Autonomous Robots manuscript No.
(will be inserted by the editor)

Learning Grasp Affordances Through Human Demonstration

Charles de Granville · Andrew H. Fagg

Received: date / Accepted: date

Abstract When presented with an object to be manipulated, a robot must identify the available forms of interaction. How might an agent acquire this mapping from object representation to action? In this paper, we describe an approach that learns a mapping from objects to grasps from human demonstration. For a given object, the teacher demonstrates a set of feasible grasps. We cluster these grasps in terms of the position and orientation of the hand relative to the object. Individual clusters in this pose space are represented using probability density functions, and thus correspond to variations around canonical grasp approaches. Multiple clusters are captured through a mixture distribution-based representation. Experimental results demonstrate the feasibility of extracting a compact set of canonical grasps from the human demonstration. Each of these canonical grasps can then be used to parameterize a reach controller that brings the robot hand into a specific spatial relationship with the object.

Keywords Grasp affordance · learning from demonstration · clustering · mixture models · probabilistic densities of 3D rotations

1 Introduction

Gibson (1966, 1977) proposed that objects in the environment can be represented by an agent in terms of the actions that can be performed with those objects. This *affordance* representation captures a combination of the *interaction-relevant* physical properties of the object and the capabilities of the agent's own body. Furthermore, Gibson suggested that this representation should be distinct from one that explicitly captures the semantics of the objects. Instead, an affordance representation should provide a detailed, task-neutral "menu" of possible actions that can be taken by the agent. Thus, this knowledge of agent-object interaction is available whether the task is well known to the agent or one that the agent is just learning to perform.

One important form of interaction is that of grasping. For a given object, how might an agent come to represent the set of feasible grasps that may be made? Ultimately, one must establish a mapping from perceivable features to a set of parameterized grasping actions (specific positions and orientations for the hand, as well as configurations for the fingers) that are expected to be successful if executed. We would like for this set to be small so as to facilitate both planning and exploratory learning.

One approach is to establish a *direct* mapping between the perceived features of objects and the grasps available to the agent. Coelho, Piater, and Grupen developed an approach that automatically learns a direct mapping from constellations of visual features to hand orientations and finger configurations in a planar grasping task (Coelho et al. 2000; Piater and Grupen 2002). Given a set of object images and corresponding quality grasp configurations, the visual learning algorithm attempts to find sets of geometrically-arranged image features that consistently predict the relative orientation of the hand and the associated finger configurations. In novel situations these affordance maps can then be used to position and configure the hand in such a way that a successful grasp is probable.

Charles de Granville
E-mail: chazz184@ou.edu

Andrew H. Fagg (corresponding author)
E-mail: fagg@ou.edu

Present address: Symbiotic Computing Laboratory; School of Computer Science; University of Oklahoma; Norman, OK

Similarly, Sweeney and Grupen (2007) present an approach that learns a collection of shared grasp affordances for a set of objects. Each affordance is represented as a joint probability distribution over coarse visual features, hand position, and hand orientation. Successful examples of grasps are demonstrated by a human teleoperator. When the visual features exhibited by a novel object are similar to those in the training set, the learned affordance map can be used to derive actions that are likely to result in a successful grasp.

One challenge that arises out of taking a direct approach to learning grasp affordances is that the appearance of an object may change as a function of its pose even though the set of grasps that may be made relative to the object does not. One possible solution to this problem is to establish an *indirect* mapping from sensory features to grasps through an intermediate object-centered representation. This approach separates the problem into one of first estimating an object's pose and identity (and/or shape), and then estimating the set of feasible configurations of the hand relative to the object.

Stoytchev (2005) presents a developmental approach to the latter of these two problems. This approach discovers sequences of actions that result in a successful "binding" of the object with the robot (i.e., a grasp). For each of several objects, the robot performs a random sequence of exploratory actions. Subsequences of actions that lead to simultaneous movement of the object and a component of the robot are deemed as "interesting." Short subsequences that reliably achieve this interesting bound configuration are identified as viable grasping actions, and are associated with the object through the affordance map. Hence, when a known object is subsequently presented to the robot, it is able to generate a sequence of actions that will likely lead to a successful grasp.

De Granville et al (2006) present a technique for learning the canonical hand orientations that can be used to grasp specific objects. Given a set of demonstrated grasps by a human teacher, a small number of canonical hand approach orientations are identified through a clustering process on S3 (the 4D unit hypersphere). Assuming a similarity in the morphology between the human teacher and the robot, the learned clusters represent a set of hand approach orientations that can lead to feasible grasps.

In contrast to making use of object identity, Bekey et al (1993) and Miller et al (2003) rely on a shape-based description of an object. Objects are modeled as a collection of shape primitives such as cylinders, rectangular prisms, and cones. Each primitive is associated with a set of relative hand poses and corresponding finger configurations. Given a set of primitive shapes that describe a novel object, the planner can quickly generate a set of candidate grasps. In Miller's case, each candidate grasp is executed in simulation, and evaluated in terms of the amount of force required to dislodge the object from the hand. In the case of Bekey et al., the set of candidates is evaluated as a function of the semantics of the grasp and task. For example, when using a wrench to turn a nut, grasps that provide a large amount of torque are preferable.

In this paper, we extend the approach of de Granville et al by constructing a more complete grasp affordance representation. In particular, we focus on the problem of describing the position *and* orientation of the hand as it approaches the object. Individual clusters are represented by a joint probability distribution over position and orientation. In our experiments, a single demonstration trial consists of a human teacher haptically exploring an object, pausing briefly in configurations that correspond to quality grasps. Experimental results demonstrate the feasibility of extracting a compact set of canonical grasps from the human demonstration. The learned representation can then be used to parameterize controllers that are capable of driving a hand to an appropriate pose for grasping, or to interpret the actions of other agents in the environment.

2 Representing Grasp Affordances

Having a compact representation that describes the ways in which an object may be grasped greatly improves an agent's ability to efficiently plan and execute grasping actions. Our goal is to compress a large number of examples provided by a human teacher into a small number of clusters that are meaningful in terms of describing the functionally different ways that an object may be grasped. Our approach represents each of these clusters as a probability density function (PDF) defined over both orientation and position. A set of clusters is then captured using a mixture model approach.

2.1 Modeling Hand Orientation

Unit quaternions are a natural representation of 3D orientation because they comprise a proper metric space, a property that allows us to compute measures of similarity between pairs of orientations. Here, an orientation is represented as a point on the surface of a 4D unit hypersphere. This representation is also antipodally symmetric: pairs of points that fall on opposite poles represent the same 3D orientation. The Dimroth-Watson distribution captures a Gaussian-like shape on the unit hypersphere, while explicitly acknowledging this symmetry (Mardia and Jupp 1999; Rancourt et al 2000). The probability density function for this distribution is as follows:

$$f\left(\mathbf{q}|\mathbf{u},k\right)=F\left(k\right)e^{k\left(\mathbf{q}^{T}\mathbf{u}\right)^{2}},\tag{1}$$

where $\mathbf{q}\in\mathbb{R}^4$ represents a unit quaternion, $\mathbf{u}\in\mathbb{R}^4$ is a unit vector that represents the "mean" rotation, $k\geq 0$ is a concentration parameter, and $F(k)$ is a normalization term that is given in the appendix. Note that $\mathbf{q}^T\mathbf{u}=\cos\theta$, where θ is the angle between $\mathbf{q}$ and $\mathbf{u}$. Hence, density is maximal when $\mathbf{q}$ and $\mathbf{u}$ are aligned, and decreases exponentially as $\cos\theta$ decreases. When $k=0$, the distribution is uniform across all rotations; as k increases, the distribution concentrates about $\mathbf{u}$. Figure 1(a) shows a 3D visualization of the Dimroth-Watson distribution, and highlights its Gaussian-like characteristics. The high density peaks correspond to $\mathbf{u}$ and $-\mathbf{u}$.

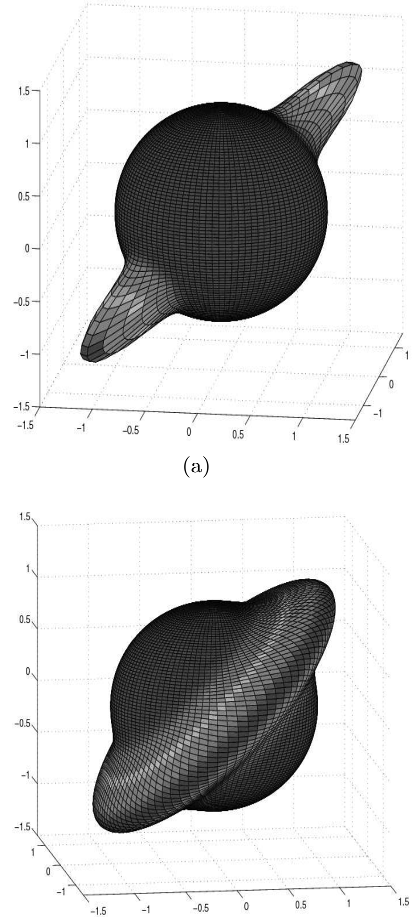
A second cluster type of interest corresponds to the case in which an object exhibits a rotational symmetry. For example, an object such as a cylinder can be approached from any orientation in which the palm of the hand is parallel to the planar face of the cylinder. In this case, hand orientation is constrained in two dimensions, but the third is unconstrained. This set of hand orientations corresponds to an arbitrary rotation about a fixed axis, and is described by a great circle (or girdle) on the 4D hypersphere. We model this set using a generalization of the Dimroth-Watson distribution that was suggested by Rivest (2001). The probability density function is as follows:

$$\bar{f}\left(\mathbf{q}|\mathbf{u}_1,\mathbf{u}_2,k\right)=\bar{F}\left(k\right)e^{k\left[\left(\mathbf{q}^{T}\mathbf{u}_1\right)^{2}+\left(\mathbf{q}^{T}\mathbf{u}_2\right)^{2}\right]},\tag{2}$$

where $\mathbf{u}_1\in\mathbb{R}^4$ and $\mathbf{u}_2\in\mathbb{R}^4$ are orthogonal unit vectors that determine the great circle, and $\bar{F}(k)$ is the corresponding normalization term that is derived in the appendix. Figure 1(b) illustrates the girdle distribution on S2 (the 3D unit sphere). First, note that all points on the great circle are assigned maximal density. This corresponds to the set of points for which $\left(\mathbf{q}^T\mathbf{u}_1\right)^2+\left(\mathbf{q}^T\mathbf{u}_2\right)^2=1$. However, as the angle between $\mathbf{q}$ and the closest point on the circle increases, the density decreases exponentially.

For a given set of observations, the parameters of the Dimroth-Watson and girdle distributions are estimated using maximum likelihood estimation (MLE). The axes of the distribution are derived from the sample covariance matrix, $\mathbf{\Lambda}\in\mathbb{R}^{4\times 4}$:

$$\mathbf{\Lambda}=\frac{\sum_{i=1}^{N}\mathbf{q}_i\mathbf{q}_i^T}{N},\tag{3}$$



(a)

(b)

Fig. 1 Three dimensional representations of the Dimroth-Watson (a) and girdle (b) distributions on S2. In both cases, the surface radius is $1+p$, where p is the probability density at the corresponding orientation

where $\mathbf{q}_i$ is the orientation of the ith sample, and N is the total number of samples. The MLE of $\mathbf{u}$ is parallel to the first eigenvector of $\mathbf{\Lambda}$ (Mardia and Jupp 1999; Rancourt et al 2000). The orthogonal vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ span the same space as the first and second eigenvectors of $\mathbf{\Lambda}$ (Rivest 2001).

For the Dimroth-Watson distribution, the MLE of the concentration parameter, k, uniquely satisfies the following (see de Granville (2008) for the derivation):

$$G\left(k\right)\equiv\frac{F'\left(k\right)}{F\left(k\right)}=-\frac{\sum_{i=1}^{N}\left(\mathbf{q}_i^T\mathbf{u}\right)^2}{N}.\tag{4}$$

In the case of the girdle distribution, the MLE of k uniquely satisfies (see de Granville (2008) for the derivation):

$$\bar{G}\left(k\right)\equiv\frac{\bar{F}'\left(k\right)}{\bar{F}\left(k\right)}=-\frac{\sum_{i=1}^{N}\left[\left(\mathbf{q}_i^T\mathbf{u}_1\right)^2+\left(\mathbf{q}_i^T\mathbf{u}_2\right)^2\right]}{N}.\tag{5}$$

For computational efficiency, we approximate $G^{-1}()$ and $\bar{G}^{-1}()$ when solving for k (see the appendix for details).

2.2 Modeling Hand Position

The position of the hand is represented as a 3D vector in Cartesian space. We choose to model position using a Gaussian distribution:

$$p\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\frac{1}{\left(2\pi\right)^{\frac{d}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{6}$$

Here, $\mathbf{x}\in\mathbb{R}^{d}$ denotes a point in a d dimensional Cartesian space, while $\boldsymbol{\mu}\in\mathbb{R}^{d}$ and $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$ correspond to the mean vector and covariance matrix of the Gaussian distribution. For our purposes, $d=3$, $\boldsymbol{\mu}$ describes the mean position of the hand, and $\boldsymbol{\Sigma}$ captures covariance in hand position.

2.3 Modeling Hand Pose

Hand pose is represented as a joint probability distribution over position and orientation. We assume that the position and orientation of the hand are independent within a single cluster. Since there are two choices for the orientation component (Dimroth-Watson and girdle), the joint distribution takes on one of the following forms:

$$g\left(\mathbf{x},\mathbf{q}|\theta\right)=p\left(\mathbf{x}|\theta_{p}\right)f\left(\mathbf{q}|\theta_{f}\right) \tag{7}$$

and

$$\bar{g}\left(\mathbf{x},\mathbf{q}|\bar{\theta}\right)=p\left(\mathbf{x}|\theta_{p}\right)\bar{f}\left(\mathbf{q}|\theta_{\bar{f}}\right), \tag{8}$$

where θ and $\bar{\theta}$ are the parameters for the two joint distributions, θ_{p} consists of the parameters for the position density, and θ_{f} and $\theta_{\bar{f}}$ are the parameters of a Dimroth-Watson and girdle distributions, respectively.

The $g()$ and $\bar{g}()$ density functions essentially encode two different grasp types. The first constrains all six degrees of freedom of hand pose, while the second constrains all but one rotational degree of freedom.

2.4 Mixtures of Hand Pose Models

An individual hand pose distribution can capture a single cluster of points, but a set of grasps is typically fit best by multiple clusters. Furthermore, the use of multiple clusters captures any covariance that may exist between the position and orientation of the hand when grasping a particular object. We therefore employ a mixture model-based approach. Here, the density function of the mixture, $h\left(\right)$, is defined as:

$$h(\mathbf{x},\mathbf{q}|\boldsymbol{\Psi})=\sum_{j=1}^{M}w_{j}c_{j}(\mathbf{x},\mathbf{q}|\theta_{j}), \tag{9}$$

$$\boldsymbol{\Psi}=(w_{1},...,w_{M},\theta_{1},...,\theta_{M}), \tag{10}$$

and

$$\sum_{j=1}^{M}w_{j}=1, \tag{11}$$

where M denotes the number of component densities, and c_{j} is one of the two density functions describing hand pose ($g\left(\right)$ or $\bar{g}\left(\right)$). Each element of the mixture represents a single cluster of points, and is weighted by w_{j}. Estimation of the parameters of the individual clusters and the cluster weight variables is accomplished using the Expectation Maximization algorithm (Dempster et al 1977).

3 Learning Grasp Affordances

The previous section presented an approach for representing the grasp affordances of an object by a mixture of parametric distributions over the position and orientation of the hand. This section focuses on the experimental procedures and algorithms employed in this work to learn grasp affordances from human demonstration.

3.1 Data Collection Procedure

A human teacher wears a P5 glove (Essential Reality, Inc.) equipped with a Polhemus Patriot sensor near the wrist (Polhemus, Inc.).[1] These components continuously capture the pose of the hand at $15Hz$. A fixed transformation relative to the wrist is used to estimate the point between the tips of the thumb and index finger. Because the human teacher primarily uses precision grasps, this point is used as a description of hand pose. In addition, a Polhemus sensor is mounted on the object, which allows us to compute the pose of the hand in an object centered coordinate frame. Each trial consists of approximately 5 minutes of haptic exploration of the object. Throughout the course of a trial, the object may be translated and rotated in the global coordinate frame. This allows the human teacher to execute grasps that might not be possible if the object were in a fixed

[1] The experimental protocol was approved by the University of Oklahoma Internal Review Board (IRB #11909).

location of the workspace. During the trial, the teacher largely maintains contact with the object in configurations that correspond to quality grasps, although some samples fall along transitions between valid grasps. After the trial, the observations are subsampled and split into a training set and two validation sets. A training set consists of 1000 samples, each validation set contains 250 samples, and a total of 10 trials are performed for each object.

3.2 Model Selection

For a given set of observations, it is unclear *a priori* how many or of what type of cluster is appropriate. Our approach is to construct all possible mixtures that have a maximum of M clusters (we choose $M = 10$) and to choose the mixture that best matches the observations. For this purpose, we make use of the Integrated Completed Likelihood (ICL) criterion (Biernacki et al 2000) to evaluate and order the different mixture models. Like the Bayesian Information Criterion, ICL prefers models that explain the training data, but punishes more complex models. In addition, ICL punishes models in which clusters overlap one-another. These features help to select models that describe a large number of grasps with a small number of clusters.

Because the EM algorithm is a gradient ascent method in a likelihood space containing many local maxima, each candidate mixture model was fit a total of Ω different times using the available training data (for our purposes, $\Omega = 80$). For a given mixture, this ensures that a variety of different initializations for the EM algorithm are explored. The model that performs best on the first validation set according to ICL is subsequently evaluated and compared with other mixtures using the second validation set (again using ICL).

Due to our data collection procedure, some samples do not correspond to quality grasps, and instead correspond to transitions between grasps. It is desirable that our clustering algorithm be robust to this form of noise. However, when a large enough number of mixture components is allowed, the EM algorithm tends to allocate one or more clusters to this small number of “outlier” samples. We explicitly discard these mixture models when an individual cluster covers a very small percentage of the samples (indicated by a small magnitude cluster weight parameter, w_j). In particular, a model is discarded when:

$$\frac{\max_j(w_j)}{\min_j(w_j)} \geq \lambda, \tag{12}$$

where λ is a threshold. For our experiments, we chose $\lambda = 5$ because it tends to result in the selection of high quality, compact models. Of the models that have not been removed by this filter step, the one with the best ICL measure on the second validation set is considered to be the best explanation of the observed data set.

3.3 Assessing Model Quality

In general, it is difficult to assess the performance of unsupervised learning techniques because there is no inherent notion of a ground truth. Thus, to assess the quality of a model that is produced by our clustering algorithm, we compare the learned clusters to a set of heuristically chosen clusters. Our heuristic is based on knowledge of the object and the types of grasps demonstrated by the human teacher. For example, when grasping the handle of an object, such as the heat gun shown in figure 2(d), a single grasp type is usually employed. The orientation of the hand relative to the object is typically fixed in a configuration orthogonal to the handle’s major axis, and the hand position tends to cluster around the handle’s center. This approach provides a reasonable estimate of the number of clusters that exist for an object, as well as an expectation for each cluster’s shape or type.

Performance of the clustering algorithm is quantified in terms of a contingency table that counts the number of “true” positives (TP), “false” positives (FP), and “false” negatives (FN) present in a solution produced by our clustering algorithm. A true positive is scored when the algorithm identifies a cluster that corresponds to a heuristically derived cluster. A match occurs when the position component of a cluster covers the appropriate region of an object, and the orientation component captures the set of hand orientations used by the human teacher (i.e. whether or not a rotational symmetry exists). A false positive is scored when the algorithm identifies a cluster that does not match the heuristic. This can happen when the algorithm identifies multiple clusters where a single cluster should have been found. A false negative is scored when the algorithm fails to identify one of the heuristically chosen clusters.

Once a contingency table has been constructed for a grasp affordance model its true positive rate ($TPR = TP/(TP + FN)$), precision ($PRC = TP/(TP + FP)$), and false discovery rate ($FDR = FP/(TP + FP)$) are computed. The true positive rate reports the fraction of the desired clusters that were correctly identified, while the precision describes the fraction of correctly identified clusters out of the clusters actually learned by the algorithm. The false discovery rate is a measure of how much our algorithm is overfitting the data.

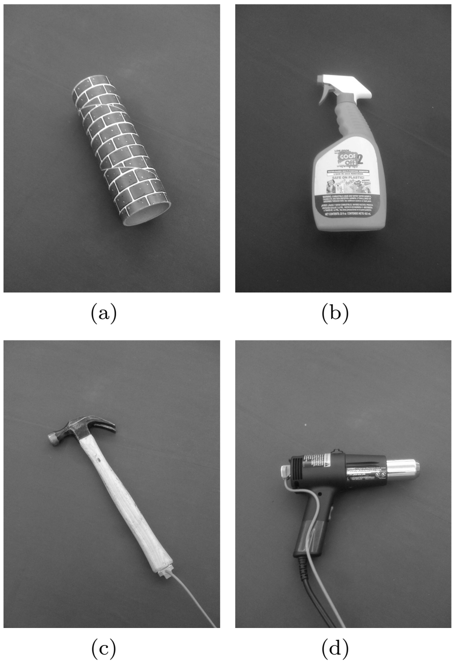
4 Experimental Results

In order to illustrate the capabilities of our clustering approach, we perform multiple grasping experiments using a variety of objects (see Figure 2). Each object has its own unique set of grasps that may be modeled as a mixture of joint distributions over the position and orientation of the hand.

4.1 Cylinder

First, we consider the cylinder shown in figure 2(a). A total of four feasible grasps exist for this object: One for each end, and two along the major axis (corresponding to "overhand" and "underhand" configurations). Figures 3(a) and 3(b) depict a typical set of samples collected for the cylinder. Samples located at the extremes of the X axis correspond to cases in which the palm of the hand is approximately orthogonal to the object's major axis. Intermediate samples correspond to cases in which the hand is exploring the lateral surface of the cylinder in either an overhand or underhand configuration.

In figure 3(a), the 3D position of the hand is shown throughout the course of the experiment, while figure 3(b) provides a visualization of the corresponding hand orientations. Orientation of the hand is represented as a single point on the surface of the unit sphere: imagine that the object is located at the origin of the sphere; the point on the surface of the sphere corresponds to the intersection of the palm with the sphere. Note that this visualization technique aliases the set of rotations about the line perpendicular to the palm. For example, in figure 3(b), there is no way to distinguish grasps using an overhand configuration from those that use an underhand configuration.


(a)

(b)

(c)

(d)

Fig. 2 The set of objects used in the clustering experiments. (a) Cylinder; (b) Spray bottle; (c) Hammer; (d) Heat gun.

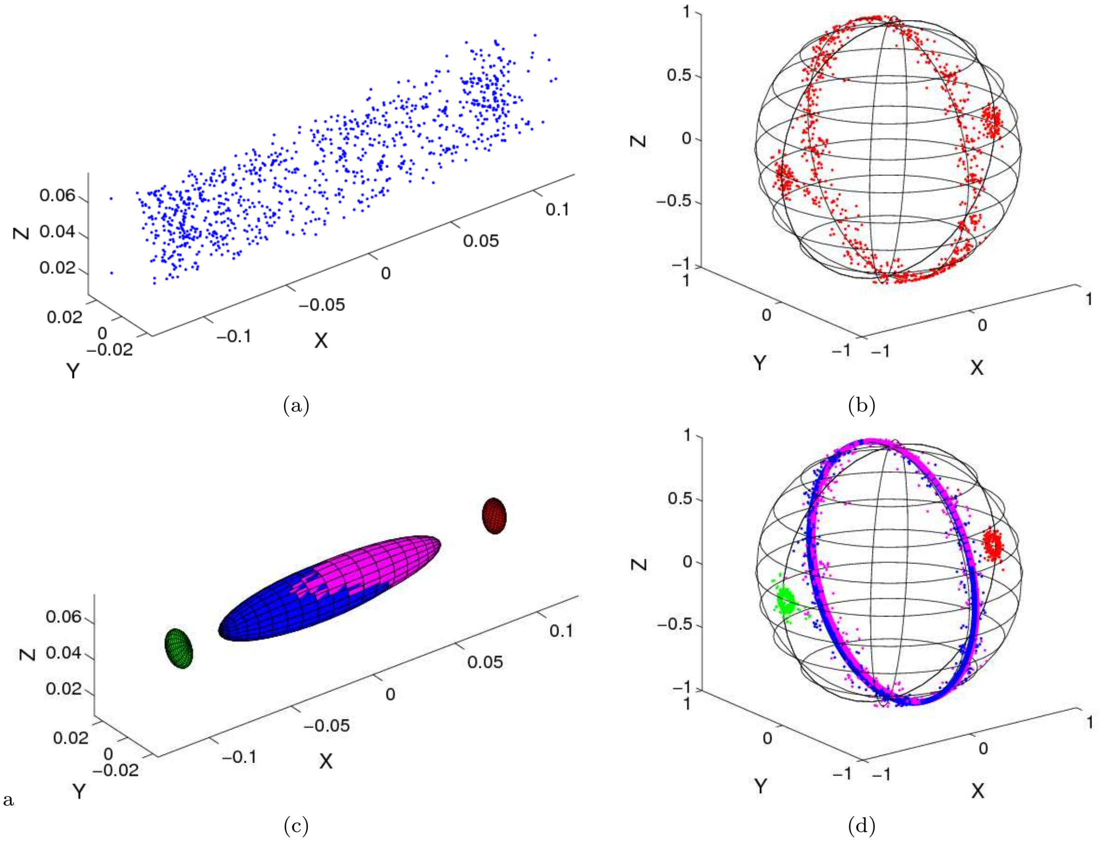
Figures 3(c) and 3(d) show the most common solution discovered by our algorithm for the cylinder. The position and orientation components of the model are shown independently, with similarly colored ellipsoids and circles constituting a single cluster in the pose space of the hand. Each circle represents the great circle on S3 defined by a girdle distribution, while each ellipsoid corresponds to the first standard deviation boundary of the corresponding multivariate Gaussian.

A single cluster was learned for each heuristically-identified grasp type. Due to the significant rotational symmetries present in the object, a girdle distribution was selected as the orientation component for each grasp. The end grasps correspond to the green and red clusters, while grasps along the major axis are represented by the blue and magenta clusters. Because all positions along the major axis of the cylinder are viable for grasping, the corresponding ellipsoids have a larger volume, and are more elongated than those representing the end grasps. Also, notice that the position components of the overhand and underhand grasps overlap significantly. However, the algorithm selects two clusters to represent them because their corresponding orientation components are best described by two different girdle distributions (even though our visual representation of orientation aliases this fact).

Figure 4 shows the mean true positive rate, precision, and false discovery rate for a variety of objects over the course of 10 trials. Focusing our attention on the cylinder, we see that on average the true positive rate and precision are above 0.9. Thus, a majority of the learned models matched our heuristic. However, there is a slight overfitting issue. For example, on one occasion the algorithm identified three clusters for one of the grasps along the lateral surface of the cylinder. Each of these clusters used a Dimroth-Watson component for orientation, instead of a girdle distribution.

4.2 Spray Bottle

The next object we consider is the spray bottle shown in figure 5. Four of the feasible grasps are shown in the figure, with three of them having symmetric grasps achieved by rotating the object 180 degrees about its major axis. There is one grasp that may be made near



(a)

(b)

(c)

(d)

Fig. 3 The training examples and learned affordance model for the cylinder within an object-centered coordinate frame. (a) The position of the hand; (b) The orientation of the hand; (c) The position component of the learned affordance model; (d) The orientation component of the learned affordance model.

the trigger (a), two from the side (b), two from the top (c), and two from the bottom (d). Each of these possibilities was extensively explored by the human teacher, and the collected training examples are shown in figures 7(a) and 7(b).

The affordance model learned by our algorithm for the spray bottle is shown in figures 7(c) and 7(d). Notice that there are a total of seven clusters (one for each demonstrated grasp type), where each of the colored line segments represent the mean rotation vector of a Dimroth-Watson distribution. The red and gray clusters represent the two symmetric grasps near the bottom of the spray bottle shown in figure 5(d). The green and brown clusters describe the set of grasps along the object's major axis as it is approached from the right and left, respectively (b). Notice that hand orientation is approximately orthogonal to the spray bottle's major axis in both cases. Also, the elongated nature of the ellipsoids captures the large variation in hand position along the major axis. Grasp approaches from the top of the spray bottle are represented by the orange and magenta clusters (c). Because the nozzle of the spray bottle is much smaller in comparison to its base, hand position and orientation are more constrained. This can be seen by comparing the relative volumes of the ellipsoids representing grasps near the base with those near the nozzle. Finally, the blue cluster captures the trigger grasp. Notice that in figure 7(a) the set of hand positions used to grasp the trigger seem to be comprised of two distinct sets of points. However, because the orientation of the hand does not vary much for grasps involving trigger, the algorithm allocates only a single cluster.

Over the ten experiments, an average true positive rate of 1.0 was achieved (figure 4). Therefore, every cluster we heuristically identified was in fact learned by the algorithm in each of the experiments. However, notice that the precision is slightly below 1.0. This is a result of overfitting: the algorithm split one cluster into two in a single experiment.

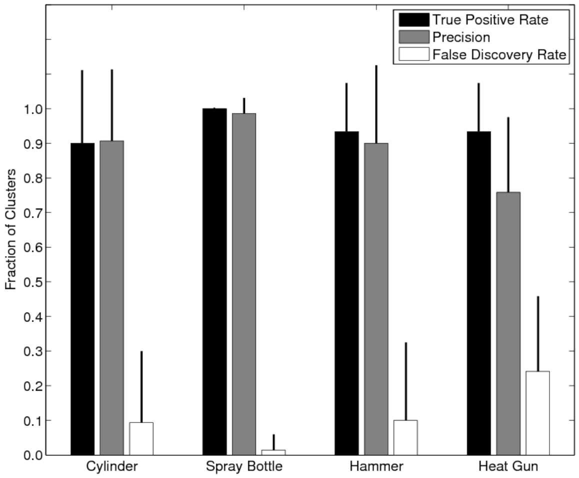


Fig. 4 Contingency table summary for each of the objects used in the clustering experiments.

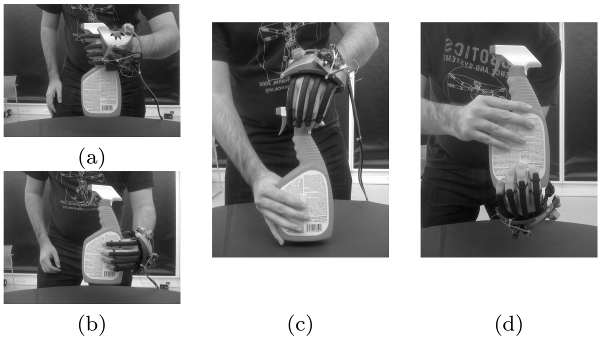


Fig. 5 Four of the feasible grasps demonstrated by the human teacher for the spray bottle. (a) The trigger grasp; (b) The grasp from the right; (c) The top grasp; (d) The bottom grasp. Note that the grasps shown in (b), (c), and (d) have symmetric grasps that are achieved by rotating the spray bottle 180 degrees about its major axis.

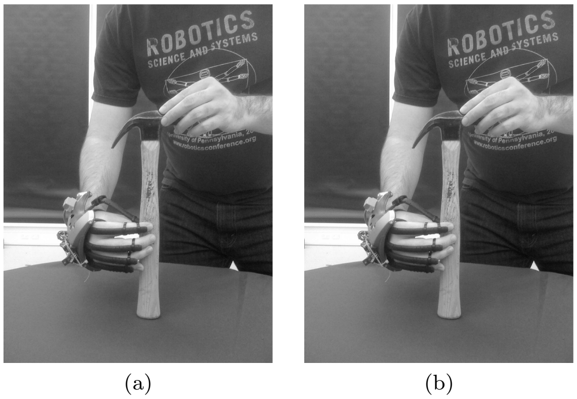


Fig. 6 The feasible grasps demonstrated by the human teacher for the hammer. (a) The handle grasp; (b) The head grasp.

4.3 Hammer

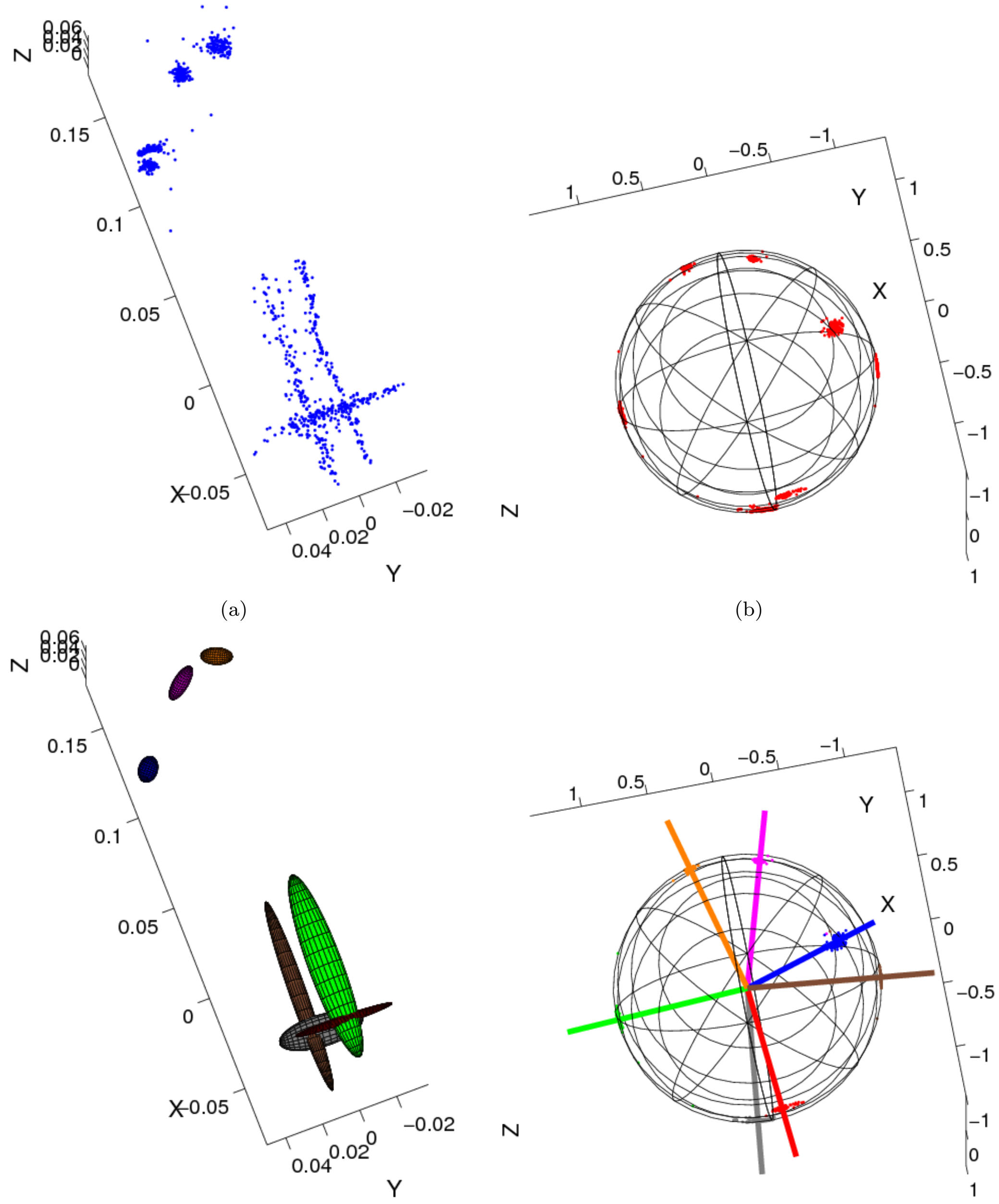
The third object presented to the human teacher was the hammer shown in figure 2(c). This object was selected for a variety of reasons. First, it presented an interesting mixture of orientation constraints. Second, a hammer might be useful to a robot performing a real world task such as building a structure. The possible grasps include those near the handle and the head of the hammer, each of which was demonstrated by the human teacher, and are shown in figures 6(a) and 6(b), respectively. Note that the hand orientations used to grasp the handle always resulted in the thumb being closer to the head of hammer, and that the grasp shown in figure 6(b) has a symmetric grasp that is achieved by rotating the hammer 180 degrees about its major axis. Hence, a total of three grasps are expected by our heuristic.

A typical collection of training examples for these grasps is shown in figures 8(a) and 8(b). The corresponding model learned by our algorithm is shown in figures 8(c) and 8(d). The green cluster represents the set of grasps that may be made with respect to the handle of the hammer. The elongation of the green ellipsoid along the major axis essentially encodes the handle's length. This means that any position on it is viable for grasping. In addition, the hammer may be grasped as long as the orientation of the hand is approximately orthogonal to the handle. Thus, a girdle distribution was learned for this portion of the object. The other two clusters (red and blue) involve grasps near the head of the hammer. These grasps are symmetric in that they are accomplished by grasping the object from the top, rotating the hammer 180 degrees, and then re-grasping. Since the head of the hammer is much smaller in comparison to its handle, the red and blue ellipsoids have less volume than the green ellipsoid. Also, because grasping this portion of the object involves very constrained hand orientations, the algorithm selected Dimroth-Watson distributions to model them.

4.4 Heat Gun

The heat gun shown in figure 2(d) was selected for many of the same reasons as the hammer. It has a variety of orientation constraints, and is similar in shape to other real world objects such as a drill. The feasible grasps for this object are shown in figure 9. In these experiments, the barrel of the heat gun was only grasped below the handle. Furthermore, the hand orientations used to grasp the heat gun's nozzle always resulted in the thumb being closer to the handle (see figure 9(b)).

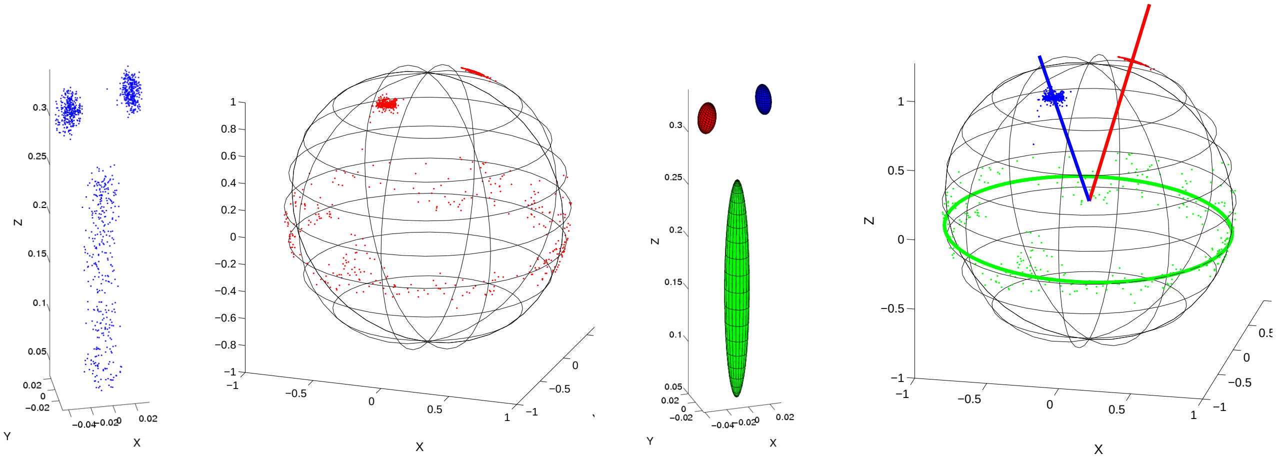
Even though three primary grasps were demonstrated, our algorithm tends to choose solutions with four clus-



(a) (b) (c) (d)

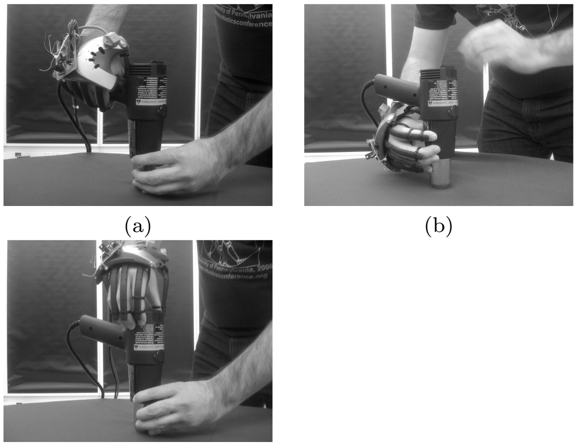
Fig. 7 The training examples and learned affordance model for the spray bottle. (a) The position of the hand; (b) The orientation of the hand; (c) The position component of the learned affordance model; (d) The orientation component of the learned affordance model.

ters (figures 10(c) and 10(d)). This is reflected by the higher false discovery rate shown in figure 4. The red cluster corresponds to grasps along the handle of the heat gun. A Dimroth-Watson distribution is used to model the orientation of the hand because the human teacher only grasped the handle in such a way that affords the use of the trigger. The blue cluster represents approaches from the top where the hand may be arbitrarily rotated about the object's major axis, with position being relatively constrained. The remaining clusters (green and magenta) describe the set of grasps along the lateral surface of the heat gun's nozzle. Ideally this would be a single cluster, but the algorithm preferred to separate it into two. This may be due to the fact that the nozzle widens as one approaches the handle, which means the distribution in position varies more in this region. However, both clusters did capture the rotational symmetry in hand orientation afforded



(a) (b) (c) (d)

Fig. 8 The training examples and learned affordance model for the hammer. (a) The position of the hand; (b) The orientation of the hand; (c) The position component of the learned affordance model; (d) The orientation component of the learned affordance model.



(a) (b) (c)

Fig. 9 The feasible grasps demonstrated by the human teacher for the heat gun. (a) The handle grasp; (b) The nozzle grasp; (c) The top grasp.

by the nozzle. Thus, while not the expected solution, the algorithm learned a reasonable solution a majority of the time. This is supported by the high true positive rate exhibited by the heat gun.

This result for the nozzle may be due to several different factors. First, Gaussian distributions may not be suitable for modeling the positions of the hand when grasping objects exhibiting a conical geometry. Second, the algorithm may also be overfitting. Finally, it is possible that the heuristic is incorrect, and the algorithm is actually finding an underlying structure in the pose space of the hand.

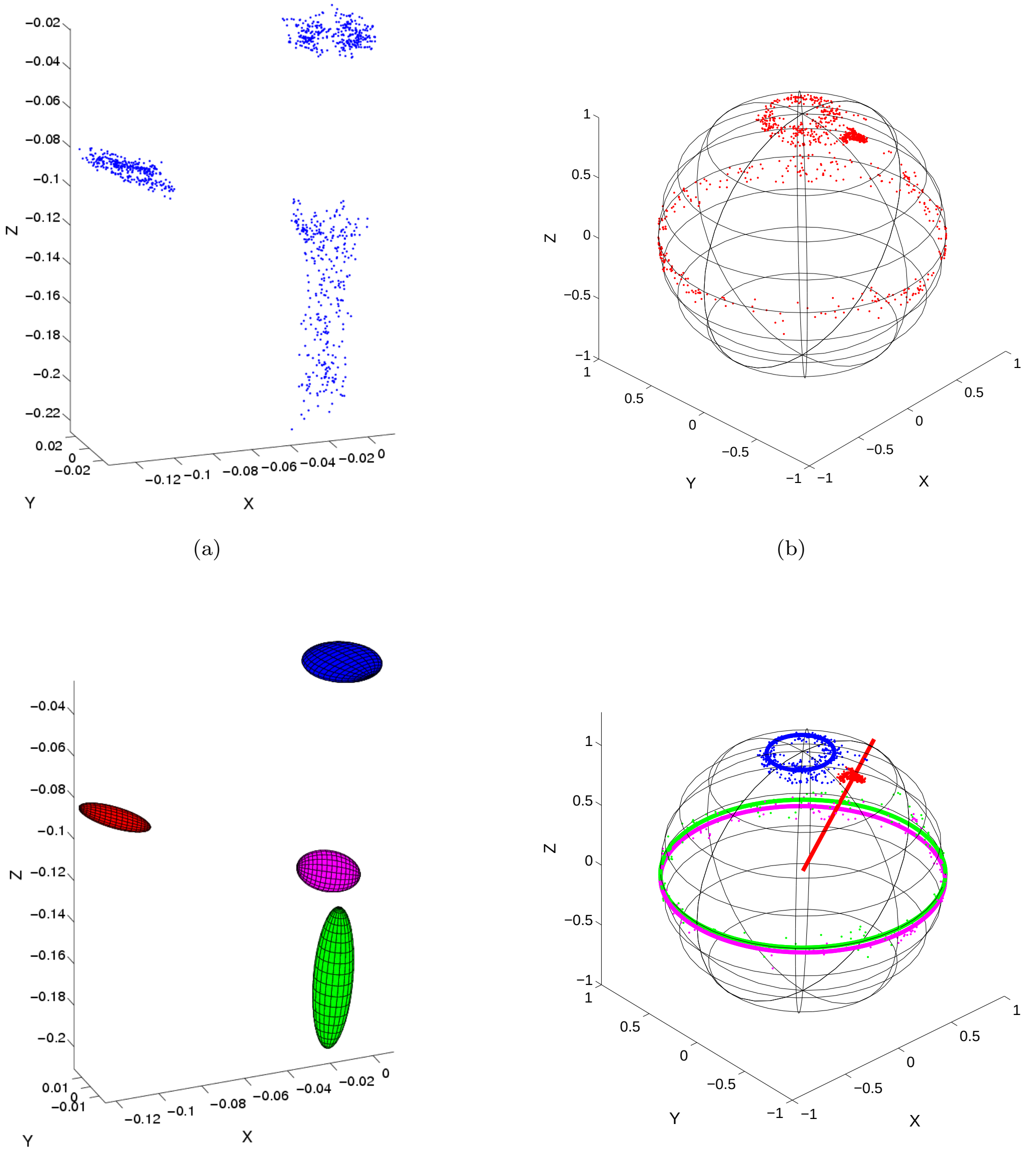
5 Sensitivity Analysis

In order to gain a better understanding of the grasp affordance learning algorithm, its performance is compared to that of a baseline algorithm which removes the filtering step described in section 3.2. Figure 11 shows the mean true positive rate and mean precision for the filtered and unfiltered versions of our algorithm. Notice that the mean true positive rate drops slightly for most of the objects when filtering is introduced. This is typically due to the elimination of solutions where a learned cluster matches one of the heuristically defined clusters, but it explains a small number of training examples.

However, for the spray bottle, the absence of filtering causes more false negatives to be scored because girdle distributions are learned as the orientation component of a cluster, even though Dimroth-Watson distributions were specified by the heuristic. When filtering is introduced, the algorithm tends to throw out these solutions, and as a result, the true positive rate increases.

Notice that for each of the objects the mean precision is lower when no filtering is performed. This occurs because the unfiltered version of the algorithm tends to allocate more clusters than the filtered version of the algorithm. While many of these clusters may match the heuristic (explaining the high mean true positive rates), some of them are unnecessary. When filtering is performed, many of the solutions with extra clusters are eliminated, which results in a higher mean precision for each of the objects.

When all of the objects are considered together, the mean decrease of the true positive rate from the un-



(a)

(b)

(c)

(d)

Fig. 10 The training examples and learned affordance model for the heat gun. (a) The position of the hand; (b) The orientation of the hand; (c) The position component of the learned affordance model; (d) The orientation component of the learned affordance model.

filtered to the filtered version of the algorithm is 0.02 ($p < 0.36$ according to a paired bootstrap test). However, the overall effect of introducing filtering produces a mean increase in precision of 0.15 ($p < 10^{-4}$). This suggests that filtering facilitates the selection of models that better match our heuristic.

6 Discussion

In this paper, we have presented a technique for learning canonical hand positions and orientations for reach-to-grasp actions. Compact representations are constructed from many example grasps made by clustering the pose of the hand. For a given object, we want the set of affordances to be small. This property enables the use of affordances as a way to access "primitives" in higher-level activities, including planning, learning, and the recog-

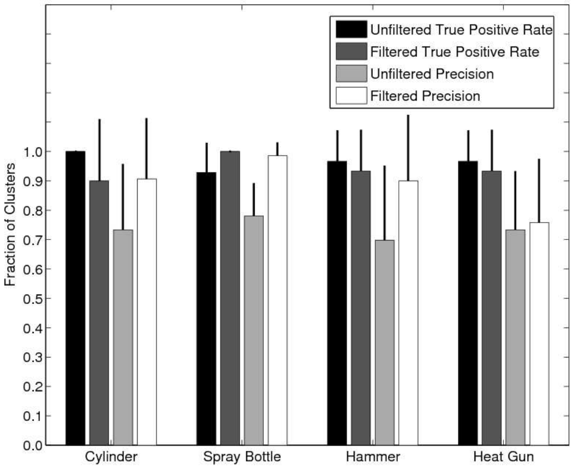


Fig. 11 Contingency table comparison of the unfiltered and filtered versions of the grasp affordance learning algorithm for each of the objects used in the clustering experiments.

nition of motor actions by other agents (Brock et al 2005; Fagg et al 2004).

In particular, the clusters that have been learned map directly onto resolved-rate controllers that can bring a robot hand to a specific position and orientation relative to the object. Note that this control step makes two assumptions: first, that the robot has a similar hand morphology to the human demonstrator; second, that haptic exploration methods are available to refine the grasps once the hand is approximately in the right configuration (Coelho and Grupen 1997; Platt et al. 2002).

Grasping experience for training the models does not have to be derived from observation of human behavior, but can come from a robot performing the grasping task. Experience may be generated through either automatic control or guidance from a human teleoperator. Preliminary results using NASA's humanoid robot *Robonaut* have demonstrated the viability of extending our approach to actual robot systems. In future experiments, we plan to track the pose of objects visually, and learn the corresponding affordances in an object centered coordinate frame. We are also currently designing experiments that focus on using the learned grasp affordance representation as a means of controlling and planning robot grasping actions.

Our analysis presented here has focused entirely on the pose of the hand. In current work, we are taking steps to include the configuration of the fingers into the model. This is desirable in that it leads to a more complete representation of grasp affordances. Instead of using joint distributions defined only over the position and orientation of the hand, we create a joint distribution over hand pose and finger configuration. In order to address the high dimensionality of the finger configuration space, we describe clusters in an *eigengrasp* space (Santello et al 1998; Ciocarlie et al 2007). The learned clusters may then be used in conjunction with other methods of low level control to successfully grasp an object.

We are also interested in bridging the gap between vision and grasping. Wang (2007) has recently presented an approach that recognizes the identity and pose of objects based on visual features. By using this intermediate object representation one can establish an indirect connection between visual features and the hand pose clusters learned by our algorithm. This approach could also allow one to learn representations that are not specific to any particular object, but to components of objects. Thus, if a novel object is composed of parts similar to those in previous experience, the robot should still be able to grasp the object.

Note that this affordance representation captures the syntax of grasping (i.e., what grasps are possible for a given object), and does not take into account the semantics of grasping (how an object is to be used in the larger context of a task). This distinction, which is drawn by Gibson, is a critical one for a learning agent. When a new task is presented, the syntax of interacting with a specific object can be readily accessed and used. The learning agent is then left with the problem of selecting from a small menu of possible grasping actions to solve the new task. This abstraction can have important implications for the agent quickly learning to perform in these novel situations.

Acknowledgements This research was supported by NSF/CISE/REU award #0453545 and by NASA EPSCoR grant #7800. We would also like to thank Rob Platt, Mike Goza, Myron A. Diftler, and William Bluethmann at the Dexterous Robotics Laboratory at NASA's Johnson Space Center for their support in providing access to Robonaut.

References

Bekey GA, Liu H, Tomovic R, Karplus WJ (1993) Knowledge-based control of grasping in robot hands using heuristics from human motor skills. IEEE Transactions on Robotics and Automation 9(6):709–722

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7):719–725

Brock O, Fagg AH, Grupen RA, Karuppiah D, Platt R, Rosenstein M (2005) A framework for humanoid control and intelligence. International Journal of Humanoid Robotics 2(3):301–336

Ciocarlie M, Goldfeder C, Allen P (2007) Dexterous grasping via eigengrasps: A low-dimensional approach to a high-complexity problem. In: Proceedings of the Robotics: Science & Systems 2007 Workshop - Sensing and Adapting to the Real World, Electronically published

Coelho, Jr JA, Grupen RA (1997) A control basis for learning multifingered grasps. Journal of Robotic Systems 14(7):545–557

Coelho, Jr JA, Piater J, Grupen RA (2000) Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot. Robotics and Autonomous Systems Journal, special issue on Humanoid Robots 37(2–3):195–219

de Granville C, Southerland J, Fagg AH (2006) Learning grasp affordances through human demonstration. In: Proceedings of the International Conference on Development and Learning, electronically published

Dempster A, Laird N, Rubin D (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1):1–38

Fagg AH, Rosenstein MT, Platt, Jr R, Grupen RA (2004) Extracting user intent in mixed initiative teleoperator control. In: Proceedings of the American Institute of Aeronautics and Astronautics Intelligent Systems Technical Conference, Electronically published

Gibson JJ (1966) The Senses Considered as Perceptual Systems. Allen and Unwin

Gibson JJ (1977) The theory of affordances. In: Shaw RE, Bransford J (eds) Perceiving, Acting, and Knowing, Lawrence Erlbaum, Hillsdale

de Granville C (2008) Learning grasp affordances. Master's thesis, School of Computer Science, University of Oklahoma, Norman, OK

Mardia KV, Jupp PE (1999) Directional Statistics. Wiley Series in Probability and Statistics, Wiley, Chichester, West Sussex, England

Miller A, Knoop S, Christensen H, Allen P (2003) Automatic grasp planning using shape primitives. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp 1824–2829

Piater J, Grupen R (2002) Learning appearance features to support robotic manipulation. In: Proceedings of the Cognitive Vision Workshop, electronically published

Platt, Jr R, Fagg AH, Grupen RA (2002) Nullspace composition of control laws for grasping. In: Proceedings of the International Conference on Intelligent Robots and Systems, pp 1717–1723

Rancourt D, Rivest LP, Asselin J (2000) Using orientation statistics to investigate variations in human kinematics. Applied Statistics 49(1):81–94

Rivest LP (2001) A directional model fo the statistical analysis of movement in three dimensions. Biometrika 88(3):779–791

Santello M, Flanders M, Soechting JF (1998) Postural hand synergies for tool use. Journal of Neuroscience 18(23):10,105–10,115

Stoytchev A (2005) Toward learning the binding affordances of objects: A behavior-grounded approach. In: Proceedings of the AAAI Spring Symposium on Developmental Robotics, electronically published

Sweeney JD, Grupen R (2007) A model of shared grasp affordances from demonstration. In: IEEE-RAS International Conference on Humanoid Robots, session FRa2, electronically published

Wang D (2007) A 3D feature-based object recognition system for grasping. Master's thesis, School of Computer Science, University of Oklahoma, Norman, OK

A Normalization Terms

The normalization terms of the Dimroth-Watson and girdle distributions are as follows:

$$F\left(k\right) = \frac{2}{\pi \int_0^{\pi} \int_0^{2\pi} \sin\theta e^{k \cos^2 \xi_1 \sin^2 \frac{\theta}{2}} d\xi_1 \, d\theta},$$

and

$$\bar{F}\left(k\right) = \frac{k}{2\pi^2 \left(e^k - 1\right)}.$$

See de Granville (2008) for the derivations of both normalization terms.

B Maximum Likelihood Estimates of the Concentration Parameters

For computational efficiency, the maximum likelihood estimates of k for the Dimroth-Watson and girdle distributions are approximated by:

$$k = G^{-1}(z) \approx 1.9090 + 3.4599 \log\left(z\right) \log\left(1.6376\, z\right),$$

and

$$k = \bar{G}^{-1}(\bar{z}) \approx 4.2648 + 4.0254 \log\left(\bar{z}\right) \log\left(5.2532\, \bar{z}\right),$$

where

$$z = -\frac{\sum_{i=1}^{N} \left(\mathbf{q}_i^T \mathbf{u}\right)^2}{N},$$

and

$$\bar{z} = -\frac{\sum_{i=1}^{N} \left(\mathbf{q}_i^T \mathbf{u}_1\right)^2 + \left(\mathbf{q}_i^T \mathbf{u}_2\right)^2}{N}.$$