Automated Spatial-Semantic Modeling with Applications to Place Labeling and Informed Search

Pooja Viswanathan   David Meger   Tristram Southey   James J. Little   Alan Mackworth

University of British Columbia
Laboratory for Computational Intelligence
201 - 2366 Main Mall, Vancouver, BC, Canada
{poojav,dpmeger,tristram,little,mack}@cs.ubc.ca

Abstract

This paper presents a spatial-semantic modeling system featuring automated learning of object-place relations from an online annotated database, and the application of these relations to a variety of real-world tasks. The system is able to label novel scenes with place information, as we demonstrate on test scenes drawn from the same source as our training set. We have designed our system for future enhancement of a robot platform that performs state-of-the-art object recognition and creates object maps of realistic environments. In this context, we demonstrate the use of spatial-semantic information to perform clustering and place labeling of object maps obtained from real homes. This place information is fed back into the robot system to inform an object search planner about likely locations of a query object. As a whole, this system represents a new level in spatial reasoning and semantic understanding for a physical platform.

1. Introduction

As the proportion of older adults in our population continues to grow, there is an increasing burden on the healthcare system to monitor and care for the elderly. Robotic assistants have been proposed to help older adults perform daily tasks, thus enabling them to live in their own homes, and increasing overall quality of life and independence. In order to perform successfully, however, robots need to understand the world through the same language as their human co-habitants. For example, an intelligent wheelchair needs to know the location of the kitchen in order to guide its cognitively-impaired driver at lunch time. In addition, the com-
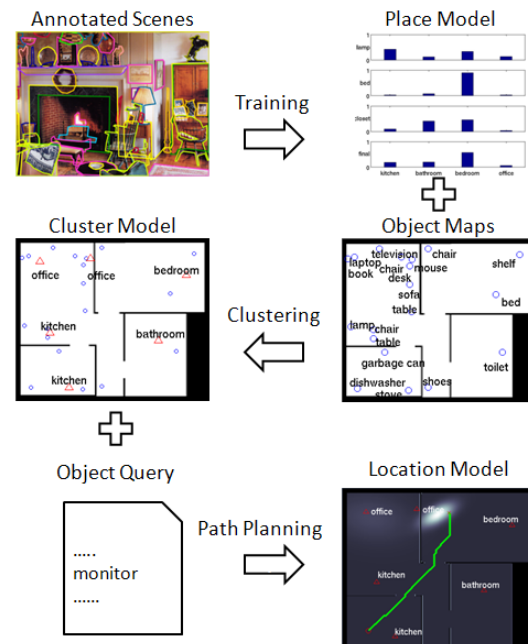

Figure 1. A flowchart of our system's capabilities. The Spatial-Semantic Model enables numerous useful tasks.

mand "Robot, find the computer!" requires the robot to understand the term "computer" and to know that computers are often located in offices.

Figure 1 provides an overview our system's ability to automatically obtain the spatial-semantic model necessary to plan a path in response to the user's command. As seen in the figure, the spatial-semantic model has the following components: *Place Model*, *Cluster Model*, and *Location Model*. We define a place as an area, consisting of one or more objects, that is used to perform

a set of related tasks. Our system models kitchens, bathrooms, bedrooms, and offices, while other example places might include lounges, libraries, and laundry rooms. The Place Model employs semantic information about objects (places they usually occur in) to determine their corresponding place labels. The Cluster Model uses spatial and semantic information about objects (places they usually occur in and their observed locations) to determine their cluster and place labels. We define a location as a 2D coordinate on a map, that may or may not contain an object. Finally, the Location Model determines the likely locations of objects by exploiting their semantic information as well as information about spatial-semantic clusters on the map.

While several existing systems have demonstrated similar modeling capabilities [9, 4, 16, 10], they have often required extensive engineering efforts or access to specialized data sources. We seek to provide a robotic system with the ability to gather such experience, or at least to obtain it in an extremely automated and scalable fashion, without extensive effort from a system designer.

To this end, we have utilized the information present in the LabelMe [11] database: a free online datasource which provides a large and growing amount of human-labeled visual data, much of which contains indoor scenes suitable for place labeling and object recognition. Each LabelMe scene provides an image annotated with a variety of semantic information such as the place type (e.g. kitchen, bathroom, bedroom), segmentations of unique objects in images in the form of polygons, and semantically meaningful object labels (see the top left image of figure 1). Our system interfaces with the semantic information present in LabelMe to construct Place Models. We do not directly analyze the images in the dataset, and instead focus on the textual annotations for each image.

Our work is motivated by our previous efforts to develop a platform for embodied visual search known as Curious George [9]. This system has demonstrated a world-class ability to find query objects in indoor environments by winning the robot league of the Semantic Robot Vision Challenge (SRVC) [2], an international competition to evaluate embodied recognition systems, for two consecutive years. The spatial-semantic modeling methods in this paper are designed to enhance Curious George. Thus, throughout this paper, we assume the existence of a robot capable of performing successful object recognition in the real world, and do not discuss the object recognition problem directly. For example, we use manually constructed home floorplans annotated with object locations as system input, since

Curious George is able to produce such maps.

In section 3.2 we describe our Place Model and demonstrate classifications of LabelMe scenes. In section 3.3, we describe our Cluster Model and apply it to the task of identifying places within real homes based on object locations. In section 3.4, we demonstrate the use of the Location Model and other components of the spatial-semantic model to improve object search in a simulated environment. Experiments and results are discussed in section 4. We conclude with future directions for this research.

2. Related Work

In the robotics community, topological mapping has long been seen as an alternative to the precise geometric maps that are produced by Simultaneous Localization and Mapping (SLAM). Topological maps are meant to encode "places" and their connections, although the level of semantic meaning available about each place is often quite limited. Kuipers [6] proposes the *Spatial Semantic Hierarchy* which represents space at a number of levels, each with varying spatial resolution and detail. Numerous authors have explored the properties of grid-like topological maps in terms of theoretical ability to localize [1], practicality of space decomposition [8], and as an integrated mapping and localization system [7]. Kröse *et al.* have developed a series of practical systems [14, 5] in which the visual similarity between images is used to cluster regions in the environment. Place labels for the clusters, however, are provided by a human through speech.

Several robotic systems have attempted to produce environment representations containing human-like semantic information. Shubina and Tsotsos [12] consider the active vision problem of optimizing robot paths while performing object recognition. Torralba *et al.* [15] describe specific room and room type classification of images collected by a wearable testbed, and subsequently use room information as a prior for object recognition. This work has in part inspired our Location Model as presented in section 3.4. The Stair [4] robot has demonstrated the ability to attend to, recognize, and grasp objects in an environment, though it has not, to our knowledge, been demonstrated to build large scale semantic maps. Sjo *et al.* [13] describe an embodied object recognition system which is very similar to our own in hardware and overall goals. To our knowledge, these authors have not yet incorporated place labeling and object-location priors into their systems, and most certainly have not considered automated extraction of semantic information from online data sources.

Our work is directly inspired by Vasudevan *et al.* [16] who have considered labeling places and functional regions of the environment based on object occurrences. Their system demonstrates impressive performance in realistic environments, but uses manually labeled training images while we access an ever-growing online database containing thousands of types of objects. Our work further improves upon their method by employing an iterative clustering scheme that is more robust to initially incorrect clusters, and exploiting place information in informed search.

3. Methods and Models

This section provides formalisms for the representations and methods used to create our spatial-semantic model, which is subsequently applied to the problems of place classification, clustering and labeling object maps, and informed object search. Detailed results for each of these applications are provided in the Experiments section.
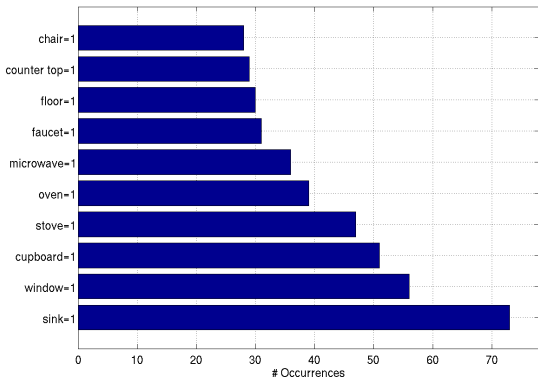

Figure 2. The histogram representation of the Count Model for the place "kitchen". Please note that only the 10 most frequently occurring object-count pairs are displayed, for clarity.

3.1. LabelMe Data Extraction

In order to build a spatial-semantic model, we first need to learn a model of objects and their occurrence frequencies in each place type. We can obtain this information from the LabelMe database by querying scenes and recording the number of annotated occurrences of each object in the scene, as in Vasudevan *et*

al. [16]. For example, we separately model the frequency of exactly 1 chair, exactly 2 chairs, etc., occurring in a kitchen. Figure 2 shows an example of the ten most frequently occurring object-count pairs in annotated kitchen images from LabelMe. The counts table $ct_p(o, c)$ contains the number of times object o occurs c times in images of place type p.

If the number of images of place type p is n_p, the likelihood of observing c occurrences of object o in place p is computed as:

$$P(o, c|p) = \frac{ct_p(o, c)}{n_p} \qquad (1)$$

This probability is smoothed over nearby counts with Gaussian noise to account for sparse training data. That is, we add a small probability to counts $c-1$ and $c+1$ for each occurrence of count c. We refer to this likelihood as the Count Model, which is used to build the different components of the spatial-semantic model.

3.2. Place Model

Given a Count Model, the Place Model is used to predict the most likely place type of the observed objects. The prior probability for each place type p is the proportion of training examples with label p, as follows:

$$P(p) = \frac{n_p}{\Sigma_i n_i} \qquad (2)$$

We can compute the posterior probability of the place type p given an object o and its count c using Bayes' theorem:

$$P(p|o, c) \quad = \quad \frac{P(o, c|p)P(p)}{\Sigma_i P(o, c|p_i)P(p_i)} \qquad (3)$$

We add additional noise to this posterior for each place type, inversely related to the proportion of training examples of the place type. This allows inference to handle the occurrence of previously unseen objects in places for which we have little or no training data.

In order to predict the place type given an image, we need to combine the possibly conflicting predictions of all objects present in the image. We refer to this problem as Place Classification and propose the following scheme to obtain a solution:

1. For each image, generate a set of object-count pairs (o, c) based on each object type o found in the test image, and number of times it appears in the image, c.
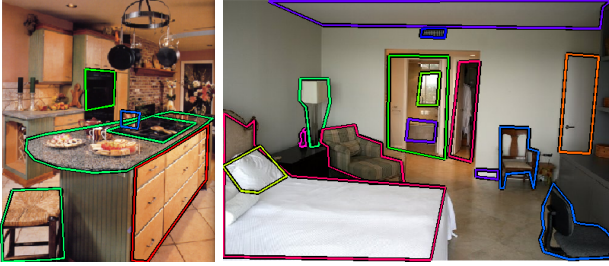
Figure 3. Sample place labeling results. The left image was correctly labeled as a kitchen, while the right image is a bedroom that was incorrectly assigned to place type "unknown". The kitchen image is annotated with the following objects: drawer, oven, pot, stool, stove, table top. The bedroom image is annotated with: bathtub, armchair, bed, ceiling, chair, door, lamp, molding, phone, pillow, vent, wardrobe and window.

2. Compute a hypothesis probability distribution over place types p, $h_{oc}(p) = P(p|o,c)$ using equation 3 for every (o,c) pair.

3. Compute the final hypothesis probability $h(p) = \Sigma_{oc}(h_{oc}(p))$. This is equivalent to allowing every object-count pair (o,c) observed in the test image to contribute a vote for each place type p weighted by its probability $h_{oc}(p)$.

4. If the entropy of the resulting hypothesis is lower than a pre-specified threshold (1.36 in this paper), the predicted label for the image is the place type with the highest weighted sum of votes, i.e. $argmax_p h(p)$. Otherwise, the image is labeled with "unknown" place type.

Figure 3 shows some correct and incorrect results of place classification. The kitchen scene contains several objects that are commonly seen in kitchens, thus producing a strong prediction for that place type. The bedroom scene, however, contains labels that are common to both bathroom and bedroom scenes. It is interesting to note that this ambiguity can be resolved if the model is able to identify that the image contains two groups of objects that are spatially separated, thus requiring two different labels. This example highlights that spatial information is crucial for accurate place labeling, and is therefore included in the Cluster Model described in the next sub-section.

3.3. Cluster Model

As a robot explores its environment and recognizes objects, the Cluster Model can be employed to group objects based on their place types and spatial locations. Labeling places in realistic home environments is challenging due to object clutter and high variation of floorplan layouts. Thus, combined use of spatial and semantic information leads to more meaningful place maps. Construction of place maps involves two steps:

1. Clustering of objects to form places based on both spatial and semantic information

2. Labeling each cluster with its most likely place type

Step 2 is equivalent to the Place Classification method described in the previous sub-section if each cluster is treated as an image. Thus, we focus on the clustering algorithm here. We use a K-means algorithm for clustering, and use the Bayesian Information Criterion (BIC) to select the optimal number of clusters (between 4 and 8). Each cluster contains a centroid (mean of object coordinates) and a hypothesis place probability distribution (computed using steps 1-3 of Place Classification). Assignments are chosen to minimize a weighted linear combination of spatial distance and place probability distribution dissimilarity, with free parameter α. Note that uniform place type priors are used to compute the hypothesis place probabilities during clustering instead of equation 2, since we expect to see all place types with equal probability on average in maps of single floor homes.

For each k:

1. Initialize cluster centres to k objects randomly selected from distinct regions in the map.

2. Assign each object to a cluster, attempting to minimize both its distance to the centroid, and the L1-Norm between the hypothesis place probability distributions of the object and cluster.

3. Recompute the cluster centroids and hypothesis place probability distributions.

4. Repeat steps 2 and 3 until cluster centres stabilize.

5. Repeat steps 1 - 4 50 times to generate different initial cluster centres, and select the final configuration with the least amount of intra-cluster variance.
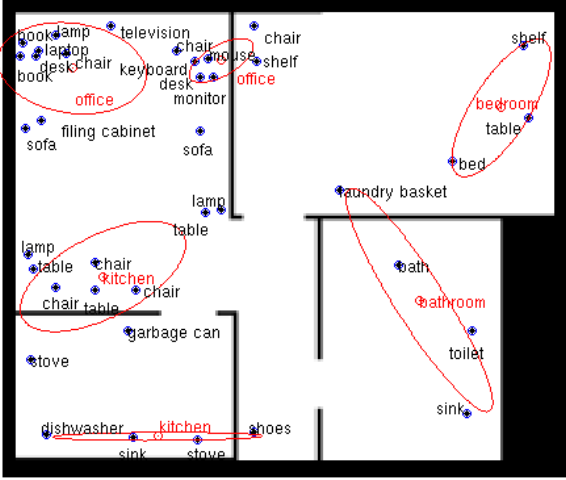
Figure 4. A sample clustering result.

As mentioned, we employ the BIC score to choose the number of clusters. BIC indicates that the best value for k is one that maximizes the log-likelihood of the data and minimizes model complexity (i.e. the number of free parameters). Cluster labels are assigned according to step 4 of Place Classification. An example of the result of this procedure can be seen in figure 4. The method accurately groups objects that are both spatially close to one another and semantically similar (typically found in the same place). More results for clustering and place labeling can be found in the Experiments section.

Place labels can be extremely useful for navigation assistance and search systems. For example, Viswananthan et al. [17] describe an intelligent wheelchair system, called NOAH, to assist cognitively-impaired users. The system requires annotated maps in order to guide the user effectively to his/her desired destination. Automated place labeling as described in this section, using recognized objects and their locations, will allow NOAH to identify different places in its environment without the need for manual input, thus allowing it to adapt to new environments automatically. Place labels can also be used to inform visual search for objects, as described next.

3.4. Location Model

A robot performing object search requires a sequential decision-making planner to determine locations for the robot, and viewing directions for the camera, at each instant. As mentioned previously, we are motivated by augmenting the planner developed previously in [9] with spatial-semantic information to improve object search performance. This planner attempts to attain coverage of an environment, to identify potential objects of interest in a low-resolution peripheral camera, and to obtain numerous viewpoints of each of these objects with a high-resolution foveal camera. We will modify the space coverage portion of this planner by constructing an object-specific prior over spatial locations that guides the robot's search.
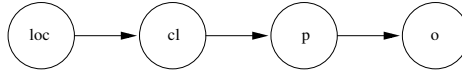

Figure 5. The independence relations between locations, clusters, places and objects present in our system can be encoded as a probabilistic graphical model.

Our Location Model combines the Place Model and Cluster Model to compute a likelihood function for the location of each object type. This likelihood is computed using the conditional independencies between variables described by the graphical model shown in figure 5:

$$P(o|loc) \;=\; \Sigma_p P(o|p)\Sigma_{cl}P(p|cl)P(cl|loc) \quad (4)$$

We marginalize equation 2 over object counts c to compute $P(o|p) = \Sigma_c P(o,c|p)$, since we are interested in finding any number of occurrences of the query object. We have determined through experimentation that selecting the most likely place label for each cluster gives reliable performance due to the high accuracy of cluster and place labels. Thus, we replace $P(p|cl)$ with a hard assignment:

$$P(o|loc) \;=\; \Sigma_p P(o|p)\Sigma_{cl}\left[\theta(p,cl)\ P(cl|loc)\right] \quad (5)$$

Where $\theta(p,cl)$ is an indicator variable that is 1 if cluster cl is assigned the label p, and 0 otherwise. Each cluster cl is represented by a Gaussian distribution over locations loc with mean μ_{cl} and covariance σ_{cl}. The likelihood $P(o|loc)$ acts as a prior model for the occurrence of an unseen object at any location in the world, and provides a direct source of guidance during the search task. In the remainder of this paper, we will refer to equation 5 as the object-location prior.

Figure 6 displays the Location Model and a sample path. The object prior indicates the potential for towels to occur in either of the bathroom or kitchen,
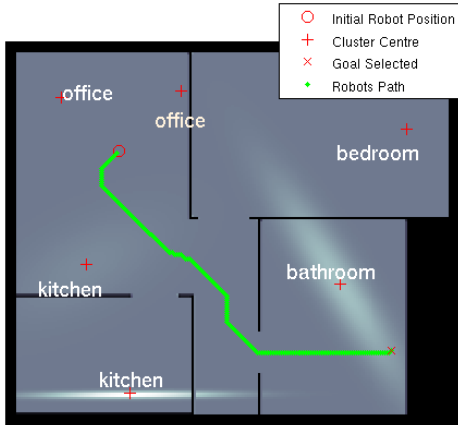
Figure 6. A sample path generated by the informed search planner for previously unseen query "towel". Background colouring demonstrates location prior density at each location.

and the random sampling procedure chooses a search location within the bathroom. The reader is reminded that the robot has never observed a towel in the testing environment but can identify likely locations for one using learned object-place associations and place clusters.
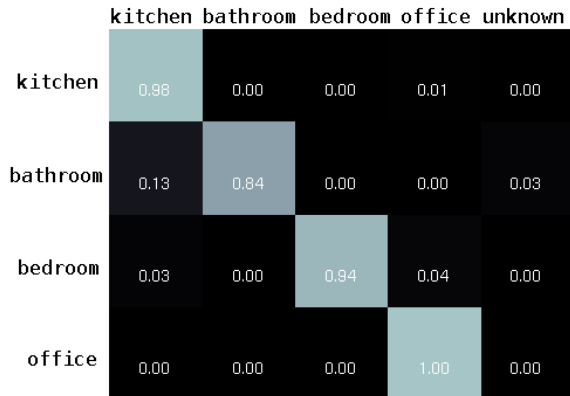
4. Experiments
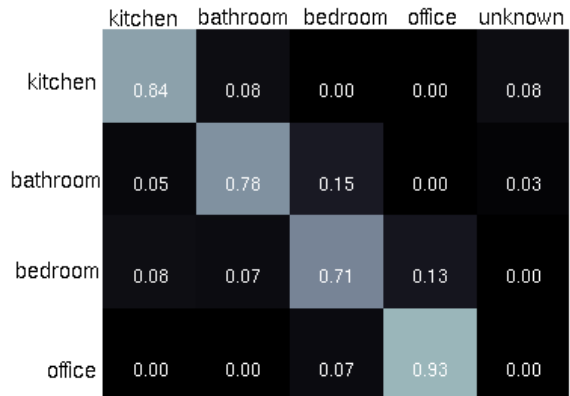
4.1. Place Classification

As an initial validation of our system's ability to model place type from object presence, we have attempted to infer place type based on the objects annotated in a LabelMe image. Specifically, we split the scene images into non-overlapping training and test sets, and for each test image we compute the most likely place type conditioned on the object annotations (i.e. we do not look at the pixels of the image, only the text labels accompanying it). Query results are obtained for four scene types: Kitchen (176), Bedroom

Room Type	Precision	Recall
Kitchen	0.97	0.98
Bathroom	1.00	0.84
Bedroom	0.97	0.94
Office	1.00	1.00

Table 1. Average Precision and Recall Rates

Figure 7. Confusion matrices for: (top) place labeling evaluated on LabelMe images and (bottom) assignment of objects to places evaluated on our realistic floorplans. Rows indicate the ground truth image/objects label and columns indicate our system's predictions.

(37), Bathroom (31), Office (824). We filter omni-directional images and those without annotations.

Table 1 shows average recall and precision rates over 50 trials, while figure 7(a) shows the average confusion matrix for 50 trials. As seen in the results, precision and recall rates are quite high for all place types. Lowest recall rates are observed for bathroom and bedroom place types due to the fewer number of training examples (resulting in lower place priors) and fewer annotations per image available (resulting in insufficient information to make the correct prediction) for these place types. We expect to see higher recall rates as more annotations become available for bathroom and bedroom scenes.

4.2. Clustering and Place Labeling

We constructed a dataset of 7 home models based on real home environments in order to evaluate the clustering and labeling method. This data consists of a selection of layouts (studio apartment, regular apartment, bungalow, etc.) of varying complexity along with the locations and labels of some of the contained objects (see figure 4 for an example). The objects are grouped into clusters and each cluster is assigned a place label and centroid location. Figure 7(b) shows the confusion matrix for the final place labeling of object clusters. As seen, our clustering algorithm produces assignments that closely match the ground-truth place labels.
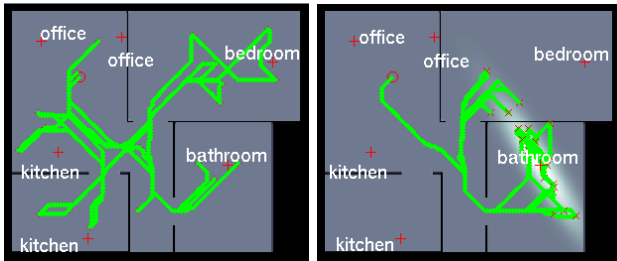
4.3. Informed Object Search

Figure 8. 20 step paths of the two proposed planning methods for the query "shampoo": (left) uninformed coverage vs (right) planning based on place labeling information.

The informed planning procedure described earlier is evaluated with a realistic robot simulator developed during preparation for the SRVC competition. This simulator, based partly on Player Stage [3], allows evaluation of the robot behaviour resulting from a planned path within the home layouts described previously. The first tests performed are qualitative evaluations of the areas visited by the robot using the informed planner, compared with the traditional method based only on coverage of the environment. Figure 8 shows a typical result in which the robot's paths focus more directly upon bathrooms, which are likely to contain the query object, "shampoo", when using informed search compared to a planner based only on coverage of the environment.

We have also performed a quantitative comparison of the informed search method by simulating the robot's camera and recording the frequency with which planned paths capture a view of the query object, as
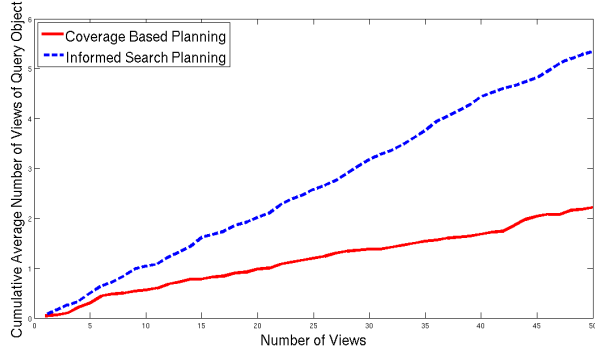

Figure 9. Average number of views of the query object captured by the robot per planning step.

shown in figure 9. This comparison is averaged over 50 trials of 50 planning steps each, for 2500 total robot poses. Between each trial, an initial robot location and query object are selected at random and each of the two planning methods is evaluated. The results demonstrate the fact that the informed search planner is able to obtain views of the query object more quickly than a coverage-based planner and that informed search planning continues to collect additional viewpoints of the object with a higher frequency. These are both desireable behaviours for a robot performing visual object search since few current recognition algorithms achieve viewpoint invariance, so the robot must gather a canonical view of each object to allow recognition. The performance of the informed search planner in simulation strongly suggests that it will facilitate improved object recognition performance on a real platform.

5. Conclusions

We have presented a method for automatic modeling of spatial-semantic information and its application to a variety of tasks. In the future, we plan to extend our models to incorporate place recognition, by recognizing common qualitative spatial relationships that exist between objects, such as distance, orientation and containment. These relationships can help improve object recognition, and allow us to better determine the regional extent of places. In addition, we hope to incorporate geometric obstacle information during the clustering process (e.g. the presence of walls between cluster members) to enforce more realistic clusterings.

We will soon integrate our existing visual search platform, Curious George, with the spatial-semantic model. A major challenge facing integration is successful object category recognition in realistic environ-

ments. We will need to evaluate the effects of recognition errors on clustering accuracy. The integrated system will build object maps that can be used for place labeling, exploit the Place Model as a prior to guide search, and in turn improve object recognition performance. This will make it an exemplar system to inspire application developers in areas such as assistive technology and home robotics. In the longer term, the automatic extraction and use of spatial-semantic information will facilitate large-scale deployment of assistive robots that are capable of reasoning and acting intelligently.

References

[1] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes. Robotic exploration as graph construction. *IEEE Transactions on Robotics and Automation*, 7(6):859 – 86, 1991.

[2] A. Efros and P. Rybski. Website: http://www.semantic-robot-vision-challenge.org/.

[3] B. Gerkey, R. Vaughan, K. Stoy, A. Howard, G. Sukhatme, and M. Mataric. Most valuable player: A robot device server for distributed control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1226–1231, Wailea, Hawaii, 2001.

[4] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarch, S. Chung, and A. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

[5] B. Kröse, O. Booij, and Z. Zivkovic. A geometrically constrained image similarity measure for visual mapping, localization and navigation. In *Proceedings of the 3rd European Conference on Mobile Robots*, pages 168 – 174, Freiburg, Germany, 2007.

[6] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119:191 – 233, 2000.

[7] B. Kuipers and Y. Byun. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Journal of Robototics and Autonomous Systems*, 8:47 – 63, 1991.

[8] B. Lisien, D. Morales, D. Silver, G. Kantor, I. Rekleitis, and H. Choset. The hierarchical atlas. *IEEE Transactions on Robotics*, 21(3):473 – 481, June 2005.

[9] D. Meger, P.-E. Forssn, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems Journal, Special Issue From Sensors to Human Spatial Concepts*, June 2008.

[10] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *Robotics: Science and Systems*, 2007.

[11] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision (special issue on vision and learning)*, 77(1-3):157 – 173, 2008.

[12] K. Shubina and J. K. Tsotsos. Visual search for an object in a 3d environment using a mobile robot. Technical report, CSE Departmental Technical Report, April 2008.

[13] K. Sj, D. G. Lopez, C. Paul, P. Jensfelt, and D. Kragic. Object search and localization for an indoor mobile robot. *Computing and Information Technology (accepted)*, 2007.

[14] T. Spexard, L. Shuyin, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Krose. Biron, where are you? enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proceedings of Intelligent Robots and Systems*, pages 934 – 940, Beijing, 2006.

[15] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.

[16] S. Vasudevan and R. Siegwart. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems*, 56(6):522 – 537, June 2008.

[17] P. Viswanathan, A. Mackworth, J. J. Little, J. Hoey, and A. Mihailidis. Noah for wheelchair users with cognitive impairment: Navigation and obstacle avoidance help. In *Proceedings of AAAI Fall Symposium on AI in Eldercare: New Solutions to Old Problems*, pages 150 – 152, 2008.