A Model of Shared Grasp Affordances from Demonstration

John D. Sweeney and Rod Grupen
Laboratory for Perceptual Robotics
University of Massachusetts Amherst
{sweeney,grupen}@cs.umass.edu

Abstract— **This paper presents a hierarchical, statistical topic model for representing the grasp preshape affordances of a set of objects. The affordances are shared among the objects, and learned from training data provided by teleoperation. Each affordance defines a distribution over visual appearance and the position/orientation of the hand. The parameters of the model are learned using a Gibbs sampling method. After training, the affordances can be used to compute grasp preshapes for novel objects. The model is evaluated experimentally on a set of objects for its ability to generate grasp preshapes that lead to successful grasps, in comparison with a baseline approach.**

I. Introduction

For robots performing object manipulation tasks, the notion of grasp affordances provide a useful way of describing and categorizing the ways in which the robot can interact with objects [1]. The grasp affordances of an object are the ways in which the object can be grasped in order to be used for a particular function. For example, a coffee mug has at least two distinct affordances: one for drinking (typically by using the handle), and another for transporting. Note that while one could transport the mug using the same grasp that was used for drinking, the converse may not be true (grasping the top of the mug). The physical characteristics of an object, e.g. visual appearance, provide a way of inferring its affordances. Further, the grasp affordances of an object provide a natural categorization of objects based on function rather than appearance, which may vary drastically among similar objects.

In this paper, we describe how a form of statistical topic model can be used to model grasping strategies based on a set of affordances common to all objects. Traditionally, topic models, such as Latent Dirichlet Allocation (LDA), have been used in the information retrieval community to model the distribution of words in a text corpus [2]. These models assume that each document is a collection of latent "topics," where each topic is defined as a multinomial distribution over the vocabulary. These models have been applied in other domains, such as visual object recognition [3], [4]. The generative process for documents under the topic model begins with choosing a document-specific distribution over topics. Next, for each word in the document, a topic is sampled using the document-specific distribution, and then a word is sampled using the distribution defined by the chosen topic.

In our model, each topic corresponds to a grasp affordance, where the affordance defines a distribution over variables that can be used to describe its appearance and the position and orientation of the hand for a grasp preshape. Each time the object is grasped, an affordance of the object is sampled from the distributions implied by that object. Furthermore, these affordances may be shared across objects. Since the model is learned from actual grasps of objects performed by teleoperation, the clustering process is in effect modeling the affordances that were used in the training process.

One difference between the model presented in this paper and traditional topic models is that the "words" in this case are the conditionally independent component distributions used to describe the appearance and physical location of a grasp. Furthermore, these distributions are continuous, whereas in the traditional topic model, each topic is a multinomial distribution over a set of words.

The experimental platform used in this paper is Dexter, the UMass bimanual humanoid, shown in Figure 1. In order to collect training data, we teleoperate Dexter to perform grasps on objects, and use its stereo vision system to compute visual features.

II. Related Work

This paper is influenced by the hierarchical, part-based model of Sudderth et al. [4]. In that work, they describe a visual object classifier that models each object by computing a multinomial distribution over a set of globally shared "parts." Each part describes a cluster of image features. The parts in their model are analogous to the affordances described here. One difference is that in applying the model to new objects, we do not have

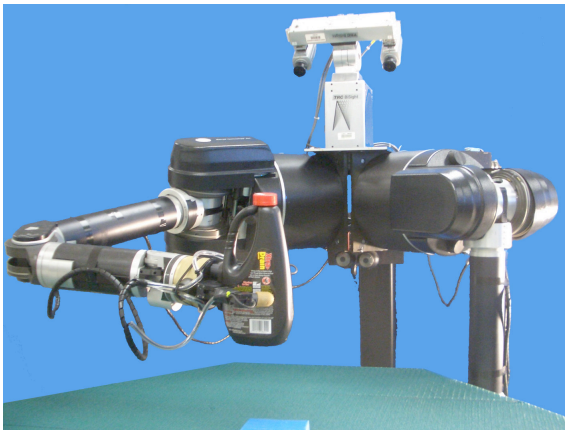

Fig. 1. This picture shows Dexter performing a grasp of the `black_drano` object, as described in Section VI. Each of Dexter's arms has 7 DOF and is equipped with a three-fingered hand with four total degrees of freedom. The stereo head has four degrees of freedom.

a full set of features, and instead use only the visual appearance to infer object class. Unlike their model, we are not interested in the classification problem; rather, we use the affordances learned by our model to generate new grasps on novel objects.

The model of Sudderth is an adaptation of statistical topic models such as the author-topic model [5] and Latent Dirichlet Allocation (LDA) [2]. In their original paper describing LDA, Blei et al. proposed using a variational method for performing approximate inference of the model [2]. Since then, Griffiths and Steyvers [6] proposed using a Gibbs sampling method for inference, which is the approach used in this work.

There has been work on how to generate grasps using visual information. In Saxena et al. [7], they are interested in computing a grasp point for an object by analyzing visual features. Their model performs a regression that estimates likely grasp positions in a 2D image based on the features. This is similar in spirit to the motivation of this work, which is to be able to estimate likely grasp preshapes from visual information alone. The difference being that this work explicitly models multiple grasp hypotheses for each object.

Platt [8] describes a scheme for generating hypothesis grasps based on analyzing the first and second moments of the foreground blob segment. We use similar visual features in this work, but we generate grasp hypotheses by using a model of the training set of demonstrated grasps.

III. Representing Affordances in the Model

As previously described, each affordance is represented as a tuple of three parameterized probability distributions: the visual appearance of the object, the hand's position, and the hand's orientation during grasping. We discuss how each of these features of an affordance are represented in turn.

A. Visual Appearance

To compute the visual appearance of an object used in the model, we first segment the object into a foreground blob using background subtraction. The object's centroid $\hat{O}_m \in \mathbb{R}^3$ is measured by performing a stereo triangulation on the first moments of the left and right foreground blobs. The visual feature b_m, is the average second moment of the left and right foreground segments, and is modeled as a two-dimensional inverse-Wishart distribution, parameterized by scale ψ with u degrees of freedom:

$$p(b_m \,|\, \psi, u) = \text{Inv-Wishart}_u(b_m \,|\, \psi). \tag{1}$$

The distribution is unimodal, and generally used to model priors on multivariate covariance matrices. This feature provides a proof of concept of the model, but other types of visual features can be used. For example, in other work, a multinomial distribution over a set of SIFT features is used to allow each topic to have many likely appearances [3], [4]. The main requirement is that there must be a probability distribution to describe the likelihood of features in an affordance.

B. Grasp Position

The position of the hand with respect to oriented features of the object when initiating the grasp determines which affordance of the object is being used. In theory, each grasp affordance describes an entire region in the hand's twist space relative to the object. It is this notion we wish to capture in our representation of an affordance. It is convenient, computationally, to model position and orientation as independent distributions.

From each training grasp point $p \in \mathbb{R}^6$, we model the position of the hand $x_p \in \mathbb{R}^3$ in a frame centered at the centroid of the object, $\hat{O}_m$, using a normal distribution with mean μ and covariance Σ:

$$p(x_m \,|\, \mu, \Sigma) = \mathcal{N}(x_m \,|\, \mu, \Sigma). \tag{2}$$

C. Grasp Orientation

Given p, we perform a discrete quantization of the rotational component to extract the orientation feature of the grasp point. This quantization matches the orientation of the hand, represented as a unit quaternion $\mathbf{q}_p \in \mathbb{H}$, to a set of Q canonical hand orientations. These orientations are represented by a set of Dimroth-Watson distributions, bimodal, symmetric distributions

over the space of unit quaternions [10], [11]. Each DW distribution has a mode $\hat{\mathbf{q}}$ and concentration parameter κ. The canonical orientations are a set of these DW parameters, $\mathcal{Q} = \{(\hat{\mathbf{q}}_1, \kappa_1), \ldots, (\hat{\mathbf{q}}_Q, \kappa_Q)\}$. Thus the extracted feature is the index of the DW distribution with highest likelihood:

$$w_p = \arg\max_{i=1..Q} \mathrm{DW}(\mathbf{q}_p \,|\, \hat{\mathbf{q}}_i, \kappa_i), \tag{3}$$

where $d(i, j)$ is a measure of distance between i and j.

The set $\mathcal{Q}$ is the result of computing a mixture model using DW components over hand orientations from a training set of grasps. The parameters of each component in the mixture model make up one member of $\mathcal{Q}$.

Each affordance then defines a discrete distribution over the set of canonical orientations:

$$p(w_p \,|\, \phi) = \phi(w_p), \tag{4}$$

where ϕ is a Q-vector in the $(Q-1)$-simplex such that $\phi(i)$ is the probability of selecting orientation i.

By using a multinomial model for grasp orientation, each affordance can represent multiple orientations. This is useful for dealing with the symmetries that can occur when grasping using Dexter. For example, Dexter can perform a side grasp on an object with the thumb pointing towards or away from the robot. If the training data consisted of both types of grasps, then the affordance that encodes that particular side grasp will have nonzero probability of choosing both types of orientation. Note that for each affordance, the probability table over orientation is built using the training data, so orientations that are more prevalent in the training set are more likely in the model.

IV. The Generative Model

The basic idea behind the generative model is to associate each example grasp with a latent "affordance" variable and then sample the object appearance and grasp position and orientation from the distributions specified by that affordance. Using the affordances sampled from the posterior, we can make predictive distributions over grasp position and orientation for new objects.

The generative model is illustrated in Figure 2. The nodes of the graph represent random variables, with the shaded nodes denoting the observed variables. Rectangles around variables denote replication, where the number of times is shown in the bottom right. Rounded nodes indicate fixed hyperparameters.

We organize the training data set $\mathcal{D} = (\mathbf{b}, \mathbf{x}, \mathbf{w})$ into M sets of object grasp features, where set m has N_m examples. Each object corresponds to a "document" in the text topic model setting. Datum i of object m is

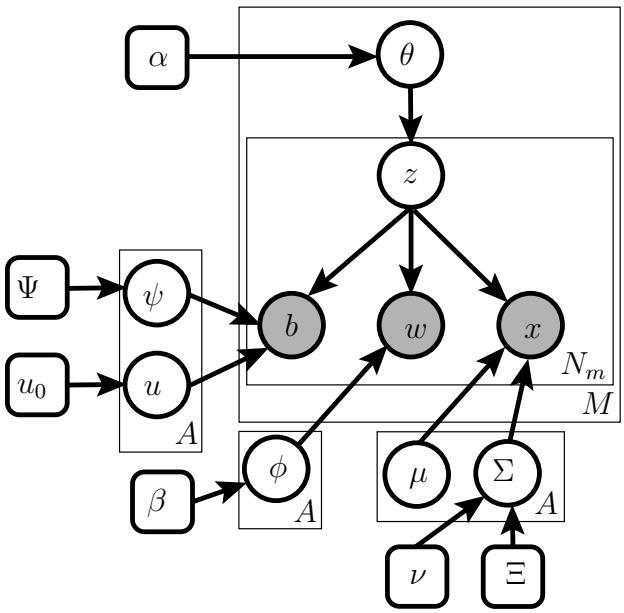

Fig. 2. The graphical model described in this paper. Circles indicate random variables, with shading indicating they are observed. The boxes around nodes represent replication; the number of times written in the bottom right corner. Rounded nodes indicate fixed hyperparameters. The edges between nodes indicates a conditional probability distribution described in the text.

the tuple $(b_{mi}, x_{mi}, w_{mi},)$, which consists of a visual feature, the position, and the orientation of the hand during the grasp, respectively.

The generative process for data point i is given below:

$$\begin{aligned}
\theta \,|\, \alpha &\sim \mathrm{Dirichlet}(\alpha) \\
z_i \,|\, \theta &\sim \mathrm{Multinomial}(\theta) \\
b_i \,|\, z_i = j &\sim \text{Inv-Wishart}_{u_j}(\psi_j) \\
w_i \,|\, z_i = j &\sim \mathrm{Multinomial}(\phi_j) \\
x_i \,|\, z_i = j &\sim \mathcal{N}(\mu_j, \Sigma_j),
\end{aligned} \tag{5}$$

where $X \sim \mathbb{D}$ means that random variable X is sampled from distribution $\mathbb{D}$. Moreover, θ is sampled from an A-dimensional Dirichlet distribution. The affordance components b_i, w_i, and x_i are sampled according to the distributions (1), (4), and (2), described in Section III.

As shown in (5), θ describes a multinomial distribution over the shared set of affordances for an object; in effect, it defines a mixture model over affordances. Once an affordance j is chosen from the set, the visual appearance, grasp position, and grasp orientation are sampled using the parameters specified by affordance j. Note that by using a multinomial distribution over the affordances, independent samples from the model for the same object can result in a different affordance being selected.

For notational convenience, let $\Omega = (\alpha, \beta, \Psi, u_0, \nu, \Xi)$ correspond to the fixed hyperparameters in the model, and let $\mathbb{C} = (\psi, u, \phi, \mu, \Sigma)$ correspond to the parameters of the component distributions for each affordance.

In order to compute the marginalized likelihood of a set of N data points from a single object, we must integrate out the component distribution parameters and the affordances mixtures $\mathbf{z}$:

$$p(\mathcal{D} \mid \Omega) = \int_{\mathbb{C},\,\theta} \cdots \int p(\theta \mid \alpha) p(\psi, u \mid \Psi, u_0) \times p(\mu, \Sigma \mid \nu, \Xi) p(\mathcal{D} \mid \theta, \mathbb{C}) d\theta d\mathbb{C} \quad (6)$$

where

$$p(\mathcal{D} \mid \theta, \mathbb{C}) = \prod_{n=1}^{N} \int p(b_n, w_n, x_n \mid \mathbf{z}, \mathbb{C}) p(\mathbf{z} \mid \theta) d\mathbf{z} = \prod_{n=1}^{N} \sum_{a=1}^{A} p(b_n \mid \psi_a, u_a) p(w_n \mid \phi_a) p(x_n \mid \mu_a, \Sigma_a) \theta(a), \quad (7)$$

and the likelihoods in (7) are (1), (4), and (2).

Thus for M objects, the marginal likelihood is

$$p(\mathcal{D} \mid \Omega) = \prod_{m=1}^{M} \int_{\mathbb{C},\,\theta} \cdots \int p(\theta_m \mid \alpha) p(\psi, u \mid \Psi, u_0) \times p(\phi \mid \beta) p(\mu, \Sigma \mid \nu, \Xi) p(\mathcal{D} \mid \theta, \mathbb{C}) d\mathbb{C} d\theta. \quad (8)$$

V. Parameter Estimation in the Model

The inference problem is to compute the posterior distribution of the latent variables given example grasp points:

$$p(\theta, \mathbf{z}, \mathbb{C} \mid \mathcal{D}, \Omega) = \frac{p(\mathcal{D} \mid \theta, \mathbf{z}, \mathbb{C}, \Omega) p(\theta, \mathbf{z}, \mathbb{C} \mid \Omega)}{p(\mathcal{D} \mid \Omega)}, \quad (9)$$

which is intractable to compute, although we can estimate it using Gibbs sampling. In order to estimate the posterior, we perform a clustering of the data into a set of affordances that are usable by all objects. The parameters of each affordance are determined by the training data points assigned to that cluster; these points are collected from all the presented objects which use that affordance Given our data set $\mathcal{D}$, we use Gibbs sampling to estimate the affordance assignments $\mathbf{z}$, which we use to provide point estimates for the other parameters θ and $\mathbb{C}$.

We assume independent, symmetric Dirichlet priors over θ and ϕ, with hyperparameters α and β, respectively. The blob covariance prior is inverse-Wishart with scale Ψ and u_0 degrees of freedom. The covariance matrices for grasp position, Σ, also have an inverse-Wishart prior with scale Ξ and ν degrees of freedom [9]. The grasp position mean is given a noninformative prior as well.

The idea behind Gibbs sampling is that while we cannot sample directly from the target state space, viz. the assignment of affordances to data points $\mathbf{z}$, we can sample each dimension of the space conditioned on the current state of the rest of the dimensions. The sampler outputs a Markov chain, so a number of iterations must be computed before samples can be considered independent. In the following, let $\mathbf{z}_{-mi}$ denote the set of all affordance assignments excluding z_{mi}, and let $\mathbf{b}_{-mi}$, $\mathbf{x}_{-mi}$, and $\mathbf{w}_{-mi}$ be defined similarly.

Using the conditional independence relationships shown in the graph of Figure 2, the posterior distribution over affordance assignments can be written as

$$p(z_{mi} \mid \mathbf{z}_{-mi}, \mathcal{D}) \propto p(z_{mi} \mid \mathbf{z}_{-mi}, o_m) \times p(b_{mi} \mid \mathbf{z}, \mathbf{b}_{-mi}) p(x_{mi} \mid \mathbf{z}, \mathbf{x}_{-mi}) \times p(w_{mi} \mid \mathbf{z}, \mathbf{w}_{-mi}). \quad (10)$$

The likelihoods of the conditional affordance assignments and hand orientation assignments are multinomials, and have been derived from standard Dirichlet integrals:

$$p(z_{mi} = j | \mathbf{z}_{-mi}, o_m = l) = \frac{n^O_{jl} + \alpha}{\sum_{j'} n^O_{jl'} + A\alpha} \quad (11)$$

$$p(w_{mi} = k | z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{w}_{-mi}) = \frac{n^W_{kj} + \beta}{\sum_{j'} n^W_{kj'} + Q\beta}. \quad (12)$$

Where n^O_{jl} is the number of times affordance j has been assigned to object l, and A is the number of shared affordances. Likewise, n^W_{kj} is the number of times orientation feature k has been assigned to feature j, and Q is the number of canonical grasp orientations.

At each iteration of the sampling algorithm, given the current assignment of data points to affordances, the posterior distribution over the position of the grasp, x_{mi}, is a multivariate Student-t distribution with $(n^A_j + \nu - 2)$ degrees of freedom, where n^A_j is the total number of features assigned to affordance j. This can be approximated with the following moment-matched normal distribution [9]:

$$p(x_{mi} \mid z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{x}_{-mi}) \approx \mathcal{N}(x_{mi} \mid \hat{\mu}_j, \hat{\Sigma}_j), \quad (13)$$

where

$$\hat{\mu}_j = \frac{1}{n_j^A} \sum_{m=1}^{M} \sum_{k|z_{mk}=j} x_{mk}$$

$$\delta_j = \frac{n_j^A + 1}{n_j^A (n_j^A + \nu - 4)}$$

$$\hat{\Sigma}_j = \delta_j \left(\Xi + \sum_{m=1}^{M} \sum_{k|z_{mk}=j} (x_{mk} - \hat{\mu}_j)(x_{mk} - \hat{\mu}_j)^T \right).$$

The conditional distribution for a blob covariance is given as

$$p(b_{mi} \,|\, z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{b}_{-mi}) = \text{Inv-Wishart}_{\hat{u}_j}(\hat{\psi}_j) \tag{14}$$

with

$$\begin{aligned} \hat{u}_j &= u_0 + n_j^A \\ \hat{\psi}_j &= \frac{1}{n_j^A} \left(\Psi_0 + \sum_{m=1}^{M} \sum_{k|z_{mk}=j} b_{mk} \right). \end{aligned} \tag{15}$$

At each iteration of the Gibbs sampler, we use (11) – (14) to compute (10). A single data point update can be computed in $O(A)$, and each sample output by the sampler requires computing this assignment for every training data point. Thus the total time to compute a sample given a training set with M objects and N grasps per object is $O(MNA)$.

A. Generating grasps for new objects

After allowing the Gibbs sampler time to converge, each sample approximates a sample from the posterior distribution. To generate grasp positions for a new object, we can use the affordances indicated by the posterior.

In general, we are interested in generating candidate grasp locations for a novel object given the visual features. Let $\hat{\Theta}^{(\mathrm{s})}$ correspond to the model parameters estimated from sample s. The generative process for new grasps given blob covariance b_t is:

$$\begin{aligned} z_t \,|\, b_t, \hat{\Theta}^{(\mathrm{s})} &\sim p(z \,|\, b_t, \hat{\Theta}^{(\mathrm{s})}) \\ w_t \,|\, z_t = j, \hat{\Theta}^{(\mathrm{s})} &\sim \text{Multinomial}(\hat{\phi}_j^{(s)}) \\ x_t \,|\, z_t = j, \hat{\Theta}^{(\mathrm{s})} &\sim \mathcal{N}(\hat{\mu}_j^{(s)}, \hat{\Sigma}_j^{(s)}). \end{aligned} \tag{16}$$

With a set of samples from the posterior distribution $p(\mathbf{z} \,|\, \mathcal{D})$, statistics that are independent of the content of individual affordances can be computed by integrating over the full set of samples. For any single sample $\hat{\Theta}^{(\mathrm{s})}$ we can estimate θ and $\mathbb{C}$ using the affordance assignments in $\mathbf{z}^{(s)}$ as described in Section V using (11) – (14). These correspond to predictive distributions over new affordances and grasp positions conditioned on $\mathcal{D}$ and $\mathbf{z}$. Note that these estimates cannot be combined across samples, since there is no guaranteed correspondence between affordances among the set of samples.

The first distribution in (16) can be computed as

$$\begin{aligned} p(z = i \,|\, b_t, \hat{\Theta}^{(\mathrm{s})}) &\propto p(b_t \,|\, z = i, \hat{\Theta}^{(\mathrm{s})}) p(z = i \,|\, \hat{\Theta}^{(\mathrm{s})}) \\ &\approx \text{Inv-Wishart}_{\hat{u}_i^{(s)}}(\hat{\psi}_i^{(s)}), \end{aligned} \tag{17}$$

where we assume that $p(z = i \,|\, \hat{\Theta})$ is uniform. An alternative is to use information about how z is allocated among the objects in $\mathcal{D}$, for example, by taking the average

$$p(z = i \,|\, \hat{\Theta}^{(\mathrm{s})}) = \frac{1}{M} \sum_{m=1}^{M} \theta_m^{(s)}(i). \tag{18}$$

By following the generative process in (16), we can produce a set of possible grasp positions for the robot to choose from given a visual feature.

VI. Experimental Results

To experimentally test the ability of the model to represent the grasp affordances shown in the training set, and generate new grasps, a set of 31 objects $\mathcal{O}$ was chosen for grasping. A random selection of these objects were designated training, and the rest test objects. The training set consisted of teleoperating Dexter to perform five grasps using each grasp affordance reachable using the right arm. Each object was presented to Dexter in the middle of the workspace, and the right arm was used to perform all grasps, as shown in Figure 1.

Because there is no notion of orientation of the object, the same object presented in multiple orientations (flat, standing up, etc.) is treated as separate objects. In the experiments, the notation `object-N` refers to the presentation of `object` in a different orientation. There are examples in the literature of how this assumption can be relaxed by incorporating the notion of rigid body transformations into the model itself [13].

For training, $N_{train} = 19$ objects were chosen randomly from $\mathcal{O}$, and grasps were demonstrated using teleoperation. This object set is shown in Figure 3. The set of grasp orientations, $\mathcal{Q}$, was computed using the training set, and a set of $Q = 6$ DW parameters were chosen. Note that in these experiments, symmetric grasps were not used, that is, the demonstrator did not perform a grasp at the same location using a different hand orientation.

Additionally, each of the objects was represented by a single blob feature. The set of test objects along with their visual feature is shown in Figure 4.

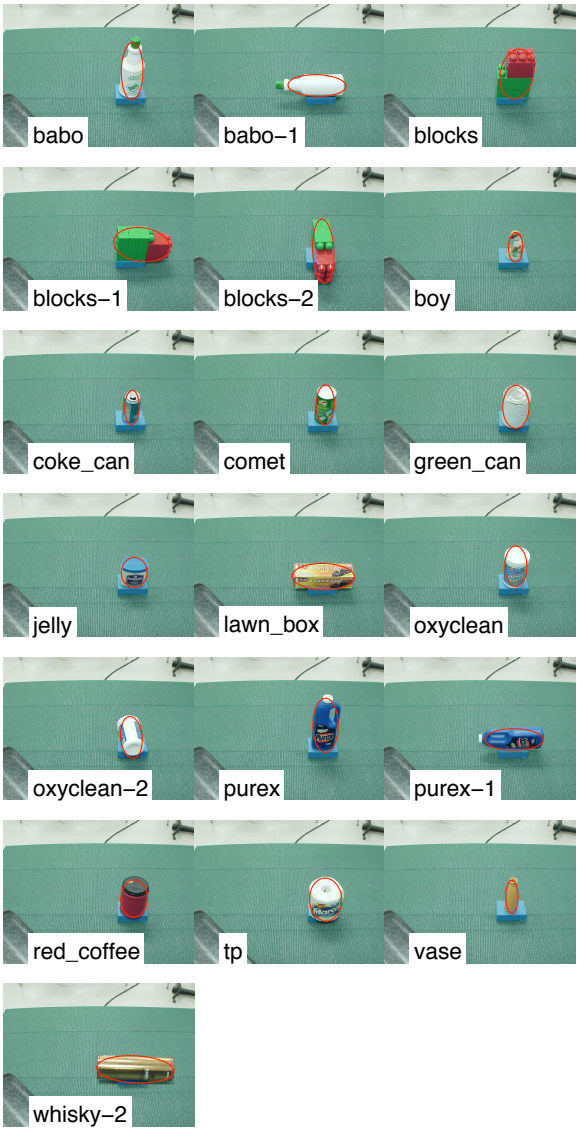

Fig. 3. This picture shows the objects in the training set. The red oval corresponds to the covariance matrix that was used as a visual feature for grasps with the object.

Fig. 4. This picture shows the objects as they were presented for generating grasps.The red oval corresponds to the covariance matrix that was computed from the average second moments of the segmented blob in the left and right cameras.

For learning the parameters of the model, $A = 10$ shared grasp affordances were used. The Gibbs sampler ran for 200 iterations of burn-in, and after which 10 samples were stored by running the sampler for 10 iterations and then storing the next sample.

Using the set $\mathcal{S}$ of 10 samples, $N_{test} = 12$ objects were presented, and the model generated 6 candidate grasps for each object; these were transformed into the base frame from the object's frame and then the robot attempted each grasp. It should be noted that technically we are generating grasp preshapes for the hand: the position and orientation to achieve before performing a grasp. In our experiments, the grasp itself is simply flexing the fingers until a sufficient force has been applied to the object. One could incorporate a grasp controller to perform the actual grasp once the hand has been moved to the hypothesized preshape location [14]. In these experiments a grasp was judged successful if the robot was still holding onto the object after moving the hand 10 cm vertically after grasping.

As an example of the types of grasps generated by the model, Figure 5 shows a composite image of six grasps generated for the `blocks-3` object.

To analyze the performance of the model, we created a naïve model which also generated grasps using the blob feature. This model performed visual processing to estimate the width and height of the object, and then generated grasps by selecting points on a spherical hemisphere centered at the object's centroid. The radius of the hemisphere was equal to half the length of the longest dimension of the object. The orientation of the hand was chosen such that the palm was normal to the ray connecting the object's centroid. Moreover, a uniform random rotation about this ray was chosen. The three fingers of the hand were spread equidistant from each other. The robot then attempted to grasp the object at each of the six locations, and grasp success

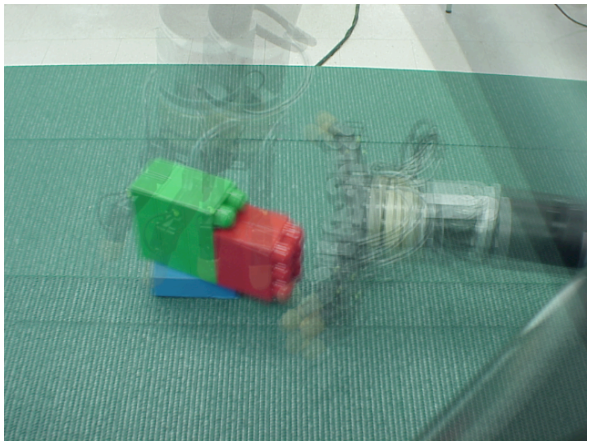

Fig. 5. A composite image showing six candidate grasp positions for the `blocks-3` object.

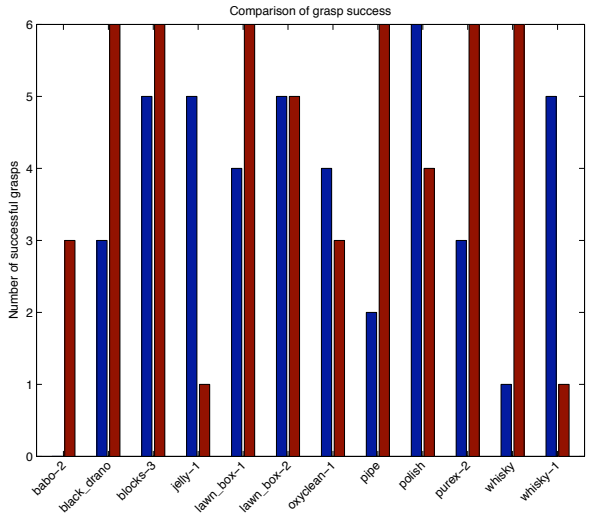

Fig. 6. This graph shows the result of using the trained grasp model on a set of test objects. Each bar measures the number of successful grasps for the labeled object. The blue bars are for the naïve model, and the red for the shared affordance model.

was judged as before. In the experiments shown here, the model is used to generate grasps for the right arm of the robot. It could be adapted for use with the left arm by suitably transforming the predicted grasps according to the symmetries between left and right arms.

The results of performing these grasps are shown in Figure 6, where blue and red bars correspond to the naïve and affordance model, respectively. Overall, the naïve model was successful 43 out of 72 total grasp attempts. In comparison, using the affordance model, 53 out of 72 attempts were successful; a statistically significant improvement ($p < 0.01$).

In most cases, our model outperformed the naïve mode approach, including the `babo-2` object, which the naïve model was unable to grasp. However, our model did have difficulty with the `jelly-1` and `whisky-1` objects. In both case, although the generated grasps were located above the object with a suitable orientation, they were too high for a successful grasp. This is a result of the fact that the model is in effect summarizing the grasps provided in the demonstration. For novel objects, the model finds the affordance with the most similar appearance, but the grasp positions suggested by that affordance may not adequately fit the actual geometries of the object.

In these experiments, there was no secondary analysis of the candidate grasps; the affordance distribution was sampled and the candidate was attempted, in order to show that successful grasps can be achieved using only the visual feature and the object's centroid. Since the model represents the affordance as a distribution, an improved method could incorporate additional knowledge about the object into the candidate selection. For example, by sampling a number of candidates and choosing the one closest to the object (that was not inside the object—the model has no notion about the size of the object being grasped), or by refining candidate hypotheses using more detailed geometric information about the object.

As a nonparametric approach, by providing the model with more training data, the variance of the affordance's position distribution can be reduced; potentially improving grasp performance. Furthermore, the success rate of the model is affected by the number of shared affordances that is used. Although we do not know a priori how many shared affordances there may be among the training data, in the current implementation we specify a fixed number of affordances. If this number is too small, the covariances for the position distribution of the affordance will be large, so it may take a number of samples to find one that is close enough to the object to successfully grasp it. The expected number of affordances used is a function of the number of data points and α. Nonparametric Bayesian approaches similar to hierarchical LDA proposed in [12] can be used to estimate the number of affordances from the data itself.

In order to see how affordances were shared among different objects, we computed Table I using a single sample of the posterior to show the composition of each affordance. Each column corresponds to an affordance, and each row denotes the training set of objects. An "x" indicates that some training grasp from this object was

	Affordance									
	1	2	3	4	5	6	7	8	9	10
babo		x				x				
babo-1			x							
blocks		x		x						
blocks-1	x		x							
blocks-2								x		
boy					x					
coke_can					x		x			
comet						x		x		
green_can		x		x						
jelly					x	x				
lawn_box			x							
oxyclean				x		x				
oxyclean-2								x		
purex		x		x						
purex-1									x	
red_coffee		x		x		x				
tp		x		x						
vase							x			
whisky-2										x

TABLE I

EACH "X" DENOTES A GRASP ON THE OBJECT IN THE ROW WAS USED BY THE AFFORDANCE DENOTED IN THE COLUMN.

used to determine the parameters of the affordance in that column. Columns with multiple "x"s indicate an affordance that used training examples from multiple objects. In this sample, it can be seen that 7 out of 10 affordances incorporate training examples from multiple objects. Once the sampler has been run for enough iterations, we can expect subsequent samples to contain very similar assignments. Different runs of the sampler produce similar sets of affordances, although the actual assignments to particular affordances will differ (e.g., the assignment found in affordance 1 in this sample may be the assignment in affordance 5 in another sample).

In these experiments, there was no secondary analysis of the candidate grasps; the affordance distribution was sampled and the candidate was attempted. Since the model represents the affordance as a distribution, an improved method could incorporate additional knowledge about the object into the candidate selection. For example, by sampling a number of candidates and choosing the one closest to the object (that was not inside the object—the model has no notion about the size of the object being grasped).

Another way to improve the model is to provide it with more training examples that correspond to the different positions where an object might be grasped. Furthermore, the success rate of the model is affected by the number of shared affordances that is used. Although we do not know a priori how many shared affordances there may be among the training data, in the current implementation we specify a fixed number of affordances. If this number is too small, the covariances for the position distribution of the affordance will be large, so it may take a number of samples to find one that is close enough to the object to successfully grasp it. The expected number of affordances used is a function of the number of data points and α. Nonparametric Bayesian approaches similar to hierarchical LDA proposed in [12] can be used to estimate the number of affordances from the data itself.

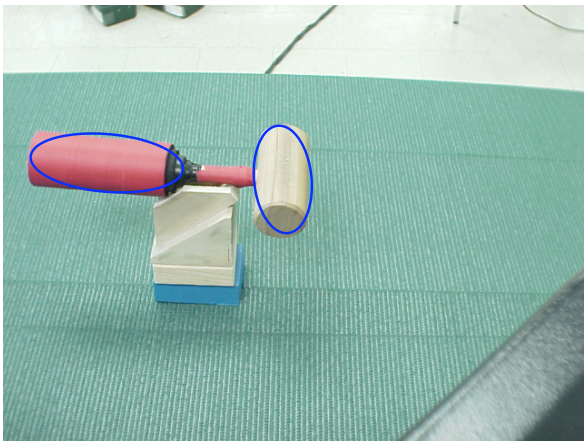

Fig. 7. This figure shows how the mallet can be segmented into multiple blobs, and each blob can be used to generate grasp positions independently.

A. More Complex Objects

Using this visual feature model, one can use multiple blob covariances to represent a single object: the model will generate grasps for each covariance, and they must then be transformed into the base frame. Again, the model has no notion of the geometry of the object being grasped, so secondary processing should be used to select those grasps that have a high likelihood of failure. As an example, we presented a mallet that was segmented into two blobs, as shown in Figure 7.

Using the model from the previous section, we generated grasps from each of the two blobs. Additional processing was required to discard some of these candidates, as they suggested grasps that collided with the mallet. For example, the model generated side grasps for the handle of the mallet that would collide with the head of the mallet. Figure 8 shows feasible candidate grasps suggested by the model.

VII. CONCLUSIONS

We have presented a hierarchical, statistical model for representing grasp affordances among a collection

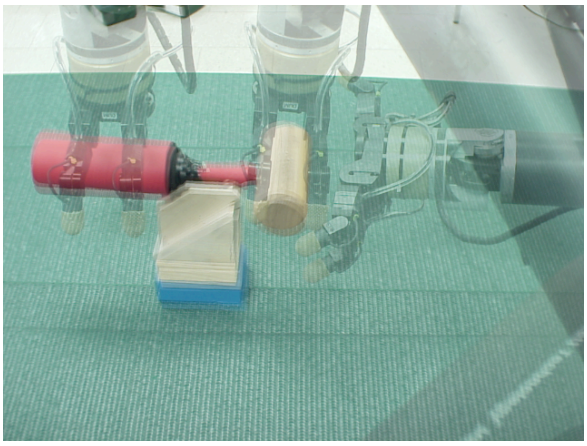

Fig. 8. This figure shows a composite image of some of the grasps generated by the system for the mallet.

of objects, based on latent topic models. The model provides a way of summarizing the data provided by a teleoperator in a way that can be applied to new objects.

Note that although we call this an affordance model, there is no notion linking each "affordance" represented in the model to the functions performed on the object. The model provides a sort of proto-affordance, in that the model gives a distribution over the likely places to grasp the object, but a higher level process must be able to choose among them to perform the task at hand. One could imagine a process whereby the functions of an object are associated with features of the object. This could be integrated with the model proposed here, where the affordance model is used to generate possible grasp locations, and based on that set, further processing is performed to associate a function with each proposed grasp location. The robot could then choose a grasp that suited the functional requirements of the task.

This document presents preliminary results using the shared affordance model. A more complete document, suitable for a peer-reviewed publication is forthcoming, and will deal with issues in this presentation. These include more complete experimental results that address the difficulties the model had with generating grasps for certain objects.

APPENDIX

A. Distributions

1) Dirichlet: A sample from a K-dimensional Dirichlet distribution is a point in the $(K-1)$-simplex; thus these samples can represent a discrete distribution over K objects, and has the following probability density using a symmetric parameter:

$$\text{Dirichlet}(\theta \mid \alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} \theta_i^{\alpha-1}, \qquad (19)$$

where $\alpha > 0$, and $\Gamma(x)$ is the Gamma function.

2) Inverse-Wishart: This distribution is parameterized by scale ψ and degrees of freedom u. The scale matrix ψ is a symmetric, positive definite matrix of size $k \times k$ [9]. The likelihood is:

$$p(b \mid \psi, u) = \mathbb{K}^{-1} |\psi|^{u/2} |b|^{-(u+k+1)/2} \times \exp\left(-\frac{1}{2}\text{tr}(\psi b^{-1})\right), \qquad (20)$$

where b is positive definite and the normalization constant is:

$$\mathbb{K} = 2^{uk/2} \pi^{k(k-1)/4} \prod_{i=1}^{k} \Gamma\left(\frac{u+1-i}{2}\right). \qquad (21)$$

3) Dimroth-Watson: This distribution is parameterized by mode q and concentration κ:

$$p(q \mid \hat{q}, \kappa) = \frac{1}{{}_1F_1(\frac{1}{2}; \frac{3}{2}; \kappa)} \exp\left(\kappa (q \cdot \hat{q})^2\right), \qquad (22)$$

where ${}_1F_1(\cdot)$ is a confluent hypergeometric function.

ACKNOWLEDGMENTS

This research was supported under contract numbers ARO W911NF-05-1-0396 and NASA NNJ05HB61A-5710001842.

REFERENCES

[1] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing* (R. Shaw and J. Bransford, eds.), ch. 3, pp. 67–82, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of the* 10^{th} *International Conference on Computer Vision (ICCV)*, IEEE, 2005.

[4] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *Proceedings of the 2005 IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1331–1338, IEEE, October 2005.

[5] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the* 20^{th} *Conference on Uncertainty in Artificial Intelligence (UAI)*, (Banff, Canada), pp. 487–494, AUAI Press, 2004.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[7] A. Saxena, J. Driemeyer, J. Kearns, C. Osondu, and A. Y. Ng, "Learning to grasp novel objects using vision," in *Proceedings of the* 10^{th} *International Symposium on Experimental Robotics (ISER)*, (Rio de Janeiro, Brazil), July 2006.

[8] R. Platt, *Learning and Generalizing Control Based Grasping and Manipulation Skills*. PhD thesis, University of Massachusetts Amherst, Amherst, MA, September 2006.
[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Texts in Statistical Science, Chapman & Hall/CRC, second ed., 2004.
[10] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, Ltd., second ed., 2000.
[11] C. de Granville, J. Southerland, and A. H. Fagg, "Learning grasp affordances through human demonstration," in *Proceedings of the International Conference on Development and Learning (ICDL)*, 2006.
[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.
[13] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scences using transformed dirichlet processes," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2005.
[14] J. Coelho, *Multifingered Grasping: Haptic Reflexes and Control Context*. PhD thesis, University of Massachusetts, Amherst, MA, September 2001.