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 preshapes of a set of objects. Observations provided by teleoperation are clustered into latent affordances shared among all objects. Each affordance defines a joint distribution over position and orientation of the hand relative to the object and conditioned on visual appearance. The parameters of the model are learned using a Gibbs sampling method. After training, the model can be used to compute grasp preshapes for a novel object based on its visual appearance. The model is evaluated experimentally on a set of objects for its ability to generate grasp preshapes that lead to successful grasps, and compared to a baseline approach.**

I. Introduction

For robots performing object manipulation tasks, affordances provide a useful means of describing how the robot can interact with objects [1]. A grasp affordance is a way of grasping an object to achieve a particular function, and is an active area of research within the neuroscience community [2], [3]. For example, a coffee mug has at least two distinct grasp affordances: one for drinking (typically by using the handle), and another for transporting. The physical characteristics of an object, e.g. visual appearance, provide a way of inferring its affordances. Further, affordances 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 to learn grasp affordance preshapes—the pose of the hand and fingers relative to the object just prior to initiating a grasping action—from demonstration data, and use those affordances to generate preshape hypotheses for novel objects based on visual appearance. We model grasping strategies based on a set of affordances common to all objects using a form of statistical topic model. Topic models, traditionally used in information retrieval, model the co-occurrence of words in a text corpus. Each document is represented as a collection of latent "topics," where each topic is a multinomial distribution over the words in the vocabulary [4].

Topic models are known as *generative* models because they represent the joint probability distribution over documents and topics. The generative process demonstrates how the joint distribution is modeled, and illustrates how a synthetic observation can be sampled from the model. As an example, to generate a single word using a topic model, a topic is sampled from a document-specific distribution. A word is then sampled from the vocabulary given the distribution specified by the chosen topic. This process is repeated for all the words in the document. Thus a document is made up of words sampled from a mixture of topics.

In our model, each "document" is an object presented to the robot. Each "word" is a single preshape of a successful grasp demonstrated on the object and represented by the position and orientation of the hand with respect to the object's centroid. The latent "topics" are the grasp affordances from which the preshapes are drawn. It is helpful to precisely define what we mean by a grasp affordance: a joint distribution over the position and orientation of the hand relative to the object, implied by the object's visual appearance. A teleoperator demonstrates grasp preshapes to the robot, and each of these observations is modeled as a sample from a (latent) affordance of the object.

We use a tuple of parametric distributions to represent an affordance, explained in Section III. A probabilistic representation is used because demonstration via teleoperation is noisy, and signal quality can have a high variance, even for simple grasping actions. In the context of grasp preshape creation, generative models are preferred to discriminative ones because preshapes can be sampled directly from the model. Furthermore, in many cases generative models converge more quickly using less data than discriminative models, which is desirable in this domain, given the cost of acquiring demonstration data [5].

To learn the parameters of the topic model, we cluster all the observations into A groups, where A is chosen arbitrarily. Each cluster is a collection of preshapes

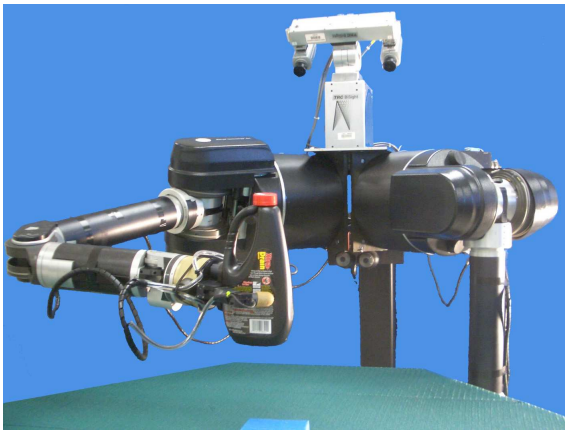

Fig. 1. This picture shows Dexter performing a grasp of the `black_drano` object, as described in Section VI. Each of Dexter's arms has 7 DOF and is equipped with a three-fingered hand with four total degrees of freedom. The stereo head has four degrees of freedom.

that can be thought of as being sampled from a single affordance, since an affordance is a joint distribution. The parameters of the distribution are inferred from the observations in the cluster using Bayesian techniques. A cluster may be made up of preshapes from several objects, thus the affordance it represents is in effect shared between those objects. The result of learning the model is a set of parameters for the distributions representing each of the A affordances.

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

Statistical topic models were originally developed in the information retrieval community for modeling documents in a corpus. The author-topic model [6] and Latent Dirichlet Allocation (LDA) [4] are two examples of this approach for discovering latent topics. Griffiths and Steyvers [7] were the first to propose using Gibbs sampling to learn the parameters of the model, which is the approach used in this work.

Topic models have also been applied in the vision domain for object recognition and classification tasks [8]. This paper is influenced in particular by the hierarchical, part-based model of Sudderth et al. [9]. 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 set of 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 a full set of features, and instead use only visual appearance to infer affordances. Unlike their model, we use the affordances learned by our model to generate new grasps on novel objects.

There has been other work on generating grasps using visual information. Saxena et al. [10] computes a grasping point for an object by analyzing visual features. Their model performs a logistic regression that estimates likely grasp positions in a 2D image based on the observed features. Our work shares a similar motivation, to grasp objects autonomously using vision, but the difference is that we are computing grasp preshapes, and we explicitly model multiple pre-grasp hypotheses for each object.

Platt [11] describes a scheme for generating pre-grasp hypotheses based on observations of the first and second moments of the foreground blob segment. We use similar visual features in this work, but we generate pre-grasp hypotheses from models 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 parametric probability distributions: the visual appearance of the object, the hand's position, and the hand's orientation just prior to grasping. These distributions are described in this section. We assume that our data set consists of M different objects with N_m grasp preshape examples for object m.

A. Visual Appearance

With each affordance we associate a probability distribution over the space of visual appearance features. This distribution is used to relate object appearance to grasp affordances; the robot should be more likely to use a particular affordance to generate preshapes if its visual appearance distribution gives high likelihood to an object's appearance.

To compute the visual appearance, we first segment the object using background subtraction. The object's centroid $\hat{O}_m \in \mathbb{R}^3$ is computed by performing a stereo triangulation on the first moment of foreground blobs in the left and right image planes. The visual feature is the average second moment of the left and right blobs, represented as a covariance matrix b_m. Since the feature space is covariance matrices, we use a two-dimensional inverse-Wishart distribution, parameterized by scale ψ with u degrees of freedom:

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

This distribution is unimodal and typically used as a prior for multivariate covariance matrices [12]. After our

model has been learned, each affordance i will have a set of parameters (ψ_i, u_i) used to compute the likelihood of an appearance according to (1).

This type of feature provides a proof of concept for the model, and other types of visual features can be used. For example, in other work, a multinomial distribution over a set of "visual words" is used to allow each topic to have many likely appearances [8], [9]. The main requirement is that there must be a probability distribution to describe the likelihood of features in an affordance.

B. Hand Position and Orientation

Each grasp affordance describes a subspace in the hand's twist space: every pose that allows the hand to successfully grasp the object for a specific task. The shape of this subspace is determined by the geometries of the hand and object. We approximate this region using a probability distribution, and the goal of our model is to infer the parameters of this distribution for each affordance.

An object can have multiple affordances: each one describes a different way of grasping the object. Although the purpose of the grasp is an integral part of the notion of an affordance as a functional relationship, we are interested in merely establishing a successful grasp. Our model provides preshapes that describe typical grasps seen in the training set; selecting which grasp to use to fulfill task constraints is beyond the scope of this paper.

To approximate an affordance's subspace in $SE(3)$, for computational convenience, instead of using a single distribution, we split the relative hand pose into position and orientation components. We define independent distributions over each of those spaces.

1) Position: The position of an affordance is described as a three-dimensional normal distribution in object-centric space with mean μ and covariance Σ. From each training grasp point $p \in \mathbb{R}^6$, we model the position of the hand $x_{mp} \in \mathbb{R}^3$ in a frame centered at the centroid of the object, $\hat{O}_m$. The likelihood of that hand position for affordance i is:

$$p(x_{mp} \mid \mu_i, \Sigma_i) = \mathcal{N}(x_{mp} \mid \mu_i, \Sigma_i). \tag{2}$$

2) Orientation: We represent the hand orientation of an affordance by using a discrete distribution over a set of Q canonical orientations. Although one can define continuous distributions over $SO(3)$ (cf. [13], [14]), we found that a discrete distribution was a simpler approach that produced satisfactory results. This technique is justified in our problem domain of grasping household objects presented on a table, as we found that a small set of orientations could be used to describe a large number of example grasps.

Let w_{mp} be the canonical orientation that most closely matches the rotational component of the relative pose p. Each affordance i defines a discrete distribution over the set of canonical orientations:

$$p(w_{mp} \mid \phi_i) = \phi_i(w_{mp}), \tag{3}$$

where ϕ_i is a Q-vector in the $(Q-1)$-simplex such that $\phi_i(q)$ is the probability of selecting orientation q.

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 consists of both types of grasps, then the affordance which 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.

After learning the model, each affordance i has a set of parameters $(\mu_i, \Sigma_i, \phi_i)$ for its position and orientation distributions. We can generate a preshape from that affordance by sampling from the distributions in (2) and (3) using those parameters, described in Section V-A.

IV. The Generative Model

In the previous section, we described how an affordance is represented as a collection of three parametric distributions. The topic model presented in this section describes the statistical relationship between the parameters for a set of A affordances and the observed training data.

The model is illustrated in Figure 2. The nodes of the graph represent random variables, with the shaded nodes denoting the observed variables. Unshaded nodes are latent variables that must be inferred from the data. Rectangles around variables denote replication, where the number of times is shown in the bottom right corner.

Our model assigns one affordance to each data observation; after all observations have been assigned, we compute the values of the affordances' parameters. We take a Bayesian approach to estimate these parameters because we can quantify our uncertainty about their values by using suitable prior distributions. Moreover, we can incorporate new data to improve our posterior estimates.

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. Datum i of object m is the tuple $(b_{mi}, x_{mi}, w_{mi},)$, which consists of a visual feature

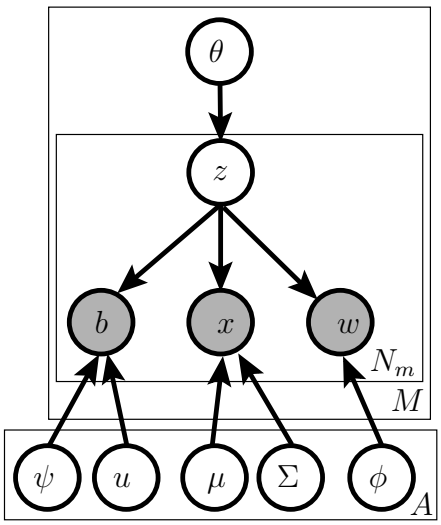

Fig. 2. The graphical model described in Section IV. Circles indicate random variables, shading indicates an observed variable, otherwise they are latent. A rectangle around nodes represents replication; the number of times written in the bottom right corner. The edges between nodes indicates a conditional probability distribution described in the text.

covariance matrix, the position, and the orientation of the hand, respectively. The variable z represents the assignment of an affordance to the observed preshape. The variable θ describes a multinomial distribution over the shared set of affordances for an object; in effect, it defines a mixture model over affordances and describes the likelihood of an object using a particular affordance. 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.

Using this model, the generative process for data point i is given below:

$$\begin{aligned} \theta \,|\, \alpha &\sim \text{Dirichlet}(\alpha) \\ z_i \,|\, \theta &\sim \text{Multinomial}(\theta) \\ b_i \,|\, z_i = j &\sim \text{Inv-Wishart}_{u_j}(\psi_j) \\ w_i \,|\, z_i = j &\sim \text{Multinomial}(\phi_j) \\ x_i \,|\, z_i = j &\sim \mathcal{N}(\mu_j, \Sigma_j), \end{aligned} \tag{4}$$

where $X \sim \mathbb{D}$ means that random variable X is sampled from distribution $\mathbb{D}$. The first line of (4) samples θ from a symmetric, A-dimensional Dirichlet prior with parameter α. In the second line, the affordance assignment z_i for this datum is sampled according to θ. The preshape components b_i, w_i, and x_i are sampled according to the distributions (1), (3), and (2), described in Section III, using the parameters of affordance z_i.

We assume independent, symmetric Dirichlet priors over θ and ϕ, with hyperparameters α and β, respectively. The second moment 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 [12]. The grasp position mean is given a uniform prior.

For notational convenience, let Ω correspond to the parameters of these priors, the so-called hyperparameters, and let $\mathbb{C} = (\psi, u, \phi, \mu, \Sigma)$ correspond to the parameters of the component distributions for each affordance.

V. Parameter Estimation in the Model

The inference problem is to compute the posterior distribution of the latent variables given example grasp points, using Bayes' rule:

$$p(\theta, \mathbf{z}, \mathbb{C} \,|\, \mathcal{D}, \Omega) = \frac{p(\mathcal{D} \,|\, \theta, \mathbf{z}, \mathbb{C}, \Omega) p(\theta, \mathbf{z}, \mathbb{C} \,|\, \Omega)}{p(\mathcal{D} \,|\, \Omega)}, \tag{5}$$

which is intractable, although we can estimate it using Gibbs sampling. Gibbs sampling is used when it is impossible to sample from a distribution directly. Instead, we sample from each dimension of the distribution conditioned on the current state of the rest of the dimensions. In this case, the distribution we are interested in is the posterior assignment of affordances to data points.

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 $\mathbb{C}$.

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

$$\begin{aligned} p(z_{mi} \,|\, \mathbf{z}_{-mi}, \mathcal{D}) \propto\; & p(z_{mi} \,|\, \mathbf{z}_{-mi}, o_m) \\ & \times p(b_{mi} \,|\, \mathbf{z}, \mathbf{b}_{-mi}) p(x_{mi} \,|\, \mathbf{z}, \mathbf{x}_{-mi}) \\ & \times p(w_{mi} \,|\, \mathbf{z}, \mathbf{w}_{-mi}). \end{aligned} \tag{6}$$

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

$$p(z_{mi} = j | \mathbf{z}_{-mi}, o_m = l) = \frac{n^O_{jl} + \alpha}{\sum_{j'} n^O_{jl'} + A\alpha} \tag{7}$$

$$p(w_{mi} = k | z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{w}_{-mi}) = \frac{n^W_{kj} + \beta}{\sum_{j'} n^W_{kj'} + Q\beta}, \quad (8)$$

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. Since the assignment of affordances to observations is a statistical process, if A is large there may be affordances that are not assigned to any observations. The expected number of affordances used is a function of the number of observations and α.

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 [12]:

$$p(x_{mi} \,|\, z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{x}_{-mi}) \approx \mathcal{N}(x_{mi} \,|\, \hat{\mu}_j, \hat{\Sigma}_j), \quad (9)$$

where

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

$$\delta_j = \frac{n^A_j + 1}{n^A_j (n^A_j + \nu - 4)}$$

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

The conditional distribution for a visual feature covariance is given as

$$p(b_{mi} \,|\, z_{mi} = j, \mathbf{z}_{-mi}, \mathbf{b}_{-mi}) = \text{Inv-Wishart}_{\hat{u}_j}(\hat{\psi}_j) \quad (10)$$

with

$$\hat{u}_j = u_0 + n^A_j$$

$$\hat{\psi}_j = \frac{1}{n^A_j} \left(\Psi_0 + \sum_{m=1}^{M} \sum_{k | z_{mk} = j} b_{mk} \right). \quad (11)$$

At each iteration of the Gibbs sampler, we use (7) – (10) to compute (6). 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 $O(MNA)$. The sampler must be run for a number of iterations before samples can be considered independent. We typically compute on the order of two hundred "burn-in" iterations before accepting a sample.

A. Generating preshapes for new objects

We are interested in generating candidate preshapes for a novel object given its visual features. Let $\hat{\Theta}^{(s)}$ correspond to the model parameters estimated from sample s. The generative process for new grasps given visual feature b_t is:

$$\begin{aligned} z_t \,|\, b_t, \hat{\Theta}^{(s)} &\sim p(z \,|\, b_t, \hat{\Theta}^{(s)}) \\ w_t \,|\, z_t = j, \hat{\Theta}^{(s)} &\sim \text{Multinomial}(\hat{\phi}^{(s)}_j) \\ x_t \,|\, z_t = j, \hat{\Theta}^{(s)} &\sim \mathcal{N}(\hat{\mu}^{(s)}_j, \hat{\Sigma}^{(s)}_j). \end{aligned} \quad (12)$$

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 $\hat{\Theta}^{(s)}$ we can estimate θ and $\mathbb{C}$ using the affordance assignments in $\mathbf{z}^{(s)}$ as described in Section V using (7) – (10). These correspond to predictive distributions over new affordances and hand 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 (12) can be computed as

$$\begin{aligned} p(z = i \,|\, b_t, \hat{\Theta}^{(s)}) &\propto p(b_t \,|\, z = i, \hat{\Theta}^{(s)}) p(z = i \,|\, \hat{\Theta}^{(s)}) \\ &\approx \text{Inv-Wishart}_{\hat{u}^{(s)}_i}(\hat{\psi}^{(s)}_i), \end{aligned} \quad (13)$$

where we assume that $p(z = i \,|\, \hat{\Theta})$ is uniform.

By following the generative process in (12), given a visual feature, we can produce a set of candidate preshapes. In this work, we assume a fixed configuration of the fingers in the hand, such that they form an opposing grasp. One could easy augment the affordance representation to take into account different finger configurations.

VI. Experimental Results

To test the ability of the model to represent the grasp affordances demonstrated in the training set and generate new pre-grasps, we trained the model using a set of household objects. Because there is no notion of orientation of the object in the model, the same object presented in a different orientation (flat, standing up, etc.) is treated as a separate object. 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

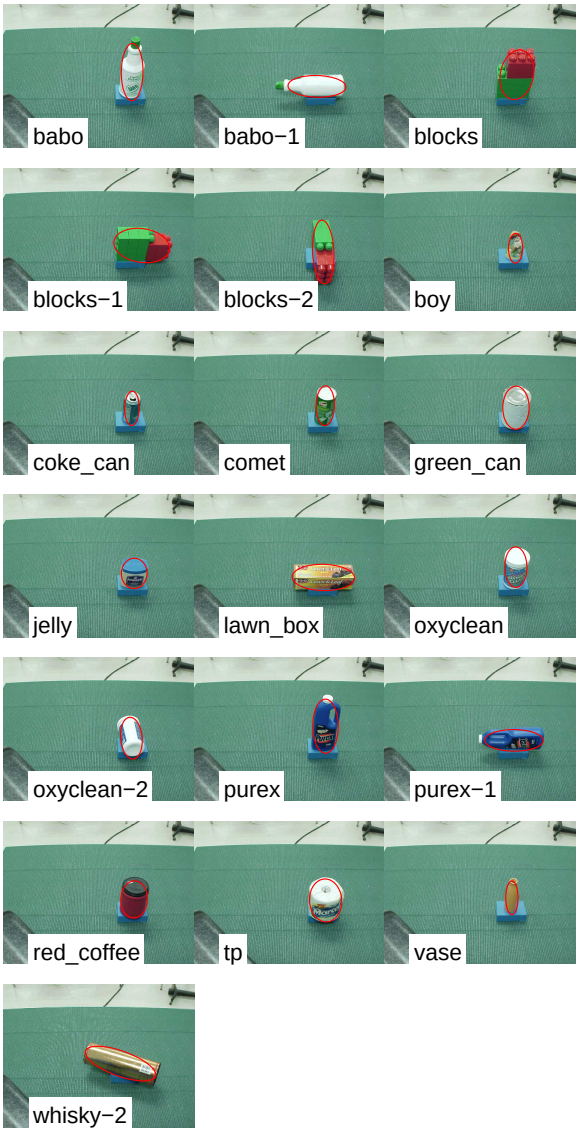

Fig. 3. This picture shows the objects in the training set. The red oval corresponds to the covariance matrix that was used as a visual feature for grasps with the object.

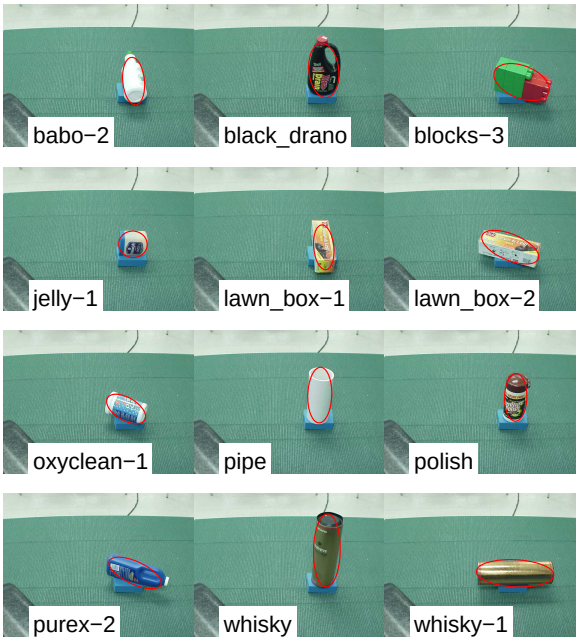

Fig. 4. This picture shows the objects as they were presented for generating grasps.The red oval corresponds to the covariance matrix that was computed from the average second moments of the segmented blob in the left and right cameras.

itself [15]. We chose a set $\mathcal{O}$ of 31 object presentations for training and testing.

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. 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. While the model specifically generated preshapes for Dexter's right arm, by applying a known affine transformation to the preshape, they can be used by the left arm. The set of canonical grasp orientations was computed using the training set, and a set of $Q = 6$ 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.

For learning the parameters of the model, $A = 10$ shared grasp affordances were used. The Gibbs sampler ran for 200 iterations of burn-in, after which the next sample was stored. Using the single sample, $N_{test} = 12$ objects were presented, shown in Figure 4, and the model generated 6 candidate preshapes for each object. The robot achieved each preshape configuration and then attempted to grasp the object. To perform the grasp, the robot simply flexed its fingers until a sufficient force had been applied to the object. In these experiments a grasp was judged successful if the robot was still holding onto the object after moving the hand 10 cm vertically.

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.

A. *The naïve model*

To analyze the performance of the model, we created a naïve model which also generated grasps using visual features. This model performed visual processing to estimate the width and height of the object, and then generated grasps by selecting points on a spherical

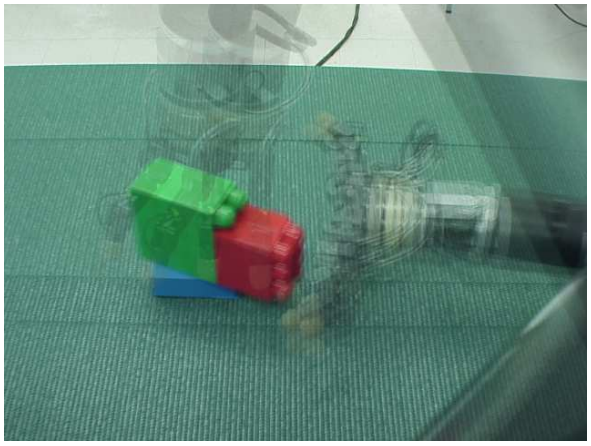

Fig. 5. A composite image showing six candidate grasp positions for the `blocks-3` object.

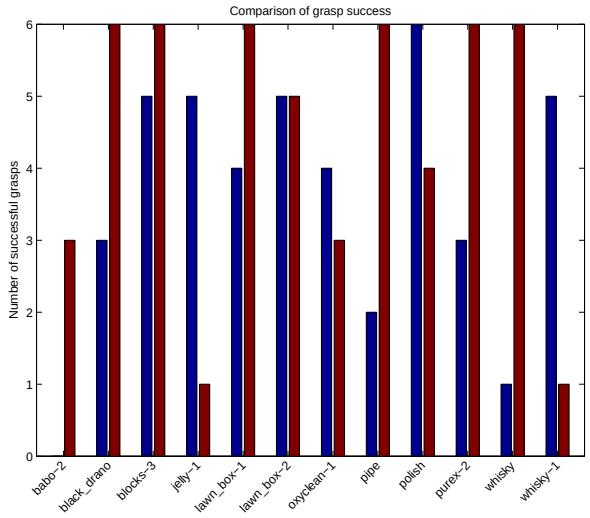

Fig. 6. This graph shows the result of using the trained grasp model on a set of test objects. Each bar measures the number of successful grasps for the labeled object. The blue bars are for the naïve model, and the red for the shared affordance model.

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 a ray connecting it to the object's centroid, and 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 starting from six different locations, and grasp success was judged as before.

The results of performing these grasps are shown in Figure 6, where blue and red bars correspond to the naïve and our 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 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 cases, although the generated preshapes 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 preshapes provided in the demonstration. For novel objects, the model finds the affordance with the most similar appearance, but the hand positions suggested by that affordance may not adequately fit the actual geometries of the object. To improve performance, one could incorporate a grasp controller to perform the grasp once the preshape was achieved [16].

Since we use a statistical model, the candidate preshapes generated by an affordance may vary in quality, and in these experiments, each candidate preshape was attempted regardless of its quality. However, as the amount of training data increases, the expected variance of the affordance distributions will decrease, potentially improving performance. In a real-world scenario the model could be used interactively, with the teleoperator providing additional training data to improve the quality of the robot's hypotheses.

Additionally, the proposed method could be improved by performing a secondary analysis of the candidate pregrasps. For example, incorporating additional information about the object geometry into candidate selection to choose the closest, non-colliding preshape.

The success rate of the model is also affected by the number of shared affordances. In the current implementation we estimate A based on the number of objects presented, although we do not know a priori the number of shared affordances represented in the training data. If A is too small, the covariances for the position distribution of the affordances will be large, so it may require sampling a number of preshapes before finding one that is close enough for a successful grasp. Nonparametric Bayesian approaches can be used to estimate the number of affordances from the data itself [17].

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"

	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.

indicates that some training grasp from this object was 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 $\hat{\Theta}^{(s)}$ to contain very similar assignments. Different runs of the sampler produce similar groupings of observations, although the actual assignments to particular affordances will differ (e.g., the assignment found in affordance 1 in this sample may be the assignment of affordance 5 in another sample).

B. More Complex Objects

Using this visual feature model, one can represent a single object with multiple second moment covariances: the model will generate grasps for each covariance, and they must then be transformed into the appropriate frame. Since the model has no notion of the geometry of the object beyond centroid and second moment, secondary processing should be used to filter low-quality preshapes. As an example, we presented a mallet that was segmented into two blobs, as shown in Figure 7.

Using the model learned in the previous section, we generated preshapes for each of the two blobs and manually filtered the candidates that collided with the mallet. For example, the model generated preshapes for

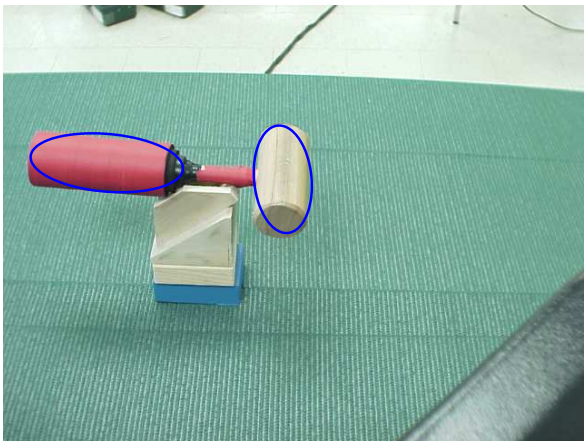

Fig. 7. This figure shows how the mallet can be segmented into multiple blobs, and each blob can be used to generate grasp positions independently.

Fig. 8. This figure shows a composite image of some of the preshapes generated by the model learned in Section VI.

side grasps of the handle that would collide with the head of the mallet. Figure 8 shows some feasible candidate preshapes suggested by the model.

VII. CONCLUSIONS

We have presented a hierarchical, statistical model for representing grasp preshapes among a collection of objects, using a latent topic model. The model provides a way of summarizing the data provided by a teleoperator in a way that can be applied to new objects. We showed that the model can generate successful grasp preshapes on novel objects and outperforms a naïve strategy.

For future work, different visual features could be used to learn affordances specific to smaller scale features of objects. As mentioned above, the model can also be improved by incorporating rigid-body transformations

into the representation of objects. Ideally, a model of grasp affordances is learned for a canonical orientation of an object, and preshapes from that affordance are transformed to match the orientation of the object as it is presented. Additionally, the model can be combined with higher-level logic that selects grasp candidates based on task constraints.

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] A. H. Fagg and M. A. Arbib, "Modeling parietal-premotor interactions in primate control of grasping," *Neural Networks*, vol. 11, pp. 1277–1303, 1998.

[3] M. A. Arbib, A. Billard, M. Iacoboni, and E. Oztop, "Synthetic brain imaging: grasping, mirror neurons and imitation," *Neural Networks*, vol. 13, pp. 975–997, 2000.

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

[5] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] R. Platt, *Learning and Generalizing Control Based Grasping and Manipulation Skills*. PhD thesis, University of Massachusetts Amherst, Amherst, MA, September 2006.

[12] 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.

[13] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, Ltd., second ed., 2000.

[14] 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.

[15] 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.

[16] J. Coelho, *Multifingered Grasping: Haptic Reflexes and Control Context*. PhD thesis, University of Massachusetts, Amherst, MA, September 2001.

[17] 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.