

where the additional term compared to $\mathbf{U}_{\mathrm{DS}}$ acts to encourage the robot to also explore areas that could contain landmark object o^j which co-occurs with the target object o with probability $\mathrm{CoOccur}(o, o^j)$. Specifically, $\mathrm{CoOccur}(o, o^j) = (1 - \mathcal{B}(R_{target,j} = Disjoint))$, and ω_n^j is the weight of the n-th Gaussian component of GMMs fitted to the belief over the location of the landmark object o^j. And $\mathbf{I}_n^j$ is 1 if the n-th Gaussian of object o^j is within the effective observation region at camera pose τ_k, otherwise 0.

$\mathbf{U}_{\mathrm{HS}}$ is inspired by the *indirect object search* strategy as studied in [6], [30]. Previous studies demonstrated that purposefully looking for an intermediate landmark object helps quickly narrow down the search region for the target object if the landmark object often co-occurs with the target object, thus improving the search efficiency.

With $\mathbf{U}_{\mathrm{HS}}$, the object search can be considered **hybrid** because we are considering promising areas represented by GMMs for both the target object (as in direct search) and landmark objects that co-occur with the target object (as in indirect search).

In our experiments, we use a A* based planner to compute d_{nav}. We empirically set $\alpha = 0.1$, $\beta = 0.4$, and $\sigma = 0.5$ such that $\arctan(\sigma d_{nav})$ plateaus as d_{nav} goes beyond $3m$.

VI. Experiments

We perform object search tasks in both simulation and real-world environments with a Fetch robot. In the simulation experiments, we quantitatively benchmark various methods, including methods that resemble previous works and our proposed method. In the real-world experiments, we demonstrate qualitatively that the proposed method scales to real-world applications. In both simulation and real-world experiments, the robot accelerates to at most 1m/s and turns at most at 1.7rad/s.

1) Simulation Experiments: The simulation experiments are performed in an apartment-like environment (10mx11m) setup in the Gazebo simulator, as shown in Figure 4. The

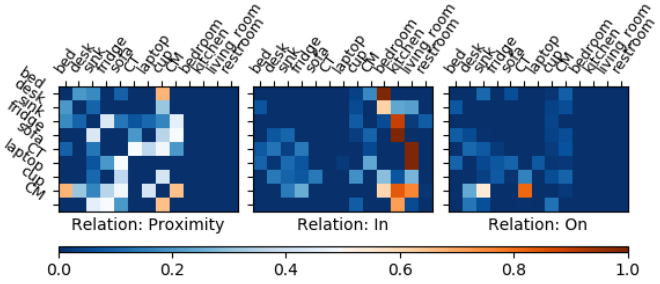

Fig. 5: Marginal belief on inter-object spatial relations, as well as object-room relations, inferred from the factor graph as explained in Sec. IV-B. *CM*: coffee machine, *CT*: coffee table

room types and considered landmark objects are annotated in Figure 4, along with the placements of target objects. The marginal belief R_{ij} inferred from the factor graph as explained in IV-B is depicted in Figure 5.

We set up an object detector in simulation that returns a detection of an object, if the object is in view, not fully occluded, and within the effective observation range. For large objects (e.g. sofa, bed, fridge), mid-sized objects (e.g. desk, table, sink), and small objects (e.g. cup, laptop, coffee machine), we assume an effective observation range of 5m, 4m, 2.5m respectively.

We benchmark following methods:

- **UDS**: Uninformed direct search (Eq.6). The robot does not account for the spatial relations between the target and landmark objects (omitting Eq. 3 in *SLiM*). This baseline represents a naive approach for object search.
- **IDS-Known-Static**: Informed direct search (Eq.6) with a known prior on landmark object locations. The robot assumes that landmark objects are static at the locations provided by the prior. This method resembles previous works [10], [13], [26].
- **IDS-Known-Dynamic**: Informed direct search (Eq.6) with a known prior on landmark object locations. This is similar to IDS-Known-Static except that the robot does not assume the landmark objects to remain at the locations expressed in the prior.
- **IDS-Unknown**: Informed direct search (Eq.6) without prior on landmark object locations. The particles for landmark objects are initialized uniformly across the environment. This method resembles previous works [17], [3].
- **IHS-Unknown**: Informed hybrid search (Eq.7) without prior on landmark object locations.

All methods except for UDS are using the full *SLiM* model. We assume that an occupancy-grid map of the environment is given. We also assume that the room types are accurately recognized across the environment. IDS-Known-* methods are provided with a noisy prior on landmark object locations which differ from the actual locations, to emulate the common cases where perfect knowledge about landmark locations is not available. For all methods, the particles for the target object are initialized uniformly across

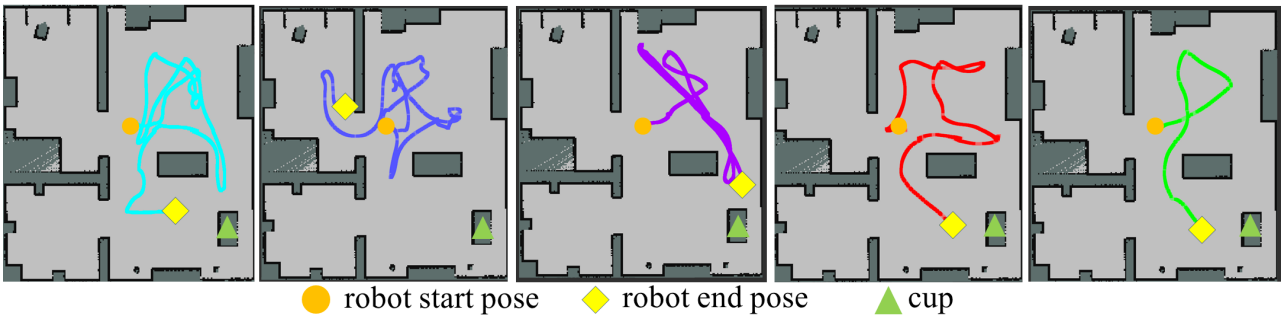

Fig. 6: Examples of search paths generated by each method while searching for *cup*. Methods from left to right: UDS, IDS-Known-Static, IDS-Known-Dynamic, IDS-Unknown, IHS-Unknown. (Best viewed in color).

Target Object	Metrics	UDS	IDS known, static	IDS known, dynamic	IDS unknown	IHS unknown
Coffee Machine	Views	7.83	6.17	4.67	6.33	**3.67**
	Search Time (s)	107	76	60	75	**50**
	Search Path (m)	8.68	6.70	5.80	6.74	**4.93**
	Success Rate	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Laptop	Views	11.00	12.50	7.17	5.67	**4.17**
	Search Time (s)	197	222	124	91	**78**
	Search Path (m)	28.27	26.86	13.13	**7.69**	8.40
	Success Rate	0.83	0.50	1.00	1.00	1.00
Cup	Views	13.17	14.50	12.67	11.83	**9.00**
	Search Time (s)	184	229	189	185	**139**
	Search Path (m)	22.64	29.81	23.40	19.68	**13.91**
	Success Rate	0.83	0.33	0.83	0.83	**1.00**

TABLE I: Benchmark results for object search in simulation experiments. Among methods that reached 100% success rate, IHS unknown successfully found target objects within the smallest number of views and least search time.

the environment.

For each target object, we run 6 trials per method. In each trial, the robot starts at the same location, depicted in Figure 4. The object search is terminated if (1) the belief over the target object location has converged, or (2) the maximum search time of 5mins has been exceeded. A trial is successful if the robot finds the target object before timeout. For each target object and each method, we measure the number of view poses, search time, distance travelled by the robot, and search success rate averaged across all trials.

The benchmark result is as shown in Table I. Examples of the resulting search path from each method are depicted in Figure 6. As we can see, UDS is not as efficient because it is not making use of the spatial relations between the target and landmark objects in the environment. Given a noisy prior on landmark object locations, IDS-Known-Dynamic outperforms IDS-Known-Static because it accounts for the uncertainty of the landmark object locations, whereas IDS-Known-Static is misled by the noisy prior.

Given no prior information, IHS-unknown outperforms IDS-unknown because it encourages the robot to explore promising regions that contain the target and/or useful landmark objects, whereas IDS-unknown only considers promising regions that contain the target object. With IHS-unknown, the robot benefits from finding landmark objects which help narrow down the search region for the target object.

2) Real-World Experiments:: The real-world experiment is executed in an environment (8mx8m) that consists of a kitchen and a living room. The robot stays localized in the pre-mapped environment based on its LIDAR, and navigates based on a MPEPC based path planner [21]. The target object is a cup, and landmark objects include table, sofa, coffee machine and sink. IHS-Unknown reached average success rate of 0.7 (7 out of 10 trials). The average number of view poses, search time and search path is 4.86, 103s, and 8.32m repectively. The failure cases were due to false negative detection of the cup due to lighting (we used Faster R-CNN [22] trained on COCO dataset [16]). Examples of real-world experiments with a Fetch robot is available in online video `https://youtu.be/uWWJ5aV6ScE`.

VII. Conclusion

In this paper we present an efficient active visual object search approach through the introduction of the *SLiM* model. *SLiM* simultaneously maintains the belief over target and landmark objects locations, while accounting for the probabilistic inter-object spatial relations. Further, we propose a hybrid search strategy that draws insights from both direct and indirect object search. Given noisy or no prior on landmark objects locations, we demonstrate the benefit of modeling landmark objects locations under uncertainty in *SLiM*, and the hybrid search strategy that encourages the robot to explore promising areas that can contain the target and/or landmark objects in both simulation and real-world experiments.

REFERENCES

[1] A. Aydemir, A. Pronobis, M. Göbelbecker, and P. Jensfelt. Active visual object search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*, 29(4):986–1002, 2013.

[2] A. Aydemir, K. Sjöö, J. Folkesson, A. Pronobis, and P. Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 2818–2824. IEEE, 2011.

[3] A. Aydemir, K. Sjöö, and P. Jensfelt. Object search on a mobile robot using relational spatial information. In *Proceedings of International Conference on Intelligent Autonomous Systems*, pages 111–120, 2010.

[4] J. Elfring, S. Jansen, R. van de Molengraft, and M. Steinbuch. Active object search exploiting probabilistic object–object relations. In *Robot Soccer World Cup*, pages 13–24. Springer, 2013.

[5] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):381–396, 2002.

[6] T. D. Garvey. Perceptual strategies for purposive vision. Tech. Rep. AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025., 1976.

[7] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner, et al. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine*, 24(3):146–156, 2017.

[8] D. Joho and W. Burgard. Searching for objects: Combining multiple cues to object locations using a maximum entropy model. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 723–728, May 2010.

[9] P. Khandelwal, S. Zhang, J. Sinapov, M. Leonetti, J. Thomason, F. Yang, I. Gori, M. Svetlik, P. Khante, V. Lifschitz, et al. Bwibots: A platform for bridging the gap between ai and human–robot interaction research. *The International Journal of Robotics Research*, 36(5-7):635–659, 2017.

[10] T. Kollar and N. Roy. Utilizing object-object and object-scene context when planning to find things. In *Robotics and Automation (ICRA), 2009 IEEE International Conference on*, pages 2168–2173. IEEE, 2009.

[11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

[12] 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 *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4385–4390. Citeseer, 2012.

[13] L. Kunze, K. K. Doreswamy, and N. Hawes. Using qualitative spatial relations for indirect object search. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 163–168. IEEE, 2014.

[14] J. K. Li, D. Hsu, and W. S. Lee. Act to see and see to act: Pomdp planning for objects search in clutter. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 5701–5707. IEEE, 2016.

[15] B. Limketkai, D. Fox, and L. Liao. Crf-filters: Discriminative particle filters for sequential state estimation. In *Robotics and Automation (ICRA), 2007 IEEE International Conference on*, pages 3142–3147. IEEE, 2007.

[16] 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*, abs/1405.0312, 2014.

[17] P. Loncomilla, J. Ruiz-del Solar, and M. Saavedra. A bayesian based methodology for indirect object search. *Journal of Intelligent & Robotic Systems*, 90(1-2):45–63, 2018.

[18] M. Lorbach, S. Höfer, and O. Brock. Prior-assisted propagation of spatial information for object search. In *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on*, pages 2904–2909. IEEE, 2014.

[19] D. Meger, M. Muja, S. Helmer, A. Gupta, C. Gamroth, T. Hoffman, M. Baumann, T. Southey, P. Fazli, W. Wohlkinger, et al. Curious george: An integrated visual search platform. In *Computer and Robot Vision, 2010 Canadian Conference on*, pages 107–114. IEEE, 2010.

[20] J. M. Mooij. libdai: A free and open source c++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11(Aug):2169–2173, 2010.

[21] J. J. Park, C. Johnson, and B. Kuipers. Robot navigation with model predictive equilibrium point control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4945–4952, Oct 2012.

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2017.

[23] K. Shubina and J. K. Tsotsos. Visual search for an object in a 3d environment using a mobile robot. *Computer Vision and Image Understanding*, 114(5):535–547, 2010.

[24] K. Sjö, D. G. López, C. Paul, P. Jensfelt, and D. Kragic. Object search and localization for an indoor mobile robot. *Journal of Computing and Information Technology*, 17(1):67–80, 2009.

[25] K. Sjöö, A. Aydemir, and P. Jensfelt. Topological spatial relations for active visual search. *Robotics and Autonomous Systems*, 60(9):1093–1107, 2012.

[26] R. Toris and S. Chernova. Temporal persistence modeling for object search. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3215–3222. IEEE, 2017.

[27] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal. Cobots: robust symbiotic autonomous mobile service robots. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 4423–4429. AAAI Press, 2015.

[28] P. Viswanathan, D. Meger, T. Southey, J. J. Little, and A. K. Mackworth. Automated spatial-semantic modeling with applications to place labeling and informed search. In *Computer and Robot Vision, 2009 Canadian Conference on*, pages 284–291. IEEE, 2009.

[29] C. Wang, J. Cheng, J. Wang, X. Li, and M. Q.-H. Meng. Efficient object search with belief road map using mobile robot. *IEEE Robotics and Automation Letters*, 3(4):3081–3088, 2018.

[30] L. E. Wixson and D. H. Ballard. Using intermediate objects to improve the efficiency of visual search. *International Journal of Computer Vision*, 12(2-3):209–230, 1994.

[31] L. L. Wong, L. P. Kaelbling, and T. Lozano-Pérez. Manipulation-based active search for occluded objects. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2814–2819. IEEE, 2013.

[32] Y. Xiao, S. Katt, A. ten Pas, S. Chen, and C. Amato. Online planning for target object search in clutter under partial observability. In *Robotics and Automation (ICRA), 2019 International Conference on*, pages 8241–8247. IEEE, 2019.

[33] Z. Zeng, Y. Zhou, O. C. Jenkins, and K. Desingh. Semantic mapping with simultaneous object detection and localization. In *Intelligent Robots and Systems (IROS), 2018 IEEE/RSJ International Conference on*, pages 911–918. IEEE, 2018.