Bootstrapping Probabilistic Models of Qualitative Spatial Relations for Active Visual Object Search

Lars Kunze and **Chris Burbridge** and **Nick Hawes**
Intelligent Robotics Lab
School of Computer Science
University of Birmingham
United Kingdom
{l.kunze|c.j.c.burbridge|n.a.hawes}@cs.bham.ac.uk

Abstract

In many real world applications, autonomous mobile robots are required to observe or retrieve objects in their environment, despite not having accurate estimates of the objects' locations. Finding objects in real-world settings is a non-trivial task, given the complexity and the dynamics of human environments. However, by understanding and exploiting the structure of such environments, e.g. where objects are commonly placed as part of everyday activities, robots can perform search tasks more efficiently and effectively than without such knowledge. In this paper we investigate how probabilistic models of qualitative spatial relations can improve the performance in object search tasks. Specifically, we learn Gaussian Mixture Models of spatial relations between object classes from descriptive statistics of real office environments. Experimental results with a range of sensor models suggest that our model improves overall performance in object search tasks.

1 Introduction

Many proposed, near-future applications of autonomous mobile robots involve them having to find objects in everyday environments (either to fetch them, or to report something about them), usually without the robot knowing precisely where they are (Kunze et al. 2012; Holz, Iocchi, and van der Zant 2013; Aydemir et al. 2013; Williams et al. 2013; **?**). This problem of finding an object in an environment, usually termed *active visual search* or *object search*, is the topic of this paper. In particular we investigate an approach which allows a robot to exploit knowledge about the structure of the environment in order to find an object more quickly than is possible without the knowledge.

The reason that object search is considered a necessary robot ability is that objects do not all stay in fixed positions. The reason objects do not stay in fixed positions is that many objects play central roles in human activities, and these activities usually involve moving the objects in some way. In an office environment this may be as limited as moving a mouse whilst using a PC or moving a mug whilst drinking coffee whilst working, up to moving mugs, laptop and books from one room to another for a meeting. Objects are not moved randomly; they are moved based on their function and their role in an activity. Thus, whilst object positions may vary over time, many objects vary in predictable patterns. Our hypothesis is that these patterns can be captured in Qualitative Spatial Relations (QSRs), relations which capture the important structure of positional variation whilst abstracting over (unimportant) quantitative variation.

Our work is motivated by two complementary scenarios. The first is a scenario in which a human asks a robot to retrieve an object using a prepositional phrase, e.g. "fetch me the book that I left on my desk near my laptop". This requires that the robot is able to map from the qualitative linguistic description to an acceptable quantitative target location on the correct desk. Our second motivating scenario is the (non-linguistic) task of locating an object given long-term experience of the locations of objects from the same category, e.g. searching for a mug using the past observations of mug locations as a guide. Unlike the first scenario, this latter case does not explicitly require a relational model of object location, but our hypothesis is that – due to the aforementioned regularities in human activities – qualitative relational models are more compact and provide more accurate predictions than models based purely on accumulated metric position information, provided a suitable landmark is used for the relation. A suitable landmark is generally an object whose location does rarely change qualitatively in an environment such as the desk in the example above.

As described in Section 2, the problem of including prior knowledge in object search tasks has been studied previously. This paper makes the following contributions beyond this prior work:

- a set of Gaussian Mixture Models which encode a range of QSRs based on descriptive statistics taken from a real office environment that can predict the position of objects given a landmark;
- an entropy-based approach to the selection of an appropriate landmark for use in indirect search;
- and an analysis of the performance of our search approach as the field of view of the robot's sensor varies.

2 Related Work

Active visual search has become a popular topic in mobile and service robotics recently. Work done by Aydemir,

Pronobis, Sjöö and others in the CogX project (Aydemir et al. 2011; 2013; Sjöö, Aydemir, and Jensfelt 2012) introduced a novel, sampling-based approach to object search using object location probability distributions attached to a 3D volumetric representation. This approach, which forms the foundation of the work in this paper, provides an effective and flexible approach to active visual search. Whilst the general case of object search is computationally complex (Tsotsos 1992), the task can be made more tractable through the use of an intermediate landmark object which constrains the location of the target object (Wixson and Ballard 1994). This is known as *indirect search*. The CogX work (Sjöö, Aydemir, and Jensfelt 2012) used the spatial relations "in" and "on" to define object targets relative to landmarks. We go beyond this work by using more restrictive spatial models to provide more tightly defined location predictions. Other recent work on object search has tackled larger scale space but used predefined view cones within rooms (Kunze et al. 2012), or has allowed searching over rooms or scenes for unknown objects without constraining their location in 3D (Joho, Senk, and Burgard 2011; Kollar and Roy 2009). We differ from these approaches as we are able to generate arbitrary views in 3D, but constrained to legal robot poses and likely object positions.

In order to take qualitative descriptions of object positions and use them to predict 3D object poses, a robot must be able to mediate between such qualitative and quantitative representations. Previous work has either used hand-coded spatial models (Brenner et al. 2007) or has learnt models from experience in 3D (Burbridge and Dearden 2012; Sjoo and Jensfelt 2011). We take the latter approach, and, in line with prior work, make use of simulation to generate training data for our system.

Other approaches to predicting object positions have used conceptual knowledge to make coarsely predictions, i.e. at the room level (Hanheide et al. 2011; Samadi, Kollar, and Veloso 2012), or have utilised learnt object location predictors in the image plane (Kollar and Roy 2009; Aydemir and Jensfelt 2012). These other approaches provide complementary information to our approach and could be integrated into a single system. As can the work on learning more general *organisation principles* of everyday environments in order to support efficient robot task performance (Schuster et al. 2012).

3 Bootstrapping Probabilistic Models of Qualitative Spatial Relations

To perform indirect search using QSRs, a system requires models of QSRs which describe the locations of target objects relative to landmarks. Such models can either be manually specified based on scene geometry using one of the many available QSR calculi for describing relative position (Cohn and Hazarika 2001) or they can be learnt from observations of collections of objects (Burbridge and Dearden 2012; Sjoo and Jensfelt 2011). We have chosen the latter approach as relative object positions vary with the category of both target and landmark as well as with relation, and existing calculi are not well suited to capture the specificities of such variations.

Our approach to learning QSR models for indirect object search first requires a collection of training data containing objects segmented and category labelled in 3D. In this paper we focus on searching for objects on desks, so we require this data to be observations of desktop object configurations. This data must then be labelled with QSRs to indicate which observations should contribute to which learnt relational model. These labels could be based on human language, i.e. provided by a human annotator or a geometric QSR calculus, in which case the learnt models will allow for object location predictions to be generated using one of these labels as input (as in the book example in Section 1). Alternatively the relations could be provided by some unsupervised process which discovers the predictive structure in the observations (Behera, Cohn, and Hogg 2012; Sridhar 2010). In either case this step discards (hopefully irrelevant) relations from the exhaustive pairwise calculation of relations between all objects in any given scene. After this step, the training data provides a collection of observations for each triple of QSR type, landmark type and target type, e.g. all the observations of the book being near the laptop. These collections can then be turned into a generative model using an appropriate algorithm. Section 4 describes how we have obtained training data for the current system, and Section 5 describes the Gaussian Mixture Model approach we have used to learn models for our current system.

4 Acquiring data

To acquire the training data for learning QSR models we use an approach comprised of three steps. At first, we produce a statistics on objects and their qualitative relations by analysing and labelling images of real-world office desks. Secondly, on the basis of this statistics, we automatically generate a multitude of novel scenes of office desks using a physics-based simulator. In a final step, the geometric relations between objects in a scene are labelled automatically using a QSR-based calculus. This approach has several advantages: first, a human has only to label objects and their relations qualitatively, second, the physics-based simulation automatically generates the geometric information of labelled objects in 3D, and third, the simulator can theoretically generate an infinite amount of training data.

Annotating images of real desktop scenes

We bootstrapped an object statistics about qualitative spatial relations by annotating 47 images of real-world office desks. Figure 1 shows some examples of these images. After a first sight of the real-world images, we decided to annotate the following 18 object categories: *Book, Bottle, Calculator, Cup (or Mug), Desktop PC, Glass, Headphone, Hole punch, Keyboard, Keys, Lamp, Laptop, Mobile phone, Monitor, Mouse, Pen (or Pencil), Stapler*, and *Telephone*.

In each of the 47 images we labelled how many instances of the above objects categories were present. Further, we specified which of the following directional and/or distant relations hold between an object and a landmark:

- *left-of*, *right-of*, *in-front-of*, *behind-of*, and

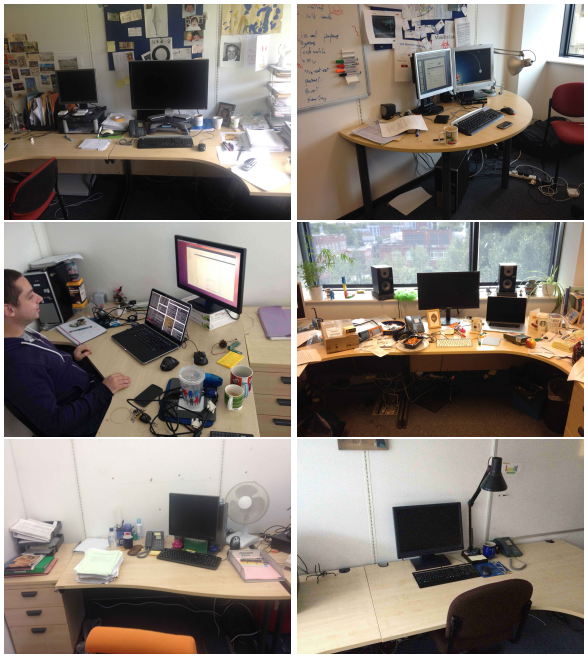


Figure 1: Image data of desktop scenes.

- *close-to*, *distant-from*.

As the number of pairwise relations between object instances is huge, we decided to only label relations with respect to two landmark objects, namely *Monitor* and *Keyboard*. For these landmark objects we also noted the approximate location on the desk as *north, east, south, west,* $center_{\mathrm{ns}}$, and $center_{\mathrm{we}}$.

Overall, 437 object instances have been identified in 47 scenes, and 1798 spatial relations have been labelled. The following tables summarize the results: Table 1 shows the probability distribution of the presence of an object in a scene. Table 2 shows the results of the approximate landmark locations on desks and Table 3 shows exemplarily some distributions of qualitative spatial relations between objects and the *Monitor* landmark.

Table 1: Probabilities of the presence of objects in a scene.

Object type	$P_{pres}(obj)$	**Object type**	$P_{pres}(obj)$
Book	0.59	Keys	0.10
Bottle	0.44	Lamp	0.14
Calculator	0.12	Laptop	0.17
Cup/Mug	0.63	Mobile phone	0.17
Desktop PC	0.38	Monitor	0.95
Glass	0.08	Mouse	0.82
Headphone	0.10	Pen/Pencil	0.63
Hole punch	0.04	Stapler	0.23
Keyboard	0.87	Telephone	0.68

Table 2: Probability distribution of the approximate locations of landmarks on a desk.

Landmark	**Location**	west	$center_{\mathrm{we}}$	east
Monitor	north	0.22	0.51	0.09
	$center_{\mathrm{ns}}$	0.09	0.07	0.02
	south	0.00	0.00	0.00
Keyboard	north	0.02	0.10	0.00
	$center_{\mathrm{ns}}$	0.22	0.35	0.12
	south	0.05	0.12	0.02

Table 3: Probabilities that a spatial relation holds between an object and the *Monitor* landmark.

Object	*left*	*right*	*front*	*behind*	*close*	*distant*
Book	0.58	0.34	0.71	0.08	0.30	0.65
Cup/Mug	0.25	0.61	0.74	0.02	0.44	0.53
Deskt. PC	0.11	0.22	0.11	0.22	0.94	0.05
Keyboard	0.06	0.12	0.91	0.00	0.82	0.10
Mouse	0.04	0.76	0.95	0.02	0.44	0.51

Generating data of simulated desktop scenes

On the basis of the statistics on the presence of objects in a scene (Table 1), the approximate locations of landmark objects (Table 2), and the qualitative spatial relations between objects and landmarks (Table 3) we generate novel scenes of office desks using a simulator. Figure 2 depicts examples of automatically generated scenes.

The scene generation works as follows: (1) we sample a set of object instances to be present in a scene whereby we assume that at least one of the two landmark objects is present (*Monitor* or *Keyboard*), (2) we sample a location for the principle landmark object on the desk, (3) we sample a set of qualitative relations for each object with respect to the landmark, (4) The qualitative relations are transformed into euclidean angles and distances by using a generative model of the ternary point calculus (Moratz, Nebel, and Freksa 2003). For a more detailed account on how the qualitative representations are transformed into metric representations and vice versa please refer to (Kunze and Hawes 2013), (5) finally, we test whether objects are on the table and/or in collision with each other and possibly apply backtracking in the scene generation process.

In total, we generated 500 scenes of simulated office desks based on the bootstrapped statistics on the real-world environment.

A successfully generated scene is fully described by the set of present object instances, their types, their 3D poses and their bounding boxes. Figure 3 visualizes the positions of instances of different object types projected onto the table surface with the dimensions of $2.0\ \mathrm{m} \times 1.2\ \mathrm{m}$. Please note the correlation between the qualitative and the quantitative positions of *Monitor* and *Keyboard* in Table 2 and Figure 3 respectively.

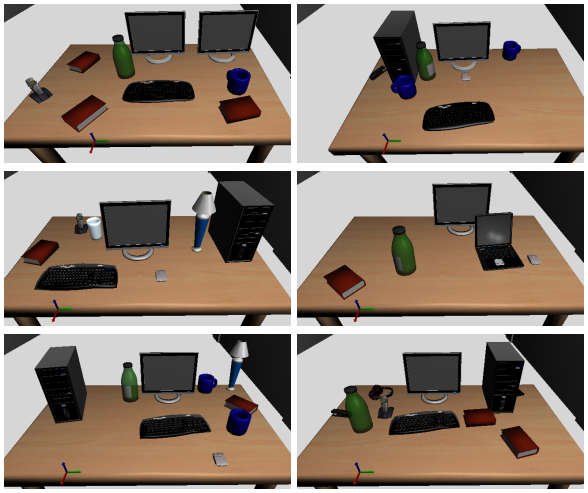


Figure 2: Automatically generated scenes of office desks.

Labelling simulated scenes with QSRs

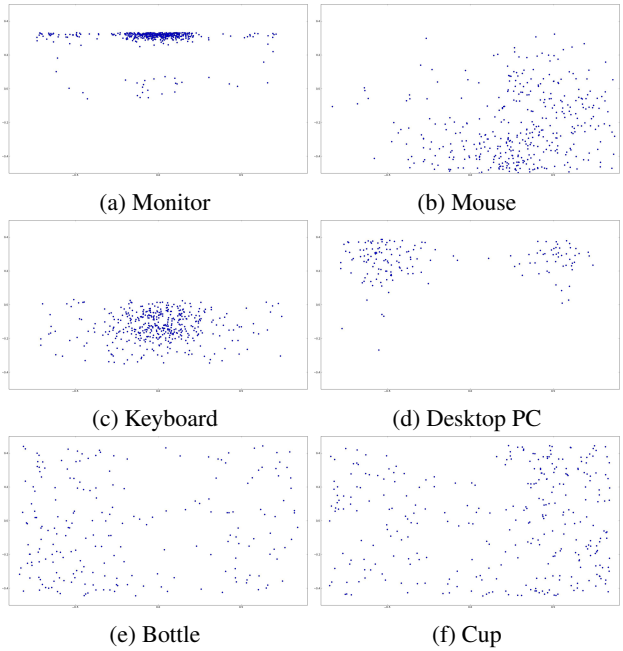
In order to learn models of QSRs from the generated scenes we label the relations between objects using the ternary point calculus (Moratz, Nebel, and Freksa 2003). We assume that a robot is standing about 2 meters in front of a generated desk and calculate a reference axis between the robot and the landmark. According to the *relative angle* and the *relative radius* we label the relation between an object and the landmark as *left-of*, *right-of*, *in-front-of*, *behind-of*, *close-to*, and *distant-from*. However, now each object instance is also considered as a landmark. Thereby we are able to generate more QSR labels between objects than those that have been manually produced in the object statistics from the real-world images. In the next section, we describe how a set of qualitative relations that holds between an object and a landmark such as *in-front-of(Object, Landmark)* $\wedge$ *close-to(Object, Landmark)* is represented by a multivariate Gaussian distribution and how different distributions (or sets of QSRs) are combined using Gaussian Mixture Models.

5 Learning Gaussian Mixture Models of Qualitative Spatial Relations

In this section, we first explain how we learn Gaussian Mixture Models (GMMs) to predict the position of an object given a landmark, and secondly, we introduce an entropy-based measure that describes the predictive capability of landmarks.

Predicting an object position given a landmark

In Section 4 we explained how we generated a labelled data set for learning QSR models on the basis of real-world office desks. As shown in Figure 3, the positions of a cup are almost uniformly distributed over the office desk. Hence, it is difficult to predict the position of a cup based on this metric information only. By considering QSRs with respect to other table-top objects (landmarks), we learn probabilistic models that predict the relative position of an object more precisely.



(a) Monitor (b) Mouse (c) Keyboard (d) Desktop PC (e) Bottle (f) Cup

Figure 3: Positions of objects on a desk of the size 2.0 m$\times$1.2 m (width $\times$ depth).

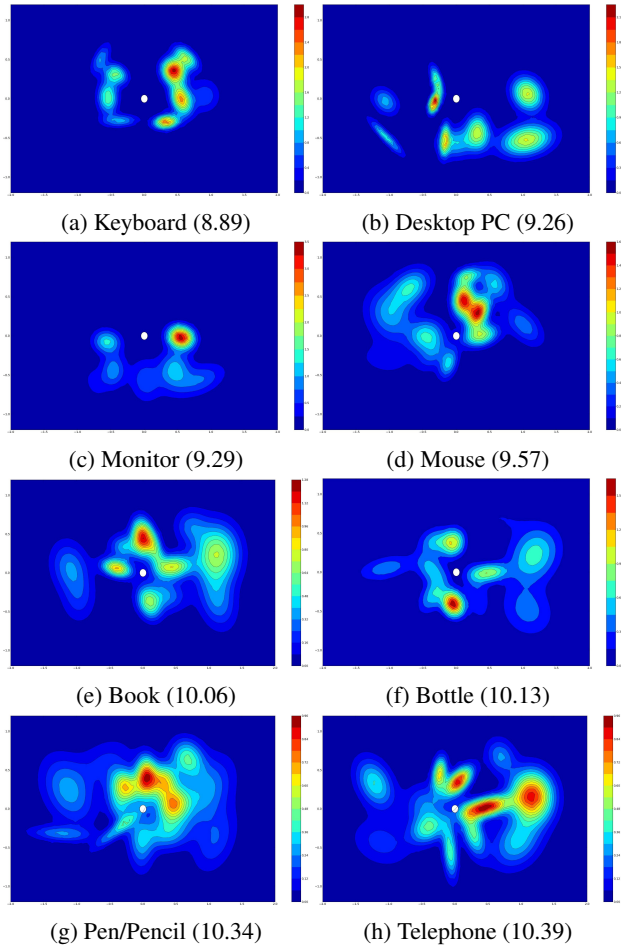
The very basic idea of our approach is to learn a probability distribution in form of a GMM that predicts the position of an object ω given a landmark λ:

$$P_{\omega,\lambda}(\mathbf{x}|\theta) = \sum_{i=1}^{m} w_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i), \qquad (1)$$

where $\mathbf{x}$ denotes the relative object position, in terms of x and y coordinates, of object ω with respect to the landmark λ and θ is a set of parameters $\{w_i, \mu_i, \Sigma_i\}$ for m Gaussian distributions each resembling a set of QSRs. Thereby, the QSRs are represented by a set of multivariate Gaussians. Any combination of directional and distance relations yields to a different multivariate Gaussian. Since some of the directional relations such as *left-of* and *in-front-of* can be combined, a GMM can maximally be represented by 16 multivariate Gaussians (8 directional $\times$ 2 distance).

If we were only given the relative metric information between an object and a landmark, we would have to employ an unsupervised learning approach such as, for example, the Expectation Maximization (EM) algorithm. By using the EM algorithm we could learn a predictive model for object positions. But since we have labelled data, we can directly learn or statistically derive the parameters $\{w_i, \mu_i, \Sigma_i\}$ of the individual Gaussians. The weight w_i of each Gaussian is determined by dividing the number of samples for a particular QSR by the total number of samples. The sum of the weights w_i (for $i = 1, \ldots, m$) is always equal to one. Similarly, we determine the parameters for μ_i and Σ_i on basis of metric object positions for a particular set of QSRs, e.g.:

in-front-of(Mouse, Monitor) $\wedge$
right-of(Mouse, Monitor) $\wedge$



(a) Keyboard (8.89) (b) Desktop PC (9.26)

(c) Monitor (9.29) (d) Mouse (9.57)

(e) Book (10.06) (f) Bottle (10.13)

(g) Pen/Pencil (10.34) (h) Telephone (10.39)

Figure 4: Learnt Gaussian Mixture Models of different landmarks for predicting the relative position of a cup. The entropy of the respective distributions is given in parenthesis.

close-to(Mouse, Monitor).

An advantage of the supervised learning approach is also that GMMs can be pruned on the basis of QSRs. For example, consider the two natural language instructions:

1. get me a cup from my desk, and
2. get me the cup right of the keyboard.

Whereas in the fist case the complete GMM is used, only a partial or pruned GMM is used in the latter case.

Figure 4 shows learnt GMMs for different landmarks for predicting the relative position of a cup.

Choosing a landmark

In the previous section we have explained how we learn GMMs for predicting the position of an object with respect to a landmark. What characterizes a good landmark, is an interesting question we are addressing in this section.

In order to choose a landmark in the object search task we introduce an entropy-based measure that allows us to differentiate between landmarks. We calculate the entropy of a mixture model for an object ω and a landmark λ as follows:

$$H_{\omega,\lambda}(X) = -\sum_{i=1}^{n} P_{\omega,\lambda}(x_i) \log_b P_{\omega,\lambda}(x_i) \qquad (2)$$

where $P_{\omega,\lambda}(X)$ denotes the probability distribution of the respective GMM. To calculate $P_{\omega,\lambda}(X)$, we discretise the metric space in a region around the landmark into a grid X, with n grid cells itx_i. The region is twice as large as the table (4.0 m $\times 2.4$ m), as the landmark could theoretically be located at the very edge of the table. We have chosen a grid cell size of 0.05 meters for each dimension as this resembles the voxel size of the 3D occupancy map later used in the robot experiments. The probabilities for each cell are calculated by dividing the value of the GMM by the sum of all values in the discretised region. With this distribution we can calculate the probability to find an object at a particular grid cell, or voxel.

Given the entropy measure above, we can find out that a specific landmark is a really good predictor for the position of an object. However, it might be the case that this landmark almost never co-occurs with the object. Therefore, we introduce a weight to account for this problem. Instead of minimising the entropy, we try to maximize the following expression:

$$\underset{\lambda \in \Lambda}{\operatorname{argmax}}\, w_{\omega,\lambda} \frac{1}{H_{\omega,\lambda}(X)} \qquad (3)$$

where $w_{\omega,\lambda}$ is determined by the conditional probability that a landmark is present given the object: $P(\lambda|\omega)$. Eventually, we compute a score for an object-landmark pair (ω, λ) as follows:

$$\text{score}_{\omega,\lambda} = P(\lambda|\omega) \frac{1}{H_{\omega,\lambda}(X)}. \qquad (4)$$

Table 4 compares different landmarks for predicting the relative position of a cup and shows the calculated measures for entropy $H_{\omega,\lambda}(X)$, the conditional probability $P(\lambda|\omega)$ and $\text{score}_{\omega,\lambda}$. As the computed $\text{score}_{\omega,\lambda}$ was best for the *Monitor*, we used it as landmark in our experiments.

Table 4: Scoring landmarks for a *Cup* using a weighted entropy-based measure.

Landmark	$H_{\omega,\lambda}(X)$	$P(\lambda\|\omega)$	$\text{score}_{\omega,\lambda}$
Monitor	9.29	0.99	0.106
Keyboard	8.89	0.85	0.096
Mouse	9.57	0.84	0.087
Telephone	10.39	0.63	0.060
Pen/Pencil	10.34	0.61	0.059
Book	10.06	0.56	0.055
Bottle	10.13	0.43	0.043
Desktop PC	9.26	0.32	0.035

6 Experiments

We conducted simulated experiments to evaluate the learnt QSR models in object search tasks. We used the open source

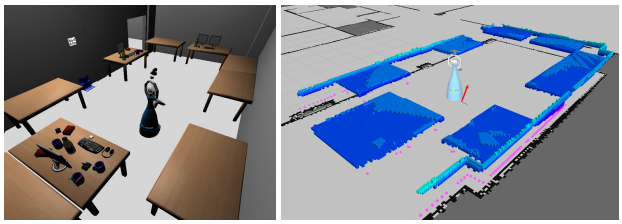


Figure 5: Left: *IRlab* environment with objects on three desks. Right: 3D occupancy grid map of the supporting planes.

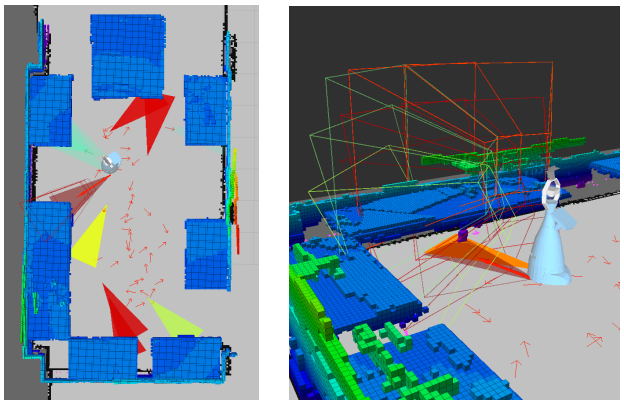


Figure 6: Left: 2D pose evaluation with a *narrow* angle of view. Right: 3D view cone evaluation with a *wide* angle of view.

robot simulator MORSE (Echeverria et al. 2011) for simulating the *IRlab*[1] environment, the SCITOS G5 robot platform[2] and its sensors. In simulation, we used a "semantic camera" to perceive objects in the environment. The semantic camera returns an object ID, the object's type, and its pose whenever an object is in sight and between the near and far plane of the cameras view frustum. In the experiments, the robot was controlled through the task-level architecture SMACH[3] and the middleware ROS[4]. The robot control program is comprised of four states: a search monitor, a particular search method, a navigation routine, and a perception routine. The search monitor assesses the overall progress of the search, i.e., whether an object was found or not and/or whether a timeout has occurred (here: 240 seconds). On this basis it decides to continue or to abort the search task. If it decides to continue the search, the search method selects the next best view pose and the navigation routine moves the robot to the goal accordingly. At the goal location the perception routine is called, the best 3D views are taken using the pan-tilt unit and the result is interpreted by the search monitor and so on.

[1]Intelligent Robotics lab, University of Birmingham, UK
[2]`http://metralabs.com`
[3]`http://wiki.ros.org/smach`
[4]`http://wiki.ros.org`

Experimental setup

Figure 5 shows the *IRlab* environment used in simulation. Overall we conducted ten searches. To set up different object search scenarios, we sampled for each scenario three office desk configurations from the generated data (Section 4) and assigned them to three random desks (out of eight). That is, we varied the number and types of objects, their spatial relationships among each other and their position in the lab environment throughout the scenarios. The task for the robot was to find a cup. Table 5 shows how many cups were present in each scenario and how they were distributed over the different workplaces. That is, in the scenarios in which the cups were distributed over less than three desks, the QSR models indicated locations at which no cups were present.

Table 5: Distribution of cups in the evaluated scenarios.

	Scenario No									
	1	2	3	4	5	6	7	8	9	10
Number of cups	3	4	2	4	2	4	2	3	1	3
Desks with cups	2	2	2	2	2	3	2	2	1	3

In the experiments we compared two search methods:

Supporting planes Within the supporting planes method 30 locations are sampled from the 2D map and evaluated with respect to the projected 3D occupancy map of voxels that had been classified as supporting planes.

QSR Within the QSR-based method also 30 locations are sampled from the 2D map and evaluated with respect to the supporting-plane-voxels of the 3D occupancy map weighted according to the QSR-based mixture of Gaussians.

We tested each of these search methods with different configurations of the semantic camera, called *wide* and *narrow*. In the *wide* configuration we set the camera's angle of view to 58°; resembling the specification of the Asus Xtion PRO LIVE camera mounted on the real robot. In the *narrow* configuration we restricted the angle of view to 29°, as we assume that objects can be better recognized if they are in the center of the field of view. Please note, that the camera configuration was applied for both the actual sensing as well as the evaluation of view cones. Figure 6 depicts the evaluation of 2D and 3D view cones in a scenario.

Experimental results

Table 6 summarizes the results of ten searches using different search methods and camera configurations. For each setup and scenario the table reports the number of visited poses, the consumed time, and whether the search was successful or not. It also provides an average of these figures for each search method over the ten trials.

For the *wide* camera configuration both search methods, QSR and supporting planes, have a comparable performance with an average number of 3.3 and 3.5 visited poses respectively. In previous experiments (Kunze and Hawes 2013) we have seen that the QSR-based method outperforms the supporting plane method when the QSR model matches with

reality. But since the QSR method is sometime misled by false information about an object, it performs not significantly better than the method based on supporting planes. In the worst case, the QSR-based method gets stuck in local maxima whereas the supporting plane method explores eventually the whole search space. Therefore, it is an interesting problem to modify the QSR-based search method in a way that it also searches the whole search space in the worst case. A possibility that would allow for a larger search space is to gradually increase the variance of the learnt models. Alternatively, one could completely remove a GMM model in the view cone evaluation step after searching locations related to a specific landmark. However, a solution to this problem is beyond the scope of this paper.

For the *narrow* camera configuration the QSR method performed better than the supporting plane method. This can be explained by the fact of a smaller view cone. The supporting plane method did not find so many objects in the first place because it would need more time to explore the space. On the contrary, the QSR method could improve its performance. Since the view cone evaluation is more focused on the regions directly influenced by the QSR models the robot finds the cups in less time.

7 Conclusions

In this paper we presented a probabilistic approach to indirect object search using Gaussian Mixture Models to project Qualitative Spatial Representations into 3D space. This approach produces spatially-situated probability distributions which encode the likelihood of a target object being present relative to a given landmark. Our analysis of a synthetic data set has demonstrated that the relative position produces more accurate predictions compared to absolute position, assuming an appropriate landmark can be found. As an initial step towards landmark selection we also presented a scoring mechanism which favours the use of landmarks which are present with higher likelihood in all scenes and which produce lower-entropy predictions in these scenes. The use of QSR-based indirect search is supported by an empirical investigation which shows that our approach improves performance as the robot's field of view decreases, i.e. when more accurate predictions of object location are required.

Whilst our experimental results validate our overall approach, there are a number of important steps we must take to extend this work. First we will replace the synthetic data with 3D object data captured from real desktops at regular intervals. This will both refine our object location predictions with respect to the collected data, allow us to check our assumptions about static landmark objects, and also allow us to reason about qualitative object locations *over time* as well as over space. This will become increasingly important as our robots run for longer periods. Our second extension will be to add planning to this framework. Choosing the order of view cones to visit in a single room, or to choose which rooms to visit in a larger building is an important problem tackled by existing work (Aydemir et al. 2013; Hanheide et al. 2011), but currently ignored in ours. An interesting extension here is including the choice of landmark objects in the planning process, factoring in both landmark location consistency over time and the predictive power of the associated QSRs. We will add the ability to interpret natural language commands into our system, allowing humans to give object location descriptions, and allowing the robot to report its progress ("I looked in front of the monitor but it wasn't there") plus scene descriptions. To ground human language into spatial scenes we will need to replace the automated annotation step in our bootstrapping process with annotation by human subjects. Finally we must tackle the tension between the use of human-understandable QSRs (as used in this paper) and the kinds of QSRs that may emerge from the data with unsupervised learning methods such as clustering. It is possible that the former type of QSRs add in unnecessary distinctions compared to those present in the data (i.e. those which are closely correlated with object function), but this may be a price worth paying in order to support natural language interactions.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 600623, STRANDS, and the EPSRC grant EP/K014293/1.

References

Aydemir, A., and Jensfelt, P. 2012. Exploiting and modeling local 3d structure for predicting object locations. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

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

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

Behera, A.; Cohn, A. G.; and Hogg, D. C. 2012. Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In *Advances in Multimedia Modeling*. Springer. 196–209.

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, 2072–2077.

Burbridge, C., and Dearden, R. 2012. Learning the geometric meaning of symbolic abstractions for manipulation planning. In *Proceedings of Towards Autonomous Robotic Systems (TAROS)*, number 7429, 220–231. Springer.

Cohn, A. G., and Hazarika, S. M. 2001. Qualitative spatial representation and reasoning: an overview. *Fundam. Inf.* 46(1-2):1–29.

Table 6: Object search results for different sensor configurations and search methods.

Camera angle of view	Search method	Metric	Scenario No 1	2	3	4	5	6	7	8	9	10	Scenario average
wide (58°)	QSR	poses (number)	1	2	5	3	8	2	5	2	3	2	**3.3**
		time (sec)	27	72	157	92	253	58	156	50	92	60	**101.7**
		object found	1	1	1	1	0	1	1	1	1	1	**0.9**
	supporting planes	poses (number)	1	2	3	2	1	2	8	2	8	6	**3.5**
		time (sec)	24	50	103	56	24	53	212	43	246	164	**97.5**
		object found	1	1	1	1	1	1	1	1	0	1	**0.9**
narrow (29°)	QSR	poses (number)	2	1	2	1	2	1	2	1	9	1	**2.2**
		time (sec)	70	30	53	29	57	23	55	51	251	24	**64.3**
		object found	1	1	1	1	1	1	1	1	0	1	**0.9**
	supporting planes	poses (number)	2	8	10	4	9	5	9	1	8	9	**6.5**
		time (sec)	55	250	253	114	265	168	253	31	258	256	**190.3**
		object found	1	0	0	1	0	1	0	1	0	0	**0.4**

Echeverria, G.; Lassabe, N.; Degroote, A.; and Lemaignan, S. 2011. Modular Open Robots Simulation Engine: MORSE. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation*.

Hanheide, M.; Gretton, C.; Dearden, R.; Hawes, N.; Wyatt, J.; Pronobis, A.; Aydemir, A.; Göbelbecker, M.; and Zender, H. 2011. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11)*, 2442–2449.

Holz, D.; Iocchi, L.; and van der Zant, T. 2013. Benchmarking intelligent service robots through scientific competitions: The RoboCup@Home approach.

Joho, D.; Senk, M.; and Burgard, W. 2011. Learning search heuristics for finding objects in structured environments. *Robotics and Autonomous Systems* 59(5):319–328.

Kollar, T., and Roy, N. 2009. Utilizing object-object and object-scene context when planning to find things. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, 2168–2173. IEEE.

Kunze, L., and Hawes, N. 2013. Indirect object search based on qualitative spatial relations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on AI-based Robotics*.

Kunze, L.; Beetz, M.; Saito, M.; Azuma, H.; Okada, K.; and Inaba, M. 2012. Searching objects in large-scale indoor environments: A decision-thereotic approach. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Moratz, R.; Nebel, B.; and Freksa, C. 2003. Qualitative spatial reasoning about relative position. *Spatial cognition III* 1034–1034.

Samadi, M.; Kollar, T.; and Veloso, M. M. 2012. Using the web to interactively learn to find objects. In Hoffmann, J., and Selman, B., eds., *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press.

Schuster, M.; Jain, D.; Tenorth, M.; and Beetz, M. 2012. Learning organizational principles in human environments. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Sjoo, K., and Jensfelt, P. 2011. Learning spatial relations from functional simulation. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, 1513–1519.

Sjöö, K.; Aydemir, A.; and Jensfelt, P. 2012. Topological spatial relations for active visual search. *Robotics and Autonomous Systems*. To appear.

Sridhar, M. 2010. Unsupervised Learning of Event and Object Classes from Video by. (December).

Tsotsos, J. K. 1992. On the relative complexity of active vs. passive visual search. *Int. J. Comput. Vision* 7(2):127–141.

Williams, T.; Cantrell, R.; Briggs, G.; Schermerhorn, P.; and Scheutz, M. 2013. Grounding natural language references to unvisited and hypothetical locations. In *Proceedings of Twenty-Seventh AAAI Conference on Artificial Intelligence*, 947–953.

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