Semantic Modeling of Places using Objects

Ananth Ranganathan and Frank Dellaert
College of Computing, Georgia Institute of Technology
{ananth, dellaert}@cc.gatech.edu

***Abstract*— While robot mapping has seen massive strides recently, higher level abstractions in map representation are still not widespread. Maps containing semantic concepts such as objects and labels are essential for many tasks in manmade environments as well as for human-robot interaction and map communication. In keeping with this aim, we present a model for places using objects as the basic unit of representation. Our model is a 3D extension of the constellation object model, popular in computer vision, in which the objects are modeled by their appearance and shape. The 3D location of each object is maintained in a coordinate frame local to the place. The individual object models are learned in a supervised manner using roughly segmented and labeled training images. Stereo range data is used to compute 3D locations of the objects. We use the Swendsen-Wang algorithm, a cluster MCMC method, to solve the correspondence problem between image features and objects during inference. We provide a technique for building panoramic place models from multiple views of a location. An algorithm for place recognition by comparing models is also provided. Results are presented in the form of place models inferred in an indoor environment. We envision the use of our place model as a building block towards a complete object-based semantic mapping system.**

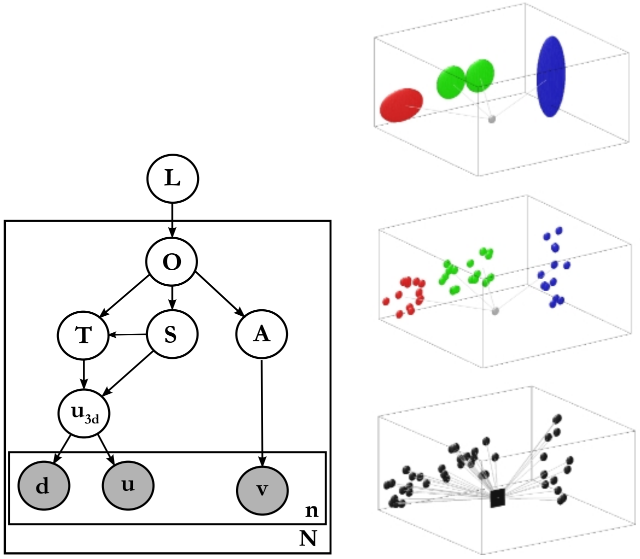

Fig. 1. A generative model for representing places in a mapped environment. The place label L generates a set of N objects O, each having a shape S, an appearance A, and a 3D location T. The objects, transformed to location T , give rise to 3D feature points u_{3d}. These features are observed in an image as n features, each with pixel location u, appearance v, and stereo depth d (shaded variables). The schematic on the right illustrates this process.

I. Introduction

Robot mapping has in recent years reached a significant level of maturity, yet the level of abstraction used in robot-constructed maps has not changed significantly. Simultaneous Localization and Mapping (SLAM) algorithms now have the capability to accurately map relatively large environments [8], [22]. However, grid-based and feature-based maps constructed using lasers or cameras remain the most common form of representation. Yet higher level abstractions and advanced spatial concepts are crucial if robots are to successfully integrate into human environments.

We have chosen objects and their location to be the semantic information in our maps based on the object-centricness of most man-made environments. People tend to associate places with their use or, especially in workspaces, by the functionality provided by objects present there. Common examples are the use of terms such as “printer room”, “room with the coffee machine”, and “computer lab”. Even in outdoor spaces, people often remember locations by distinguishing features that most often turn out to be objects such as store signs and billboards [12]. Thus, objects form a natural unit of representative abstraction for man made spaces. While we do not claim that representing objects captures all the salient information in a place of interest, it is an important dimension that is useful in a wide variety of tasks.

A major concern in constructing maps with objects is object detection, which has been a major area of research in computer vision for a significant period of time. Due to the difficulty in general purpose object detection, many semantic mapping algorithms assume object detection as a black box [5], and so, sidestep a major component of the mapping problem. However, recent advances in stable feature detection algorithms have enabled featurebased object detection methods that have revolutionized the field. In particular, use of SIFT descriptors [11] along with affine invariant features [13] has been particularly popular since binaries for detecting these features are publicly available.

In this paper, we present one of the first instances of semantic mapping in robotics that integrates state-of-the-art object detection techniques from computer vision. We present a 3D generative model for representing places using objects and develop learning and inference algorithms for the construction of these models. The model for a place is constructed using images and depth information obtained from a stereo camera. Our model is a 3D extension of the constellation models popular in the computer vision literature [4]. In particular, as illustrated in Figure 1, a place is represented as a set of objects O with 3D locations T specified in a local coordinate frame.

In turn, an object is modeled as having a particular shape and appearance, and gives rise to features in the image. This generative model is discussed in detail in Section III below.

The models for the objects are learned in a supervised manner, as will be developed in Section IV. The shape models for the objects are learned using stereo range data, while corresponding appearance models are learned from features detected on the images. Training data is provided to the learning algorithm in the form of roughly segmented images from the robot camera in which objects have been labeled. While unsupervised learning is preferable from the standpoint of automation, it requires that the objects to be learned be prominent in the training images, a condition that is not satisfied by our training images. Supervised learning results in more reliable and robust models.

Once the object models have been learned, inference for the place model is performed at runtime using the Swendsen-Wang algorithm, a cluster Markov Chain Monte Carlo (MCMC) technique [1]. Approximate inference through MCMC is required since finding the correspondence between image features and objects is a combinatorial problem. Features in the test image are connected in a Markov Random Field (MRF) which promotes clustering based on appearance and locality in 3D space. Subsequently, we employ the Swendsen-Wang algorithm to perform sampling-based inference on the MRF. Range data and feature appearance provide strong constraints on the distributions resulting in rapid convergence. The details are given in Section V.

Finally, the place models can be used to perform place recognition, for which purpose we provide an algorithm in Section VI. We also describe a technique to build 360^o panoramic models from multiple views of a place using relative robot pose.

Experiments are presented on a robot mounted with a Triclops stereo camera system. We present place models constructed using a fixed object vocabulary in an indoor environment to validate our learning and inference algorithms, in Section VII. We start with related work in the next section.

II. Related Work

The need to include semantic information in robot representations of the environment has long been recognized [10]. A number of instances of functional and semantic representations of space exist in the literature, of which we mention a few recent examples. [18] builds a map of the environment consisting of regions and gateways and augments it to include rectangular objects [17]. Supervised learning of labels on regions is performed in [16] using cascade detectors to detect objects of interest. [5] describes a hierarchical representation that includes objects and semantic labeling of places in a metric map but assumes the identities of objects to be known. [14] performs 3D scene interpretation on range data and computes a map consisting of semantically labeled surfaces. [24] lays out a detailed program for creating cognitive maps with objects as the basic unit of representation. However, their object detection technique uses simple SIFT feature matching which does not scale to larger objects. Our method is more comprehensive than the above-mentioned techniques since it incorporates inference for the 3D location, type, and number of objects in the scene.

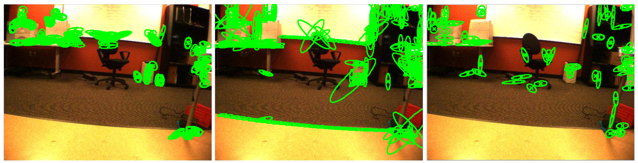

Fig. 2. Features detected on a typical training image using (a) Harris-affine corners (b) Canny edges (c) MSER detector.

More computer vision oriented approaches to the problem of scene modeling also exist. [7] gives a technique for 3D scene analysis and represents the analyzed 3D scene as a semantic net. [23] implements place recognition using features extracted from an image and uses place context to detect objects without modeling their location. More similar to our approach is the one in [20] that creates 3D models of scenes using the Transformed Dirichlet Process. In contrast to this, our approach is simpler and more robust as it uses supervised learning and Swendsen-Wang sampling.

III. Constellation Models for Places

We model places using a 3D extension of the popular constellation model in computer vision. More precisely, we use the "star" modification of the constellation model [4]. The model represents each place as a set of objects with 3D locations in a local coordinate frame. We assume that given the origin of this coordinate frame, hereafter called the *base location*, the objects are conditionally independent of each other. While full constellation models can also model relative locations between objects, the associated increase in complexity is substantial. More discussion on this subject can be found in [4].

A graphical representation of the generative place model is given in Figure 1. The place label L generates a set of objects O, where the number of objects N is variable. Each object gives rise to a set of 3D feature points u_{3d} according to a shape distribution S. Further, the 3D location of the object in local coordinates is represented using the translation variable T, which is unique for each object and also depends on the place label. Finally, the 3D points, transformed in space according to T, give rise to image features at locations u with appearance v distributed according to an object specific distribution A. The 3D points also produce range measurements d, obtained using stereo, corresponding to the observed features.

The shape distribution S and the appearance distribution A taken together model an object class. In our object models, we represent the shape of an object using a Gaussian distribution in 3D, while its appearance is modeled as a multinomial distribution on vector quantized appearance words.

A. Feature detection and representation

For appearance measurements, we use three complementary types of features in our work that are subsequently

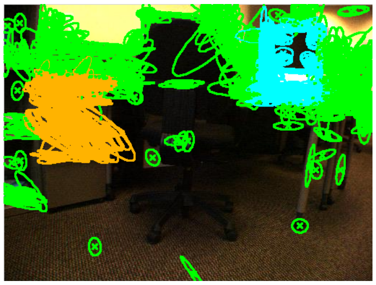

Fig. 3. An example training image with roughly segmented and labeled objects, a monitor in cyan and a drawer in brown.

discretized to facilitate learning. Following previous work [20], we use Harris-affine corners [13], maximally stable extremal regions [13], and clustered edges obtained from a Canny edge detector [2]. We used the public implementations of the first two feature detectors available at *http://www.robots.ox.ac.uk/~vgg/research/affine*. Figure 2 highlights the complementary characteristics of the detected features of each type for a sample image. As noted in [20], edge features are crucial for modeling objects with textureless surfaces such as monitors. The features are represented using SIFT descriptors in a 128 dimensional vector space. We vector quantize the SIFT descriptors using K-means clustering to produce a set of appearance "words". Each feature is subsequently described by the bin corresponding to its closest appearance word, and its 2D pixel location in the image.

IV. Supervised Learning of 3D Object Models

The individual object models in the object vocabulary of the robot are learned in a supervised manner. While unsupervised learning is the sought-after goal in most applications, learning objects in this manner poses many problems. Firstly, the objects of interest have to be displayed prominently in the images with almost no background for reliable unsupervised object discovery to be possible. Further, currently most unsupervised object recognition methods rely on some form of "topic" discovery on the set of training images to learn object models [19]. In images with varied background, the topic discovered by the algorithm may not correspond to objects of interest. Supervised learning sidesteps these issues and is accordingly more suitable for our application.

Training data is presented to the algorithm in the form of roughly segmented stereo images along with the associated stereo range data. Objects in the image are labeled while all unlabeled regions are assumed to be background. An example training image is given in Figure 3.

The shape Gaussian for an object is estimated from the range data labeled as belonging to the object, and the appearance is similarly learned from the features corresponding to the object. Learning the object models is thus straight-forward.

V. Inference for constructing place models

During the testing phase, the robot observes an image along with associated stereo range information, and has to infer the label and model for the place, i.e the types of objects and their 3D locations. We denote the set of appearance and shape models learned during the training phase as $\mathcal{A} = \{A_{1:m}\}$ and $\mathcal{S} = \{S_{1:m}\}$ respectively, where m is the number of objects in the robot's vocabulary. The pixel locations of the features observed in the image are denoted by the set $U = \{u_{1:n_f}\}$, while the corresponding appearance descriptors are written as $V = \{v_{1:n_f}\}$, n_f being the number of features. The depth from stereo corresponding to each of these features is represented as $D = \{d_{1:n_f}\}$. In the interest of brevity, in the following exposition we will compress the set of measurements to $Z = \{U, V, D\}$ whenever possible.

We infer the place model and label in a Bayesian manner by computing the joint posterior distribution on the place label, the types of objects, and their locations. This posterior can be factorized as

$$p(L,O,T|\mathcal{A},\mathcal{S},Z) \quad = \quad p(O,T|\mathcal{A},\mathcal{S},Z)p(L|O,T,\mathcal{A},\mathcal{S},Z) \qquad (1)$$

where L is the place label, O is the set of object types, and T is the corresponding 3D locations. Note that the number of objects at a place, i.e the cardinality of set O, is unknown.

The inference problem can be divided into two parts, namely place *modeling* and place *recognition*, which correspond to the two terms on the right side of (1) respectively. The modeling problem consists of inferring the objects and their locations, while the recognition problem involves finding the label of the place given the objects and their locations. In this section we focus on the modeling problem and return to the recognition problem in the next section.

Since the measurements are in the form of image features, inference cannot proceed without the correspondence between features and the object types they are generated from. We incorporate correspondence by marginalizing over it so that the model posterior of interest can be written as

$$p(O,T|\mathcal{A},\mathcal{S},Z) \quad = \quad \sum_{\Pi} p(O,T|\mathcal{A},\mathcal{S},Z,\Pi)p(\Pi|\mathcal{A},\mathcal{S},Z) \qquad (2)$$

where Π is a discrete correspondence vector that assigns each image feature to an object type. We call (2) the *place posterior*.

Since computing the place posterior analytically is intractable, we employ a sampling-based approximation. The intractability arises from the need to compute the distribution over correspondences, which is a combinatorial problem. One technique for overcoming this intractability is using Monte Carlo EM (MCEM) [21], in which a Monte Carlo estimate of the distribution over correspondences is used to maximize the posterior over the other hidden variables iteratively. In our case, this would involve a maximization over a possibly large discrete space of object types. Further, multi-modal distributions in this space cannot be discovered using MCEM, which only computes the MAP solution. These reasons motivate our use of Markov Chain Monte Carlo (MCMC) methods for computing a sampling-based approximation to the posterior.

To compute the distribution over correspondences, we note that features corresponding to a particular object type occur in clusters in the image. Hence, appearance and stereo depth provide important clues to the correspondence of a feature in

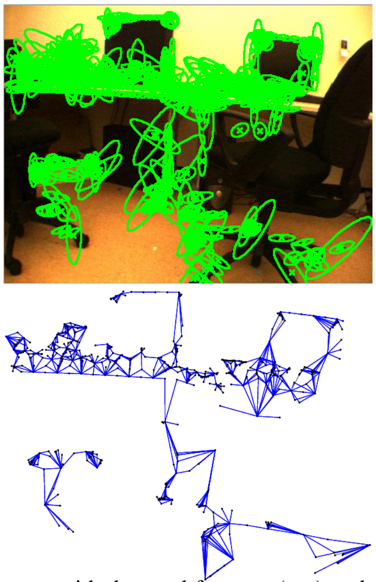

Fig. 4. A test image with detected features (top) and the corresponding MRF of features on which sampling is performed (bottom). Note that the MRF may be disconnected as shown here.

the sense that if a feature is similar to its neighboring features with respect to its appearance and 3D location, it is highly likely that it belongs to the same object type as its neighbors.

We take advantage of this spatially clustered nature of the correspondences by placing the image features in a Markov Random Field (MRF). Each feature is connected to its k closest neighbors in the image, where the neighborhood k is a parameter. Larger values of k make large scale correlations visible while increasing complexity ($k = n_f$ gives a fully connected graph). The value of k depends on the expected size of the projected objects in the images. Figure 4 shows the MRF corresponding to features in an image for $k = 10$.

We define discriminative probabilities, also called edge compatibilities, on the edges of the MRF. These are defined as functions of the depth and appearance of the features involved in the edge, where both functions are the Mahalanobis distance between the respective feature values. Denoting the functions on depth and appearance as f_d and f_a, the discriminative probability is

$$p_e \quad \propto \quad f_d(d_i, d_j) \times f_a(v_i, v_j) \tag{3}$$

where

$$-\log f_d = \left(\frac{d_i - d_j}{\sigma_d}\right)^2 \text{ and } -\log f_a = (v_i - v_j)^T \Sigma_a^{-1} (v_i - v_j)$$

and σ_d and Σ_a are depth variance and appearance covariance respectively, that encode the size of objects and their uniformity of appearance.

The overall sampling scheme to compute the posterior can now be seen to have the following form. We sample clusters in the MRF according to the edge compatibilities and subsequently assign an object type to each cluster according to some prior distribution. The sample configuration is then evaluated using the measurement likelihoods based on the learned object models.

We employ a cluster MCMC sampling algorithm to efficiently implement the above scheme. Common MCMC sampling techniques such as Gibbs sampling and Metropolis-Hastings change the value of only a single node in a sampling step, so that mixing time for the chain, i.e the expected time to move between widely differing states, is exponential. Cluster MCMC methods change the value of multiple nodes at each sampling step, leading to fast convergence.

A. The Swendsen-Wang algorithm

We now describe the Swendsen-Wang (SW) algorithm, which we use to compute the place posterior with fast convergence. The SW algorithm has been interpreted in many ways - as a random cluster model, as an auxiliary sampling method, and as a graph clustering model using discriminative edge probabilities [1]. It is in the latter manner that we use the algorithm.

A sample is produced using the SW algorithm by independently sampling the edges of the MRF to obtain connected components. Consider the graph $G = (V, E)$ of image features, as defined above, with discriminative probabilities $p_e,\ e \in E$ defined in (3) on the edges. We sample each edge of the graph independently and turn "on" each edge with probability p_e. Now only considering the edges that are on, we get a second graph which consists of a number of disjoint connected components. If the discriminative edge probabilities encode the local characteristics of the objects effectively, the connected components will closely correspond to a partition Π of the graph into various objects and the background. The distribution over partitions Π is given as

$$p(\Pi | \mathcal{A}, \mathcal{X}, Z) \quad = \quad \prod_{e \in E_0} p_e \prod_{e \in E \setminus E_0} (1 - p_e) \tag{4}$$

Samples obtained from a typical image feature graph are shown in Figure 5.

To sample over correspondence between image features and object types, we assign an object type to each component of the partition according to the posterior on object types $p(O, T | \mathcal{A}, \mathcal{S}, Z, \Pi)$. Computation of the posterior on object types involves only the appearance measurements since the other measurements also depend on the 3D location of the object T. Applying Bayes Law on the posterior, we get the distribution on object types as

$$p(O_c | \mathcal{A}, \mathcal{S}, Z, \Pi) \quad \propto \quad p(Z | O_c, \mathcal{A}, \mathcal{S}, \Pi) p(O_c | \mathcal{A}, \mathcal{S}, \Pi) \tag{5}$$

where the second term on the right is a prior on object types that can potentially incorporate information regarding the size and appearance of the component, and the frequency with which the object type has been observed in the past. We employ a uniform prior on object types for simplicity since the prior is largely overshadowed by the data in this case. The object type affects only appearance measurements and so, the measurement likelihood in (5) collapses to just the conditionally independent appearance likelihoods on the

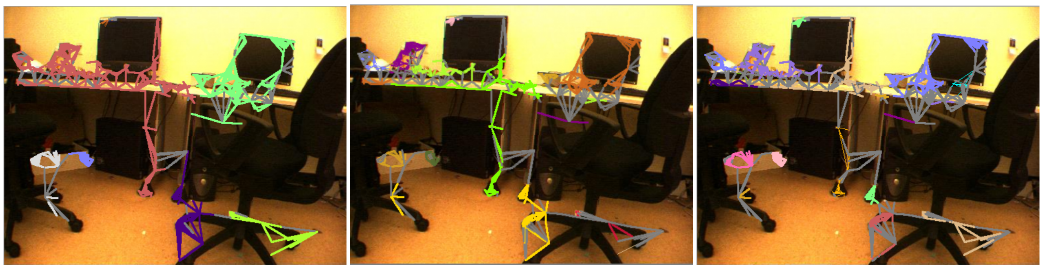

Fig. 5. Samples from the SW algorithm for the MRF corresponding to Figure 4, obtained by independently sampling the MRF edges according to edge compatibilities. Each connected component is colored differently and the edges that are turned "off" are shown in gray. The SW algorithm works by assigning an object label to each connected component and subsequently scoring the sample based on the learned object models. In practice, most components get labeled as background.

Algorithm 1 The Swendsen-Wang algorithm for sampling from the place posterior

```
1) Start with a valid initial configuration $(\Pi, O)$ and repeat for each sample
2) Sample the graph $G$ according to the discriminative edge probabilities (3) to obtain a new partition $\Pi'$
3) For each set in the partition $c \in \Pi'$, assign the object type by sampling from $p(O_c|Z, \Pi')$ as given by (5)
4) Accept the sample according to the acceptance ratio computed using (7)
```

features that are evaluated using the multinomial appearance distribution for the object type

$$p(Z|O_c, \mathcal{A}, \mathcal{S}, \Pi) \quad = \quad \prod_{i=1}^{n_f} p(v_i|O_c, \mathcal{A}, \mathcal{S}, \Pi) \qquad (6)$$

The sample thus generated, consisting of a graph partition and object type assignments to each component in the partition, is accepted based on the Metropolis-Hastings acceptance ratio [6], given as

$$a(\Pi' \to \Pi) \quad = \quad \min\left(1, \frac{\prod_{c' \in \Pi'} p(T_{c'}|O_{c'}, \mathcal{A}, \mathcal{S}, Z, \Pi')}{\prod_{c \in \Pi} p(T_c|O_c, \mathcal{A}, \mathcal{S}, Z, \Pi)}\right) (7)$$

The acceptance ratio (7) can be understood by considering the factorization of the place posterior (2) as

$$p(O_c,T_c|\mathcal{A},\mathcal{S},Z,\Pi) \quad = \quad p(O_c|\mathcal{A},\mathcal{S},Z,\Pi)p(T_c|\mathcal{A},\mathcal{S},Z,\Pi,O_c) \quad (8)$$

and noting that only the second term is involved in the acceptance ratio since the first term, the posterior on object types, has been used in the proposal distribution for the sample above. The acceptance ratio also makes the assumption that the partition components are independent of each other. Note that common factors in the calculation of the acceptance ratio can be omitted to improve efficiency. A summary of the SW algorithm for sampling from the place posterior is given in Algorithm 1.

B. Computing the target distribution

Computing the acceptance ratio (7), involves evaluating the posterior on object locations given a partition of the feature graph. The location posterior, which is the second term on the right in (8), can be factorized into a prior on object locations and the stereo depth likelihood using Bayes law

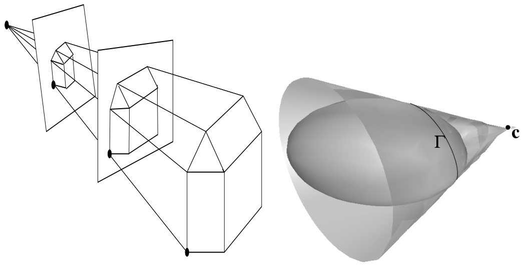

Fig. 6. An illustration of the use of scale to obtain a prior on object location. (a) The scale of the projection in the image gives a good estimate of location if the size of the object in 3D is known. (b) The projection of a 3D object model (shown as a Gaussian ellipsoid) is a 2D covariance, the size of which is used to estimate the object's location. Images taken from [9].

$$p(T_c|\mathcal{A},\mathcal{S},Z,\Pi,O_c) \quad \propto \quad p(D|\mathcal{S},\Pi,O_c,T_c)p(T_c|\mathcal{S},U,\Pi,O_c) \quad (9)$$

where appearance measurements have been neglected since they are assumed independent of the location of the object.

The prior on object locations $p(T_c|\mathcal{S}, U, \Pi, O_c)$ incorporates the 2D pixel locations of the features in the graph component and is evaluated using projection geometry. The graph component is a projection of the object ellipsoid in space and the scale of the projection thus gives an estimate of the 3D object location. Assuming a linear camera projection model, the mean location of the object is projected onto the mean of the feature locations. Analogously, the covariance of the feature locations is a projection of the 3D covariance of the object shape model, which is known. Hence, a proposed 3D location for the object can be evaluated through the norm-1 error between the predicted 2D feature location covariance and the actual covariance. Our use of scale to estimate location in this manner is illustrated in Figure 6. If the observed 2D covariance is Σ_f, the 3D object shape covariance is Σ_O and the Jacobian of the camera projection matrix at T_c is represented as P, the object location prior can be written as

$$\log p(T_c|\mathcal{S}, U, \Pi, O_c) \quad \propto \quad - \left\| \Sigma_f - P\Sigma_o P^T \right\| \qquad (10)$$

In practice, the location distribution (9) displays a peaked nature, so that its evaluation is approximated by a maximum

Algorithm 2 Inference algorithm for constructing the place model

```
For each image $I$ and stereo depth map $D$ obtained at a landmark location, do
  1) Detect features using the 3 types of feature detectors on the image $I$.
  2) Create an MRF using the edge potentials as defined in (3)
  3) Sample over partitions of the graph using the Swendsen-Wang algorithm 1 and obtain the posterior (8) over object types and their locations
  4) The posterior on object types and locations is the required place model
```

a priori (MAP) estimate. This also saves us from sampling over a large continuous space. The maximum of the posterior defined in (9) is obtained by performing a line search along the projection ray of the mean of the feature locations of the graph component. The map value for the object location, $T_c^\star$, is given as

$$T_c^\star = \underset{T_c}{\text{argmax}} \; p(D|\mathcal{S}, U, \Pi, O_c, T_c)p(T_c|\mathcal{S}, U, \Pi, O_c) \quad (11)$$

We compute the individual stereo depth likelihoods by marginalizing over the "true" depth of the feature as measured along the projection ray

$$p(D|\mathcal{S},\Pi,O_c,T_c^\star) = \prod_{i=1}^{n_f} \int_{u_{3d}^i} p(d_i|u_{3d}^i)p(u_{3d}^i|\mathcal{S},U,O_c,T_c^\star) \quad (12)$$

where u_{3d}^i is the true depth measured as d_i by the stereo. The stereo measurement model $p(d_i|u_{3d}^i)$ is modeled as a Gaussian distribution that is learned from measurements obtained in a scenario where ground-truth is known.

The prior on true depth $p(u_{3d}^i|\mathcal{S}, U, O_c, T_c^\star)$ is obtained as a 1D marginal along the projection ray of the object shape model, i.e it is the Gaussian on the projection ray induced by its intersection with the object. This is necessitated by the fact that each feature in the graph component picks its own true depth based on its projection ray. Since both the depth likelihood and true depth prior are Gaussian, the integral in (12) can be computed analytically.

We now have all the pieces to compute the target distribution (8) and the acceptance ratio (7) in the Swendsen-Wang Algorithm 1. Appearance and stereo depth likelihoods are computed using equations (6) and (12) respectively, while a MAP estimate of the object locations is given by (11). A summary of the inference algorithm is given in Algorithm 2.

C. Extension to panoramic models

The algorithm discussed thus far computes the place model for a single image but cannot find a 360^o model for a place unless a panoramic image is used. To overcome this limitation, we propose a simple extension of the algorithm.

We compute the panoramic model of a place from multiple images by "chaining" the models from the individual images using odometry information. For example, the robot is made to spin around at the place of interest to capture multiple images. We designate robot pose corresponding to the first of these images as the base location for the model and marginalize out the poses corresponding to all the other images to create a combined model from all the images.

Denoting the measurements from each of the n images of a place as $Z_1, Z_2, \ldots, Z_n$, and the corresponding detected objects and locations by $O_1, \ldots, O_n$ and $T_1, \ldots, T_n$ respectively, the panoramic model of the place is computed as

$$\begin{aligned} p(O,T|Z_{1:n},o^n) &= \int_{x_{1:n}} p(O_{1:n},T_{1:n}|Z_{1:n},x_{1:n})p(x_{1:n}|o^n) \\ &= p(O_1,T_1|Z_1) \\ &\quad \prod_{i=2}^{n} \int_{x_i} p(O_i,T_i|Z_i,x_i)p(x_i|o_{i-1}) \quad (13) \end{aligned}$$

where x_i is the pose corresponding to the ith image, $o_{i-1}$is the odometry between poses x_{i-1} and x_i, and $o^n = o_{1:n-1}$ is the set of all odometry. x_1 is assumed to be the origin as it is the base location and the pose distribution $p(x_i|o_{i-1})$ is evaluated using the odometry model. Note that (13) uses the fact that the (O_i, T_i) are conditionally independent of each other given the robot pose.

VI. Place Recognition

Place recognition involves finding the distribution on place labels given the detected objects and their locations, i.e. finding the *recognition posterior* $p(L|O, T, \mathcal{A}, \mathcal{S}, Z)$ from (1). While robust techniques for place recognition using feature matching are well-known [15], [3], the detected objects can be effectively used to localize the robot, and can be expected to improve place recognition as they provide higher-level distinguishing information. We now give a technique to accomplish this. Applying Bayes law to the recognition posterior from (1)

$$p(L|O,T,\mathcal{A},\mathcal{S},Z) \propto p(O|L,\mathcal{A},\mathcal{S},Z)p(T|L,O,\mathcal{A},\mathcal{S},Z)p(L) \quad (14)$$

If the sequence of labels observed in the past is available, a Dirichlet label prior $p(L)$ is suitable. We assume that such history is unavailable and so use a uniform label distribution.

The object likelihood $p(O|L, \mathcal{A}, \mathcal{S}, Z)$ is evaluated as the distance between the observed discrete distribution on object types and prediction assuming the label L. Denoting the object type distribution corresponding to L as O_L, the likelihood is

$$\log p(O|L, \mathcal{A}, \mathcal{S}, Z) = -||O - O_L||_2$$

The location likelihood $p(T|L, O, \mathcal{A}, \mathcal{S}, Z)$ is computed by minimizing the distance between corresponding objects in T and the model for the place label L. Nearest neighbor (NN) correspondence is used for this purpose. However, since the robot is unlikely to be in exactly the same location even if it visits the same place, we also optimize over a 2D rigid transformation that determines the current pose of the robot in the local coordinates of the place model for the label L

$$p(T|L,O,\mathcal{A},\mathcal{S},Z) = \int_{X_r} p(T|L,O,\mathcal{A},\mathcal{S},Z,X_r)p(X_r) \quad (15)$$

where X_r, for which we use a flat Gaussian prior, is the location of the robot in the coordinates of the base location of the place L. In practice, we iterate between optimizing for X_r given object correspondence and finding NN correspondence

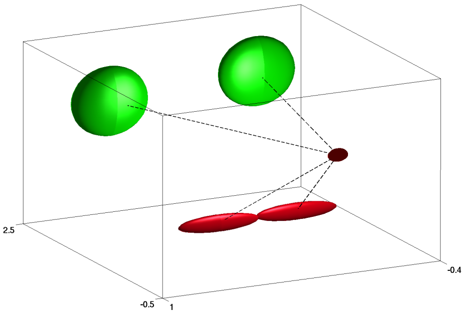

Fig. 7. The most likely model corresponding to the place shown in Figure 5. Detected computer monitors are displayed in green while chairs are shown in red. The origin, where all the dashed lines emanate from, is the location of the camera mounted on the robot. All axes are in meters.

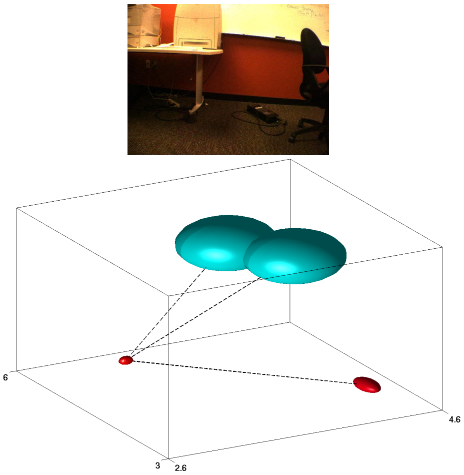

Fig. 8. Image of a place (top) and corresponding place model with two printers (cyan) and a chair (red). The camera is at the origin.

to minimize the object location errors given X_r. Subsequently, the MAP estimate for X_r is used in (15).

VII. Results

We now present results from tests conducted on a robot mounted with a Triclops stereo system. In the test setup, a total of 58 manually segmented training images, obtained from the robot cameras, were presented to the robot with the purpose of learning 5 different object models commonly visible in indoor office environments, namely computer monitors, printers, chairs, cupboards, and drawers. Subsequently, these models were used to construct place models. The problem of deciding which places to model is a non-trivial one, and is out of the scope of this paper. In these experiments, the places to be modeled are selected manually.

The features obtained from the training images were vector quantized after a preliminary classification based on shape (as described in Section III-A) into 1250 bins, which consequently was also the size of appearance histogram for the object models. We used shape and appearance variances for the discriminative edge probabilities (3) in accordance with the expected size and uniformity in appearance of the objects. For example, standard deviations of 0.25 meters for x and y directions, and 0.1 for the z-direction were used. The variability in the z-direction is less because most objects are observed from only a single direction, with the resulting lack of depth information about the object.

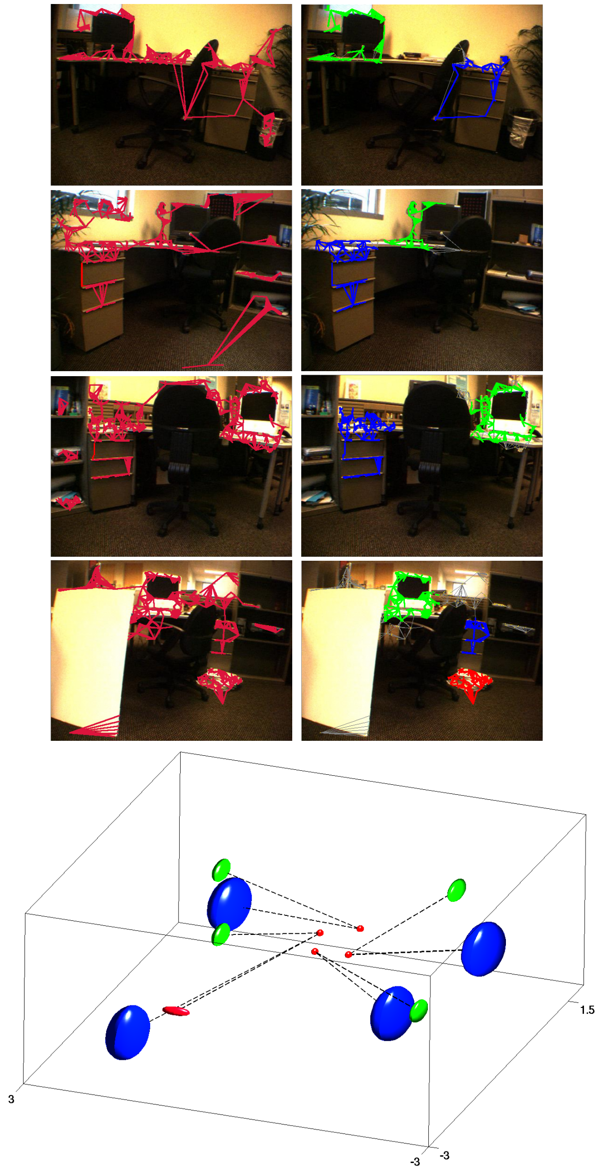

Fig. 9. A panoramic model of an office, containing four monitors (green), four drawers (blue), and a chair (red), constructed using four images. The top four rows show the four images, where the first image of each pair contains the MRF and the second shows the components associated with the objects. The robot poses corresponding to the four images are shown as small brown spheres in the model.

Typical results obtained are given in Figure 7 and 8. These results correspond to the most likely samples obtained from the Swendsen-Wang algorithm. Note that the model for chairs that was learned largely only models the base since few features are detected on the uniform and textureless upper portions. The ROC curves for the objects, shown in Figure 10, quantifies the object detection performance. A panoramic model learned from four images as described in Section V-C is given in Figure 9, which also contains the MAP robot poses.

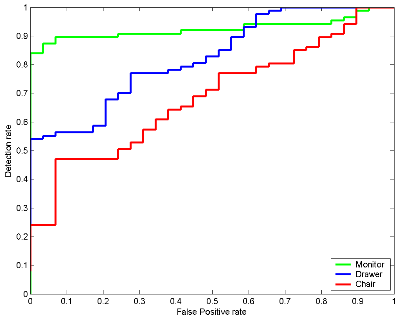

Fig. 10. ROC curves for object detection for some objects used in the experiments. An object is considered detected when it has a large posterior probability in the place model computed from a stereo image.

To test place recognition, we performed an experiment where models for six places were learned and the recognition was assessed on the recall rate of these places when the robot revisited them. A confusion matrix containing the resulting probabilites, calculated using (15), is given in Figure 11. As can be seen from Row 2 therein, a problem with the method is the eagerness of the algorithm to match any two places with the same number and type of objects regardless of their relative location. This problem arises because we do not model the variation in size of individual objects, so that objects of the same type are indistinguishable. We are currently implementing a technique to update the posterior shape of each object from measurements that will alleviate this shortcoming.

	1	2	3	4	5	6
1	0.98	0	0	0	0.02	0
2		0.58	0.42	0	0	0
3			0.58	0	0	0
4				1	0	0
5					0.98	0
6						1

Fig. 11. Symmetric confusion matrix for a place recognition experiment with 6 places. Each row gives the recognition result for a place as a distribution on all the labels.

VIII. Discussion

We described a technique to model and recognize places using objects as the basic semantic concept. A place is modeled as a constellation of parts with respect to a base location, where each part is an object. Each object is modeled as a combination of 3D Gaussian shape and a multinomial appearance. The object models are learned during a supervised training phase. Our algorithm uses affine-invariant features and stereo range data as measurements to compute a posterior on objects, their locations, and the place label during runtime.

This, in many ways, is preliminary work and large number of improvements and extensions are possible. As far as the model is concerned, representing the objects themselves using a parts model will increase the representational power and robustness of the model [20]. We have not used any pose information in this work. Odometry provides a rough estimate of the robot's location and, hence, a basic idea of context, which has been shown to improve object and place recognition [23]. It is also future work to use our technique to build object-based semantic maps.

References

[1] A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(8):1239–1253, August 2005.

[2] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 8(6):679–698, November 1986.

[3] G. Dudek and D. Jugessur. Robust place recognition using local appearance based methods. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1030–1035, 2000.

[4] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *Intl. Conf. on Computer Vision (ICCV)*, 2005.

[5] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.A. FernÃ¡ndez-Madrigal, and J. GonzÃ¡l ez. Multi-hierarchical semantic maps for mobile robotics. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3492–3497, 2005.

[6] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, 1996.

[7] O. Grau. A scene analysis system for the generation of 3-D models. In *Proc. Intl. Conf. on Recent Advances in 3-D Digital Imaging and Modeling*, pages 221–228, 1997.

[8] D. Hähnel, W. Burgard, and S. Thrun. Learning compact 3D models of indoor and outdoor environments with a mobile robot. *Robotics and Autonomous Systems*, 44(1):15–27, 2003.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[10] B. Kuipers. Modeling spatial knowledge. In S. Chen, editor, *Advances in Spatial Reasoning (Volume 2)*, pages 171–198. The University of Chicago Press, 1990.

[11] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. J. of Computer Vision*, 60(2):91–110, 2004.

[12] K. Lynch. *The Image of the City*. MIT Press, 1971.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Intl. J. of Computer Vision*, 65(1/2):43–72, 2005.

[14] A. Nüchter, H. Surmann, K. Lingemann, and J. Hertzberg. Semantic scene analysis of scanned 3D indoor environments. In *Proceedings of the 8th International Fall Workshop Vision, Modeling, and Visualization 2003*, pages 215–222, 2003.

[15] I. Posner, D. Schroeter, and P. Newman. Using scene similarity for place labeling. In *International Symposium of Experimental Robotics*, 2006.

[16] A. Rottmann, O. Martinez Mozos, C. Stachniss, and W. Burgard. Semantic place classification of indoor environments with mobile robots using boosting. In *Nat. Conf. on Artificial Intelligence (AAAI)*, 2005.

[17] D. Schröter and M. Beetz. Acquiring models of rectangular objects for robot maps. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2004.

[18] D. Schröter, M. Beetz, and B. Radig. RG Mapping: Building object-oriented representations of structured human environments. In *6-th Open Russian-German Workshop on Pattern Recognition and Image Understanding (OGRW)*, 2003.

[19] Josef Sivic, Bryan Russell, Alexei A. Efros, Andrew Zisserman, and Bill Freeman. Discovering objects and their location in images. In *Intl. Conf. on Computer Vision (ICCV)*, 2005.

[20] E. Sudderth, A. Torralba, W. Freeman, and A. S. Willsky. Depth from familiar objects: A hierarchical model for 3d scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[21] M.A. Tanner. *Tools for Statistical Inference*. Springer Verlag, New York, 1996. Third Edition.

[22] S. Thrun, M. Montemerlo, and et al. Stanley, the robot that won the DARPA grand challenge. *Journal of Field Robotics*, 2006. In press.

[23] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Intl. Conf. on Computer Vision (ICCV)*, volume 1, pages 273–280, 2003.

[24] S. Vasudevan, S. Gachter, M. Berger, and R. Siegwart. Cognitive maps for mobile robots: An object based approach. In *Proceedings of the IROS Workshop From Sensors to Human Spatial Concepts (FS2HSC 2006)*, 2006.