Decision Theoretic Search for Small Objects through Integrating Far and Near Cues

M Siva Karthik, Sudhanshu Mittal, Gurshaant Malik, K Madhava Krishna

***Abstract*— In an object search scenario with several small objects spread over a large indoor environment, the robot cannot infer about all of them at once. Pruning the search space is highly desirable in such a case. It has to actively select a course of actions to closely examine a selected set of objects. Here, we model the inferences about far away objects and their viewpoint priors into a decision analytic abstraction to prioritize the waypoints. By selecting objects of interest, a potential field is built over the environment by using Composite Viewpoint Object Potential(CVOP) maps. A CVOP is built using VOP, a framework to identify discriminative viewpoints to recognize small objects having distinctive features only in specific views. Also, a CVOP helps to clearly disambiguate objects which look similar from far away. We formulate a Decision Analysis Graph(DAG) over the above information, to assist the robot in actively navigating and maximize the reward earned. This optimal strategy increases search reliability, even in the presence of similar looking small objects which induce confusion into the agent and simultaneously reduces both time taken and distance travelled. To the best of our knowledge, there is no current unified formulation which addresses indoor object search scenarios in this manner. We evaluate our system over ROS using a TurtleBot mounted with a Kinect.**

I. Introduction

With the advent of indoor mobile robots for assistance and service, it is imperative to equip them with intelligent decision making frameworks to perform tasks optimally. One such scenario is where a robot searches for a set of small objects(3-10cm) among many lying scattered on floor in a large unstructured indoor environment. Here, the robot cannot comprehend the whole scene in a single attempt since all the objects would not appear clearly in a single view. Due to the small size of objects, it has to iteratively check various objects while traversing the environment. An active decision making algorithm helps since an exhaustive search over the whole space would be extremely expensive(Fig. 1).

In this work, the robot makes an initial guess about various far away objects, to prune the search space and reach them in an optimal manner for closer recognition through a decision analytic framework. Visible objects are detected using our approach in [1] and an estimate of similarity to each of the query objects [2] is computed to select interest objects. Through a Bayesian Belief Network, it computes the Existential Probability(EP) of an interest object and its pose/angle with respect to its similar query object's reference.

It navigates towards the interest objects to recognize them from a closer proximity through discriminative viewpoints for recognition. We propose Composite Viewpoint Object Potentials(CVOPs), constructed from VOPs(III-D.1) of all query objects to which an interest object is similar. It efficiently represents all common discriminative viewpoints for recognition, with respect to all the objects that an interest object is similar to.

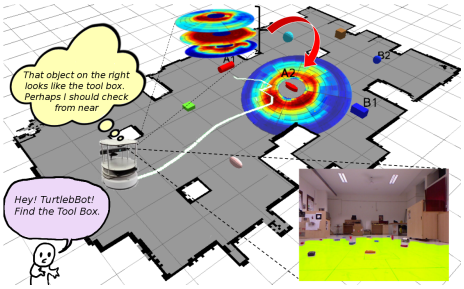

Fig. 1: The robot is required to search for objects among multiple similar looking objects spread all over the floor. It should be enabled with an efficient search strategy.

Through CVOPs, we identify potential viewpoints to reach, to check if the interest objects are indeed the query objects as anticipated from far. Using the above information, we propose a formulation of a Decision Analysis Graph(DAG) to compute the optimal order of waypoints/viewpoints to maximize the reward earned by the agent. Using the existential information about an object from far away to combine it with the object priors(VOP) for efficient decision theoretic planning is the corner stone of this paper.

A. Contributions:

- We propose a formalised adoption of human strategy of guessing about objects from far followed by planned visit to interest objects to recognize them. We propose a DAG which encapsulates the uncertainties to envisage the risk involved in various control actions to actively guide the robot and recover the query objects through a minimal cost.
- We build an efficient object modelling abstraction called Composite VOPs based on our earlier proposal of VOPs in [2]. CVOPs aid in choosing common high accuracy viewpoints with respect multiple similar looking objects helping in clearly disambiguating objects.

II. Related Work

The problem of object search was first addressed in 1976 in [8] where Garvey proposed validation of object hypotheses and reducing the search space. The term 'Active Perception' by Bajcsy in [7]. Most of the recent works exploit the spatial topological relations between the object and structure of its surrounding environment. In [4], Sjöö et al. address the problem by identifying possible locations

*All authors are with Robotics Research Centre, IIIT-Hyderabad.(mkrishna@iiit.ac.in)

of objects in a room and further look closer towards them using a monocular camera with zooming capability. In [5], Joho et al. approach object search in structured indoor environments like supermarkets through a maximum entropy model which ascertains possible locations of a target object using attributes and spatial contexts. In [6], Kunze et al. model the semantic relations between objects and their locations to evolve decision theoretic approach to search for objects in large scale environments. Recently [9][10][11] show extensive work on how strong correlations between 3D structure of the surrounding environment and object placement can be exploited for object search. [10] models a POMDP, making use of uncertain semantics between the object and its location. In it, a probabilistic semantic mapping framework is proposed, defining joint distribution between each object category and room to estimate possible object locations. Most of these works (except [4]) make use of semantic relationships between object-object or object-scene through pre-built knowledge maps or learning the semantic relations. Also, the objects that are dealt within these scenarios are large enough to get reliable feature points to clearly discriminate them from other objects from a substantial distance.

That being said, semantic relations could sometimes be diffused or completely breakdown, leaving the robot in a state of confusion. This could happen in two cases.

- In a case where the robot reaches a large room which it believes is the plausible location of a small object. The object could be anywhere in the room, which makes it expensive to examine each of them closely.
- Agents often encounter completely unorganized and chaotic scenes leading to the breakdown of semantic relationships. Further objects like bottles, toys etc. might not share a semantic relationship with the environment.

Our approach solves certain complimentary aspects in object search which have not been specifically addressed in the earlier mentioned works in accordance to the afore mentioned points.

III. Our Approach

Let $\mathcal{O} = \{o_i\}_1^N$ be the set of N objects that exist in the environment. A robot is given the task of searching for a set of queried objects $O_q \subseteq \mathcal{O}$ among the set of visible objects O_v at a particular instance in the scene. The robot has prior information of the objects($\mathcal{O}$) in the form of their appearance models and the VOP maps.

A. System Pipeline:

The system pipeline is as follows (Fig. 2)

1) All the objects as far as $6m$ around are detected using Sec. III-B presented in [1](Fig. 2(a)). Let O_v denote the set of such objects.
2) For each object in O_v, a likelihood of it being similar to each of O_q is ascertained(Fig. 2(b)). The likelihoods further develop into Existential Probabilities(EP), over a sequence of instances through a Bayesian Belief Net(BBN). EP defines the strength of belief that an object is similar to a query object. Objects which show EPs less than a certain threshold(0.45 in our case) with respect to the query objects, are filtered out. The pose of each of O_v with respect to O_q is estimated through a similar BBN.(Sec. III-C)
3) The early inferences on objects, i.e. the EPs and poses are used to construct the Composite Viewpoint Object Potential(CVOP) which helps identify the discriminative viewpoints for an object(Fig. 2(c)).(Sec. III-D)
4) A potential map over the environment is built using the CVOPs of all objects to choose discriminative viewpoints(Fig. 2(d)). These viewpoints are used to construct a DAG which guides the robot through a set of strategic viewpoints.(Sec. III-F)
5) Once the robot is close to a viewpoint, a path is planned on the CVOP towards the object over which it recognizes the object.[12].(Sec. III-E)

Below we present various modules of the system.

B. Object Detection:

Small objects(1-5cm) on the floor, as far as $6m$ can be detected using our algorithm presented in [1](Fig. 2(a)). It superpixels an image, followed by a Graph Cut over the MRF formulation using the superpixels.

C. Inferring from far:

When a robot observes objects from far away, some objects might look similar to multiple query objects. In our findings in [2], we showed that a reliable inference about such objects can be made using multidimensional Gaussian Mixture Models(GMMs) learnt independently over each object in $\mathcal{O}$. We have a set $\mathcal{G} = \{G\}_1^N$, containing GMMs(G) of each object in $\mathcal{O}$.

1) Existential Probability of an object: The GMMs are continuous Probability Distribution Functions which give the likelihood of a visible object being similar to any of O_q.

If $\mathcal{F}_i^v$ is the feature vector of a visible object, the likelihood $\mathcal{L}_\mathcal{F}$ of $\mathcal{F}$ being similar to an object in O_q is $G(\mathcal{F})$. Hence, for a visible object $o_i^v \subseteq O_v$ We obtain the set

$$\mathcal{L}_i^v = \{G_1(\mathcal{F}), G_2(\mathcal{F}), \ldots, G_q(\mathcal{F})\} \tag{1}$$

which contains the likelihood of $\mathcal{F}_i^v$ being similar to each query object in O_q. We estimate Existential probability($p(E_{o_v}^{o_q})$) for a visible object over a sequence of images through our Bayesian Belief Net formalism. I_1 and I_2 are instances of a visible object o_v. $E_{o_v}^{o_q}$ is a binary random variable which takes 1 when o_v's likelihood $G_q(\mathcal{F})$ with respect to a query object is greater than a threshold. $p(E_{o_i^v}^{o_q})$ is the probability that the object o_i^v is similar to o_q, typically given by $G_q(\mathcal{F})$ which computes the probability that the object o_v exists in the given image I_1. Several such likelihoods are integrated over a BBN. The conditional distribution $p(E_{o_v}^{o_q}|I_1, I_2)$ can be expressed as

$$p(E_{o_v}^{o_q}|I_1, I_2) = \frac{p(I_2|E_{o_v}^{o_q}, I_1)}{p(I_2|I_1)} \tag{2}$$

Using the Markov assumption and the independence between sequence of images,

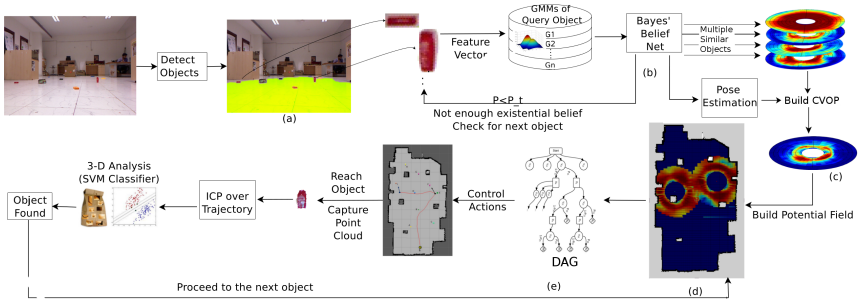

Fig. 2: Object search system overview.

$$P(E^{o_q}_{o_v}|I_1, I_2) = \eta p(E^{o_q}_{o_v}|I_2)p(E^{o_q}_{o_v}|I_1) \tag{3}$$

through Bayes' expansion of $p(I_2|E^{o_q}_{o_v})$, where η is the normalization constant.

In general when there are several images,

$$P(E^{o_q}_{o_v}|I_1, I_2, \ldots, I_n) = \eta p(E^{o_q}_{o_v}|I_1)p(E^{o_q}_{o_v}|I_2)\ldots p(E^{o_q}_{o_v}|I_n) \tag{4}$$

where each term on the right hand side computes the likelihood of object o_v being o_q in that view. Hence, we obtain the Existential probability of o_v with respect to o_q over several views. For a visible object o_v, define S^v_q as the set of query objects that the visible object is similar to.

2) Pose estimation: For each object in $\mathcal{O}$, there are uniformly distributed dictionary images representing various poses. For a visible object with high $p(E^{o_q}_{o_v})$ for a query object, we compare its shape with dictionary images of o_q and find the best match. The shape is obtained using Probability Boundary edge detector and match the shapes using Fast Directional Chamfer Matching(FDCM) [3]. Since we know S^v_q for all objects in O_v we can estimate the pose of each visible object with respect to each query object. We build a belief on $o'_v s$ pose over several images using a similar BBN as proposed above. This process helps us recover the pose reliably with an error of 1-6°.

D. CVOP based Viewpoint sampling:

The construction of CVOPs is as follows(Fig 3).

1) Construction of CVOPs: Viewing angle of an object plays a vital role in small object recognition[2] since they do not provide enough 3-D points from certain views. In case of asymmetric objects the recognition accuracy could vary drastically with viewing angle. In [2] we presented Viewpoint Object Potential(VOP) of an object, a polar map which gives the belief values for correct object recognition as a function of viewing distance and angle. This helps us reach optimal discriminative viewpoints.

A visible object similar to multiple query objects from far away, has a different pose with respect to each of them. The robot has to view an object from different viewpoints specific to various query objects for disambiguation. To avoid this cumbersome task, we propose a Composite VOP map for an object. Effectively, a CVOP contains the *combined* high accuracy viewpoints related to all the similar query objects, hence saving the robot from visiting numerous viewpoints.

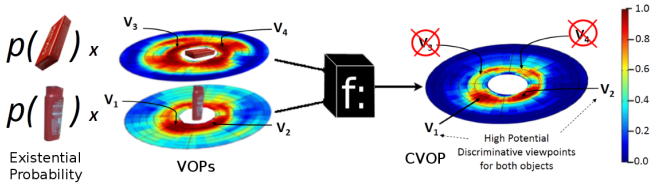

Fig. 3: Figure shows the formation of CVOP of an object which appears to be both shampoo and cookie pack from far, rotated at 0° and 30° with reference to their base frames. The VOPs of them are weighed by existential probability, rotated and merged to form the CVOP. v_1, v_2 are high accuracy regions from shampoo and so are v_3, v_4 for cookie pack. v_3 has a low potential to recognize shampoo bottle(From its own VOP) and so is the case with v_4. These viewpoints cannot cater to be high accuracy viewpoints since they help in reliably recognizing only one of the objects. On the other hand, v_1 and v_2 are high accuracy viewpoints in both the VOPs which means they have to be retained in the CVOP.

CVOP(Fig. 3) is a weighted composition of VOPs of all query objects similar to the visible objects. The weight of each VOP is decided by $p(E^{o_q}_{o_v})$, the existential belief of o_v being o_q. Further, the VOPs being polar plots are oriented according to the pose of o_v with respect to o_q and merged. Say o_v is similar to o_1, o_2, …, $o_t \in O_q$ with existential probabilities, $p(E^{o_1}_{o_v})$, $p(E^{o_2}_{o_v})$, …, $p(E^{o_t}_{o_v})$ and its angles estimated with respect to the query objects are θ_1, θ_2, …, θ_t. A CVOP is calculated as

$$\begin{aligned}\mathcal{CV}_{o_v}(\mathcal{V}_{o_1}(\theta_1), \mathcal{V}_{o_2}(\theta_2), \ldots) = min(\ & p(E^{o_1}_{o_v})\mathcal{V}_{o_1}(\theta_1), \\ & \ldots, p(E^{o_t}_{o_v})\mathcal{V}_{o_t}(\theta_t))\end{aligned} \tag{5}$$

where $\mathcal{V}_{o_i}(\theta_i)$ is the VOP of query object o_i oriented at angle θ and the function $min(.)$ assigns the minimum value from the VOPs at each viewpoint to that in CVOP.

The advantages of using a CVOP for an object are two fold. Firstly, CVOP generates a new potential field, where a viewpoint indicates the minimal probability of recognizing an object as one of the query objects it was anticipated to be. Secondly, it models a generative abstraction of the structural and discriminative properties of the object in the probability space which can be used for any kind of active recognition or manipulation tasks.

2) Trajectories towards objects: High potential(red) points in a CVOP form clusters with clear boundaries(Fig. 3). Initial Viewpoints are sampled over such outer boundaries(e.g. v_2 and v_4). With these viewpoints as the start, a trajectory $(v_1, v_2, \ldots, v_n)$ is computed towards the object, constraining the robot to contain the object in the camera frame always. From a viewpoint v_i on the boundary, select one of its 8 neighbours with the constraint that the object is contained in the camera view and the neighbour either has a

higher or equal potential compared to the current viewpoint and proceed iteratively towards the object(Fig 8(b),(f)).

3) Recognition Belief of a viewpoint: We use a BBN similar to what is proposed in Sec. III-C.1 to build a Recognition Belief(RB) for an object, over the trajectory starting from v_i. For an object o_v, which appeared similar to o_a from far, the Recognition Belief, $p(R^{o_a}_{o_v}|\{v\}^n_1)$ is

$$P(R^{o_a}_{o_v}|v_1, v_2, \ldots, v_n) = \eta p(E^{o_a}_{o_v}|v_1)p(E^{o_a}_{o_v}|v_2)\ldots p(E^{o_a}_{o_v}|v_n) \quad (6)$$

where $p(E^{o_a}_{o_v}|\{v\}^n_1)$ comes from VOP of o_a which indicates the probability of recognizing o_v as o_a from that viewpoint. This RB is used while calculating the success and failure probabilities of a chance node for the DAG(Sec. III-F).

E. Recognition from near:

In the current context, active object recognition[15] is beneficial over static recognition techniques. In our previous work [2], object recognition was based on single frame RGB-D data from Kinect, where recognition is highly dependent on viewpoint selection. In this work, we use multiple frames to form a well aligned dense cloud of the object.

While the robot approaches an object, it starts at a strategic viewpoint on the outer boundary of high potential region where it initiates the ICP module and travels towards the object over the trajectory calculated in Sec. III-D.2. The robot incrementally registers pair of clouds of the object using ICP [16][14]. This results to an aligned dense point cloud of the object with several new ones added in the updated 3D cloud for robust recognition.

F. DAG based exploration planning:

After finding the interest objects, the robot has to navigate to various viewpoints to recognize them as anticipated. Some objects showing high Existential Probabilities(EP) might be far away, some showing lower probability might be close by and vice versa. The robot might go all the way to a far object and fail to find the object while it could have gone to a closer object with lower EP for a successful recognition. Here, we propose a Decision Analytic framework through a Decision Analysis Graph(DAG)[13] which helps us find an optimal set of waypoints to maximize the reward earned by the agent.

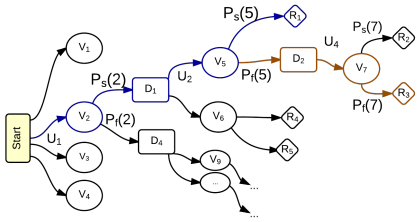

Fig. 4: Decision Analysis Graph depicting the various nodes and controls. The rectangular nodes are decision nodes where the next best control action is chosen to the next viewpoint(chance nodes) with highest utility. The circular nodes are Chance nodes and the leaves are reward nodes. Blue shows a path incurring success edges all along. The path in orange incurs two failures. Waypoints are nothing but the viewpoints of different objects the robot has to reach.

1) Construction of DAG: A DAG(Fig 4) is a Directed Acyclic Graph, $D = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ and Edges $\mathcal{E}$. wherein $\mathcal{V} = \mathcal{V}_c \cup \mathcal{V}_d \cup \mathcal{V}_r$ is the union of Chance Nodes($\mathcal{V}_c$), Decision Nodes($\mathcal{V}_d$) and Reward/Leaf Nodes($\mathcal{V}_r$). $\mathcal{E} = \mathcal{E}_p \cup \mathcal{E}_u$ is the union of Result Edges($\mathcal{E}_p$) and Control Edges($\mathcal{E}_c$).

- *Chance Nodes* represent the discriminative viewpoints sampled from the outer boundary of the high potential area of a CVOP(Fig. 3). For the visible object o_v similar to query objects $O_s = \{o_1, o_2, ..., o_k\} \subseteq O_q$, the success probability at a node is given by

$$P^V_s = p(R^{o_1}_{o_v}|v^n_1) \cup p(R^{o_2}_{o_v}|v^n_1) \ldots \cup p(R^{o_k}_{o_v}|v^n_1) \quad (7)$$

  which indicates the probability of recognizing o_a as either of the anticipated objects, where $p(R^{o_1}_{o_v}|v^n_1)$ is the probability of recognizing o_v as o_a at the end of the trajectory. The failure probability P_f is 1-P_s
- *Result Edge-Success(RE_s)* is an edge from a Chance node which indicates successful recognition of the object as one of O_s with a probability P_s.
- Similarly, *Result Edge-Failure(RE_f)* is an edge from a Chance node which indicates the failure to recognize the object as one of O_s as anticipated from far away.
- *Decision nodes* are where the decision about the next control action to reach a new object/viewpoint is decided. At every v_d, that next viewpoint is chosen which has the highest amount of utility through the *Control Edge*(E_c) leading to it.
- *Reward/Leaf Nodes* are leaf nodes of the DAG which indicate the reward earned by the robot at the end of exploration through the set of waypoints from the root node to v_r.

The root node(Start) is a Decision node which propagates a set of Control edges leading to a reward(v_r) finally. The reward at a v_r is inversely proportional to the distance travelled by the control actions to reach it. A failure edge leads to a penalty for choosing a control action that led to a failure. For instance, in Fig. 4, the reward at R_1 for the path(Fig. 4(blue)) via $Start \rightarrow U_1 \rightarrow P_s(2) \rightarrow U_2 \rightarrow P_s(5) \rightarrow R_1$ where it passes through two success edges(P_1, P_2) consecutively is

$$R_1 = 1/d(Start, V_2) + d(V_2, V_5) \quad (8)$$

where $d(Start, V_2)$ is the distance travelled from Start node to viewpoint V_2 when it executed the control action U_1. In other terms, reward at R_1 is

$$R_1 = 1/d(U_1) + d(U_2) \quad (9)$$

Whereas the reward through a path that failed to detect one or more objects progressively reduces with increasing number of non-detected objects along the path. For example, the reward along the path(Fig. 4(orange) $start \rightarrow U_1 \rightarrow P_s(2) \rightarrow U_2 \rightarrow P_f(5) \rightarrow U_4 \rightarrow P_f(7) \rightarrow R_3$ is computed to be

$$R_3 = 1/d(U_1) + k_1(d(U_2) + k_2 d(U_4)) \quad (10)$$

where k is a penalty to reduce the reward for failing to recognize the object from V_5 as an anticipated object from

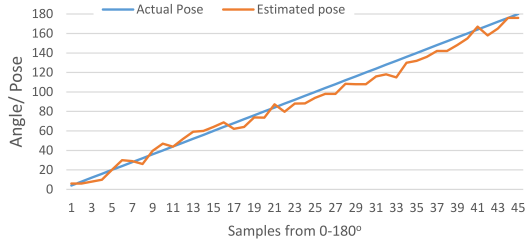

Fig. 5: Pose estimation error for a slim sidelined-wide bodied object, for samples from $0°$-$180°$ each differing by $4°$. A single instance of the sample is considered for estimation. The estimated pose is close to actual pose.(Best viewed when zoomed in.)

the query objects. However, the object viewed from V_2 was successfully recognized as one of query objects and hence there is no penalty with the distance $d(U_1)$.

Now that the reward nodes are computed, the expected reward at each chance node(from where the object needs to be viewed) is recursively computed bottom up from each leaf node. For instance, the utility at V_5 would be

$$\mathcal{U}_{V_5} = P_s(5)R_1 + P_f(5)(\mathcal{U}_{V_7}) \tag{11}$$

which recursively simplifies to

$$\mathcal{U}_{V_5} = P_s(5)R_1 + P_f(5)(P_s(7)R_2 + P_f(7)R_3) \tag{12}$$

Starting from the 'Start' node, at every v_d, control to that viewpoint is chosen which has the highest utility to get the set of waypoints for the robot to navigate.

2) Advantages of DAG: Firstly, the DAG provides a mechanism for integrating failure probabilities into the expected reward and hence eventual decision making. Secondly, it provides for alternative best paths, which can be computed a-priori. DAG efficiently encapsulates the possibility of failure in recognizing the object as it was anticipated.

IV. Experiments and Results

A. Pose Estimation from far:

In Fig. 5 shows that the error incurred in estimating the pose for an object over a single instance is $3°$-$10°$. Fig 6 depicts the increase/saturation in the belief of the object being in pose $48°$, while moving towards the object and the actual pose is $45°$, with an error of $3°$.

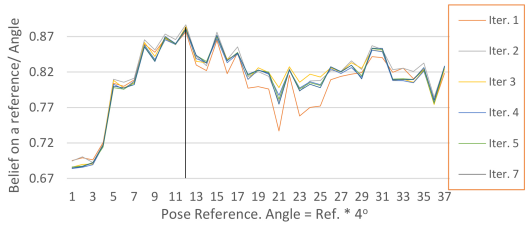

Fig. 6: Here the objects is oriented at $45°$ and we build the belief on pose through BBN over multiple instances while moving towards it. The belief over dictionary image 12(at $48°$) is the highest after 7 iterations.(Best viewed when zoomed in.)

B. Analysis of ICP based recognition:

When registering multiple point clouds, it is important to determine the optimal angular shift between two images and the number of such frames to be considered. Fig 7 shows that the recognition accuracy is the best when 4 consecutive frames were registered with an angular shift of $5°$ between

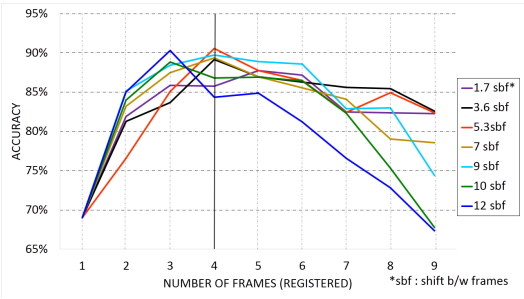

Fig. 7: A visible increase in recognition accuracy through ICP based recognition. The best accuracy is achieved when 4 frames are registered each with a shift of $5.3°$ between them.(Best viewed when zoomed in.)

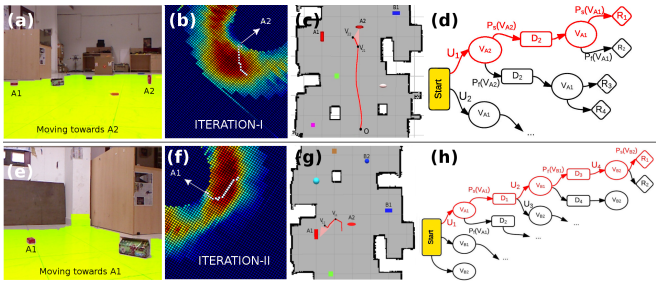

Fig. 8: (a),(e)Robot detects the interest objects and segments them. (b),(f)Robot identifies the strategic viewpoints and the paths from them for ICP. (c),(g)Robot moves to the viewpoints for recovering the objects as decided in(d),(h).

them. It can be seen that the overall recognition accuracy for a typical small object of size 11x6x$2cm$ is increased from 69% to 89% when number of registered from is increased from 1 to 4 frames.

C. Analysis of our approach:

Here, we demonstrate the functioning of the pipeline(Fig. 8) over a scene where there are multiple similar objects. Fig. 8 shows the first two iterations of the search. There are three query objects (*Red Shampoo Bottle(SB), Red Cookie Pack(RC), Blue Tool Pack(BP)*) that need to be recovered. SB and RC are similar looking from far and BP is different from the other two(Fig. 9). The robot discovers objects A_1, A_2(Fig. 8(a)) both of which have high Existential Probability(EP) for SB and RC(using Sec.III-C). After calculating their pose, it constructs CVOPs for A_1, A_2 using the VOPs of SB and RC to compute the high accuracy viewpoints V_{A_1} and V_{A_2}(Fig. 8(b),(f)). Through a DAG built over the viewpoints as in Fig. 8(d), it decides to reach V_{A_2}(as it has the highest utility value) and recognizes it(Sec.III-E) as SB, disambiguating it from RC(Fig. 8(c)). Here, it further discovers B_1 and B_2, both of them are similar to BP and hence the robot now has to explore among A_1, B_1 and B_2. The DAG is reconstructed(Fig. 8(h)) with the inclusion of the new found objects' viewpoints. Following this, it further navigates to A_1(Fig. 8(g)) since its viewpoint has the highest utility among those for A_1, B_1 and B_2. Following the same recognition strategy, it recognizes A_1 as RC, as anticipated. Following the control from the DAG, among B_1 and B_2, it traverses over to V_{B_1} where it fails to recognize it as one of the query objects and moves towards B_2 to recognize it as BP. Fig 9(a).(our approach), shows the path traced by the robot in the whole search mission. In 10 trials, the robot covered an average of $11.65m$ and recovered all the queried objects 8 times. Over all we conducted experiments in 5

Fig. 9: (Left)Paths traced by the robot through various strategies are shown. The path traced by our approach is on similar lines with the human approach. (Right)Paths traced during various human based experiments. Almost all the human subjects approach the objects that are similar to query objects. A_1, A_2, B_1, B_2 are Red Cookie Pack, Red Shampoo, Multimeter and Blue Tool Box respectively. The arena is of the size 12m x 7m (Best viewed when zoomed in.)

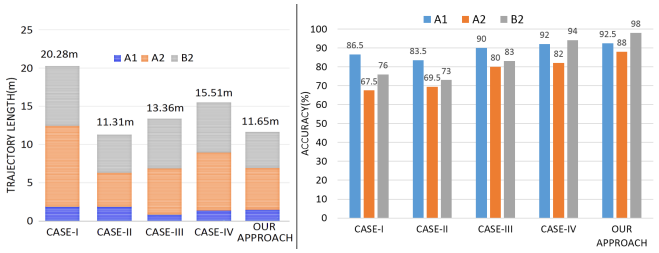

Fig. 10: (Left) Distance travelled by robot to reach a search object in each case for the above scenario. The distance travelled in our case is considerably less. (Right) The accuracy of recognizing different objects in various cases over all 50 experiments.

different setups, each with 10 trials where the robot could recover all the queried objects in 92% of the trials(Fig. 10).

D. Comparison with various strategies:

1) Case 1: Here, the robot greedily searches through every object present in the scene until the query objects are found. The robot ends up covering a large distance of $20.28m$(Fig. 10). Over several runs, the reliability is 76%(Fig. 10), which means were successfully recovered only in 76% of the cases. This is due to the poor viewpoint selection in many cases(Fig. 9.case1).

2) Case 2: When the robot is equipped with the GMM module, it eliminates most of the dissimilar objects in the scene. In this case, while the average trajectory length is reduced to $11.31m$(Fig. 10, 9.case2), the reliability of recovering objects is still very low(75.3%) due to the poor viewpoint selection. Although the GMM predicts the presence of an object, the robot cannot choose a strong viewpoint.

3) Case 3: Here, the robot is equipped with both the GMM module and VOP maps. The 3-D analysis module uses only a single point cloud. The robot infers about an object from far and moves to a strategic viewpoint. But the robot does not do this through a DAG and hence does not know what steps need to be taken when it fails to recognize an object as anticipated. The reliability of recovering objects increases to 84.3%(Fig. 10) here and the distance traversed also increases by a small amount($13.36m$)(Fig. 9.case3) since the robot reaches specific viewpoints. The reliability is still bounded since it checks an object only once in which it could fail.

4) Case 4: Here, we demonstrate the performance of an approach proposed by us in [17]. For object A_2, it moves to multiple viewpoints to recognize it. This leads to a lot of redundancy. So is the case with B_1. This leads to an increased distance traversal($15.51m$)(Fig. 9.case4) compared to the approach proposed in this paper. The object recovery accuracy is higher than the above specified cases(84.3%) but still remains lower compared to our approach.

E. Human based experiments

We explored how humans perform in a similar setting with same visual and motion capabilities as the robot. In this experiment, the person was not exposed to the arena and had remotely controlled the robot while watching the live video stream from the robot's camera. We observed that humans try to guess about objects from far and go closer to recognize it. Also, the paths traced by humans are very similar to the robot as shown in Fig. 9(right). This experiment comprises 20 trials, each with a different person. Although, the object recognition ability of humans marginally outperforms our algorithm, the average distance travelled by humans is 14.63m as compared to 11.65m by robot for the same setup. The paths are plotted in Fig. 9

REFERENCES

[1] S.Kumar, M.S.Karthik and K.Madhava Krishna, *Markov Random Field based Small Obstacle Discovery over Images*, IEEE ICRA, 2014.

[2] S.Mittal, M.S.Karthik and K.Madhava Krishna, *Small Object Discovery and Recognition Using Actively Guided Robot*, IEEE ICPR, 2014.

[3] M.Liu, O.Tuzel, A.Veeraraghavan and R.Chellappa, *Fast Directional Chamfer Matching*, in IEEE CVPR, 2010.

[4] K Sjöö, G L Dorian, P Chandana and P Jensfelt and D Kragic *Object Search and Localization for an Indoor Mobile Robot*, JCIT, 2009.

[5] D.Joho and W.Burgard *Searching for objects: Combining multiple cues to object locations using a maximum entropy model*, IEEE ICRA, 2010

[6] L.Kunze, M.Beetz, M.Saito, H.Azuma, K.Okada, M.Inaba *Searching Objects in Large-scale Indoor Environments: A Decision-theoretic Approach*, IEEE ICRA, 2012

[7] R. Bajcsy, *Active perception*, Proc. IEEE, 1988.

[8] T. Garvey, *Perceptual strategies for purposive vision*, AI Center, SRI International, Menlo Park, CA, USA, Tech. Rep. 117, Sep 1976.

[9] A. Aydemir, and P. Jensfelt, *Exploiting and modeling local 3D structure for predicting object locations*, in IEEE/RSJ IROS, 2012.

[10] A. Aydemir, A. Pronobis, M. Göbelbecker, and P. Jensfelt, *Active Visual Object Search in Unknown Environments Using Uncertain Semantics*, in IEEE Transactions on Robotics, 2013.

[11] K. Sjöö, A. Aydemir, and P. Jensfelt, , *Topological spatial relations for active visual search*, Robotics and Autonomous Systems, 2012.

[12] G. Csurka, C. Bray, C. Dance, and L. Fan, *Visual categorization with bags of keypoints*, Workshop on Statistical Learning in Computer Vision, ECCV, 2004.

[13] Finn V. Jensen and Thomas Nielsen, *Bayesian Networks and Decision Graphs (Information Science and Statistics)*, July, 2001.

[14] R. Rusu, and S. Cousins, *3D is here: Point Cloud Library (PCL)*, in IEEE ICRA, 2011.

[15] Bjorn Browatzki, Vadim Tikhanoff, Giorgio Metta, Heinrich H. Bülthoff and Christian Wallraven, *Active Object Recognition on a Humanoid Robot*, in IEEE ICRA, 2012.

[16] A.W.Fitzgibbon, *Robust Registration of 2D and 3D Points Sets*, in British Machine Vision Conference, 2001.

[17] M.S.Karthik, S.Mittal and K. Madhava Krishna, *Guess from Far, Recognize when Near: Searching the Floor for Small Objects*, in ICVGIP, 2014.