

Grasping Affordances: Learning to Connect Vision to Hand Action

Charles de Granville, Di Wang, Joshua Southerland, Robert Platt, Jr.
and Andrew H. Fagg

Abstract When presented with an object to be manipulated, a robot must identify the available forms of interaction with the object. How might an agent automatically acquire this mapping from visual description of the object to manipulation action? In this chapter, we describe two components of an algorithm that enable an agent to learn a grasping-oriented representation by observing an object being manipulated by a human teacher. The first component uses the sequence of image/object pose tuples to acquire a model of the object's appearance as a function of the viewing angle. We identify visual features that are robustly observable over a range of similar viewing angles, but that are also discriminative of the set of viewing angles. Given a novel image, the algorithm can then estimate the angle from which the object is being viewed. The second component of the algorithm clusters the sequence of observed hand postures into the functionally distinct ways that the object may be grasped. Experimental results demonstrate the feasibility of extracting a compact set of canonical grasps from this experience. 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.

Charles de Granville, e-mail: chazz184@gmail.com
Di Wang, e-mail: di@cs.ou.edu
Joshua Southerland, e-mail: Joshua.B.Southerland-1@ou.edu
Andrew H. Fagg, e-mail: fagg@cs.ou.edu
Symbiotic Computing Laboratory
School of Computer Science
University of Oklahoma
Norman, OK

Robert Platt, Jr., e-mail: robert.platt-1@nasa.gov
Dexterous Robotics Laboratory
Johnson Space Center, NASA
Houston, TX

1 Introduction

Manipulating one's world in very flexible ways is a skill that is shared only by a small number of species. Humans are particularly skilled at applying their manipulation abilities in novel situations using a range of effectors, from hands and other parts of the body, to tools. How can robots come to organize and learn knowledge representations for solving grasping and manipulation problems in unstructured environments? J. J. Gibson (1966, 1977) suggests that these representations should be partitioned into *what* can be done with particular objects and *why* an object should be manipulated in a certain way. The first of these, which Gibson terms *object affordances*, captures the details of what can be done with the object by the agent. The latter captures information about how individual manipulation skills are to be put together in order to solve a specific task. The task-neutral affordance representation is important in that it can provide an agent with a menu of actions/activities that are possible with a given object – whether the current task is well known or is a new one. Hence, the affordance representation enables the agent to potentially bring a substantial amount of knowledge to new tasks that are to be solved.

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 visual and haptic 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 these representations to be rooted in an agent's own experiences – either through direct interaction with objects or through observation of other agents' interactions.

In this chapter, we describe two efforts toward addressing this challenge. First, we describe an approach for visually recognizing the 3D orientation of an object. The models of object appearance are based entirely on sequences of image/object pose pairs as the object is being manipulated. The learning algorithm identifies robust descriptions of object appearance from different viewing angles. Second, we introduce a method of identifying descriptions of canonical grasps (that include hand pose and finger configuration) based on observation of a large number of example grasps made by a human teacher. We employ a clustering method in this *hand posture space* that identifies a small number of these canonical grasps. The resulting grasp descriptions can then be used by the agent for planning and execution of grasping actions and for interpreting the grasping actions of other agents.

2 Learning Models of 3D Object Appearance

One of our ultimate goals is for a robotic agent to learn affordance representations based on experience gathered by looking at an object as the agent manipulates it. In particular, we would like to construct visual models that enable the agent to recognize the object and the angle from which it is being viewed. This interactive ap-

proach means that although the agent is able to control many of the conditions in which this experience is gathered, the learning approach must be robust to spurious features in the visual stream, including occlusions by the robot itself and lighting effects such as shadows and specular reflections. The challenge is to discover visual operators that are sensitive to the appearance of the object at some subset of viewing angles (or *aspects*), but that are not “distracted” by these spurious effects. In our approach, individual visual operators recognize the appearance of an object for a subset of viewing aspects. A complete 3D appearance model of an object is captured by identifying a collection of visual operators that cover all possible viewing aspects.

2.1 Edgel Constellations for Describing 2D Object Appearance

A visual operator in our case recognizes a specific constellation of oriented *edgels* (Piater and Grupen, 2000; Coelho, Jr. et al, 2000; Wang, 2007). Edgels are edge image features defined at each pixel, and are described by their orientation in the image and their magnitude. Piater and Grupen (2002) define a *constellation* as a set of edgels that are arranged in some geometric configuration in the 2D image space. This geometric configuration is represented by the relative position and orientation between edgels in a constellation. By construction, a constellation is rotation-invariant in the image plane.

Fig. 1 illustrates two constellations that have been identified for two distinct viewing angles of a cup. The constellation that matches the side view (a) captures the top rim of the cup. The constellation that matches the bottom view (b) captures edgels on both the top rim and the bottom of the cup. Within a novel image, a constellation is considered to match if all of the constellation edgels can be found at the correct relative position and orientation. The highest degree of match occurs when the set of edgels in the constellation align perfectly with high-magnitude edges in the query image.

2.2 Capturing Object Appearance in 3D

Although the 2D features are invariant under rotation within the image plane, it is clear from Fig. 1 that the rotations out of this plane can dramatically alter the appearance of an object. We can represent all possible viewing aspects as the set of points on the unit sphere with the observed object as the center (Fig. 2). Imagine a camera located at some point on this *aspect sphere*, oriented toward the sphere’s origin. This point therefore constrains two orientation DOFs, leaving free the rotation about the camera’s axis. The object appearance at a single point can be described by one (or a small number of) edgel constellations. For the case of the cup, one can imagine a unique constellation that only matches a set of views surrounding the top

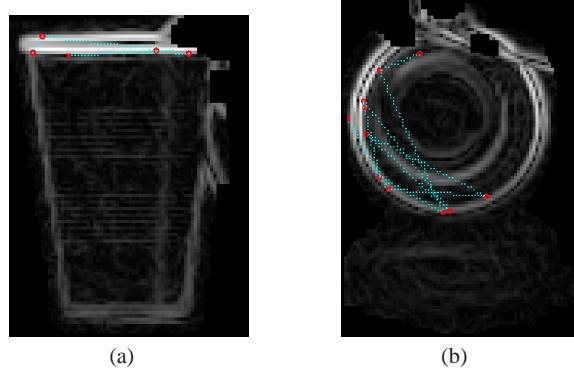


Fig. 1 Constellations matching a side view (a) and a bottom view (b) of a cup. The constellations have been “painted” on top of the edge magnitude image for each viewing direction. Individual edgels are shown using small circles; the geometric constraints between edgels are shown as dotted lines.

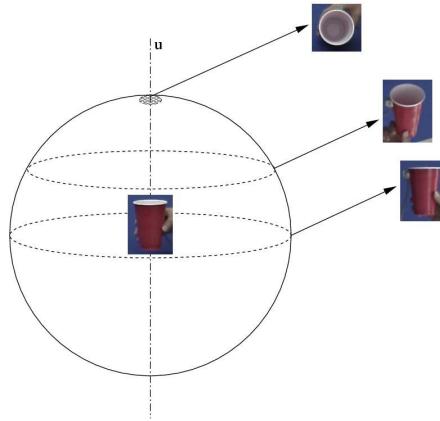


Fig. 2 The aspect sphere of a cup with rotation symmetry about \mathbf{u} .

pole of the sphere. As the viewing angle deviates from \mathbf{u} , the likelihood of observing the constellation can drop quickly. For a constellation that recognizes a “non-polar” aspect, the set of recognized aspects will fall along a circular band on the sphere. This is because the cup’s appearance does not change with rotations about the vertical axis. As the viewing angle deviates from the center of the band, it becomes less likely that the constellation will be observed.

Fig. 3 illustrates the aspects for which the constellations of Fig. 1 are found. The major axis of the cup in these figures falls along the X axis, with the top of the cup at $X = -1$. The constellation that recognizes the side of the cup is found most often along the circle for which $X = 0$ (a). The constellation that recognizes the bottom of the cup is found most often around $X = 1$, but is occasionally found around $X = -1$ (b). This is the case because this particular constellation recognizes pieces of two

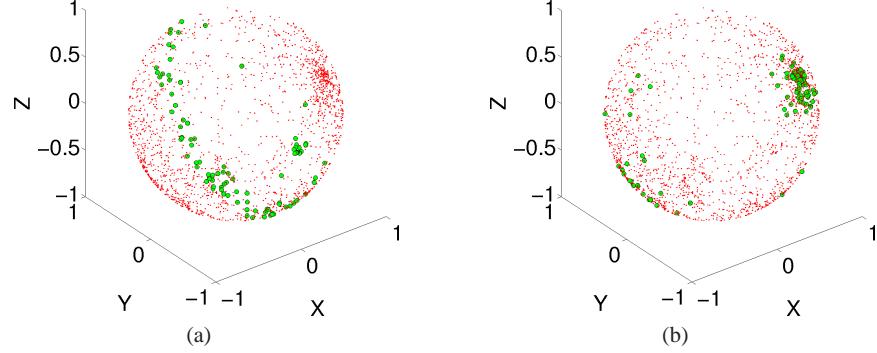


Fig. 3 Constellation matches for (a) the side constellation (of Fig. 1(a)) and (b) the bottom constellation (Fig. 1(b)). Small dots show aspects that correspond to image samples; large dots indicate where the constellation matches. In this case, the major axis of the cup is aligned with the X axis.

concentric circles of particular radii, a feature that is sometimes visible from the top of the cup.

How can we compactly represent the set of viewing angles for which a constellation is viewable? Specifically, we would like to capture the likelihood of the aspect given that a particular constellation C_i has been observed: $p(\mathbf{a}|Obj, C_i)$. Bingham and Mardia (1978) proposed the *small circle distribution*, which allows us to describe Gaussian-like distributions on the unit sphere.

$$b(\mathbf{a}|\tau, v, \mu) = \frac{1}{F(\tau, v)} e^{-\tau(\mu^T \mathbf{a} - v)^2}, \quad (1)$$

where \mathbf{a} , μ are unit vectors and μ denotes the mean direction; τ is a scalar that gives the concentration of the distribution (the higher the τ , the more concentrated the distribution); v is a scalar that determines the shape of the distribution; and $F(\tau, v)$ is a normalizing term. Note that equation 1 obtains a maximum value when $\mu^T \mathbf{a} = v$. This set of \mathbf{a} 's fall at a fixed angle about μ . By adjusting the parameters of this distribution, we can describe different shapes of clusters on our aspect sphere (Fig. 4).

2.3 Learning Complete 3D Appearance Models

Given a set of image/aspect tuples, the challenge is to discover a set of edgel constellations that cover the entire aspect sphere. Our algorithm is outlined in Fig. 5. During the training process, the algorithm samples a single constellation at a time from a specific training image (call this image P). A two-step filtering process is used to determine whether the constellation is *locally robust* and *globally discriminative*. First, the set of images from aspects surrounding P are searched for the

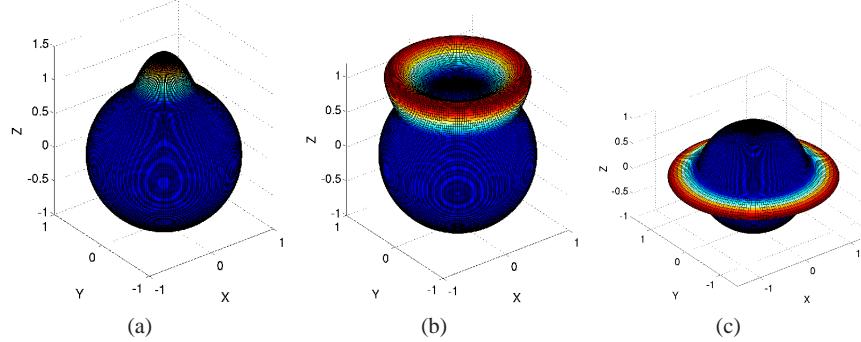


Fig. 4 Gaussian-like distributions on the unit sphere, where $\mu = [0, 0, 1]$. In all cases, the surface radius is $1 + p/(2 \times \max(p))$, where p is the likelihood at the corresponding aspect. (a) Uniform Gaussian: $\tau = 50$, $\nu = 1.2$; (b) small circle: $\tau = 100$, $\nu = 0.8$; and (c) great circle: $\tau = 100$, $\nu = 0$.

constellation. If the constellation describes transient features such as shadows, then it is unlikely to match neighboring images. If this is the case, the constellation is discarded and a new one is sampled. The second filter examines all images in the training set for the constellation. If the degree of match of the constellation distinguishes the neighboring images from most of the remaining training set, then the constellation is considered to be discriminative. Formally, the discriminative power of the constellation is measured using the Kolmogorov-Smirnov distance (KSD) between the neighboring and complete population of images (Piater and Grupen, 2002). Should the constellation satisfies both filters, the algorithm then finds the parameters of a probability density function that describes the set of aspects in which the constellation is observed. This training process continues iteratively until the entire set of generated constellations cover most of the training images.

Given a novel image, we would like to accurately estimate the aspect from which it is being viewed. More specifically, assuming that a set of constellations C_1, C_2, \dots, C_N are either observed or not in an image, we would like to find the aspect, a that maximizes $p(\mathbf{a}|Obj, C_1, \dots, C_N)$. Making the naïve Bayes assumption, we can estimate this likelihood accordingly:

$$p(\mathbf{a}|Obj, C_1, \dots, C_N) = \prod_{i=1}^N p(\mathbf{a}|Obj, C_i). \quad (2)$$

In practice, we make use of a local gradient ascent search with multiple starting locations to identify the maximum likelihood \mathbf{a} .

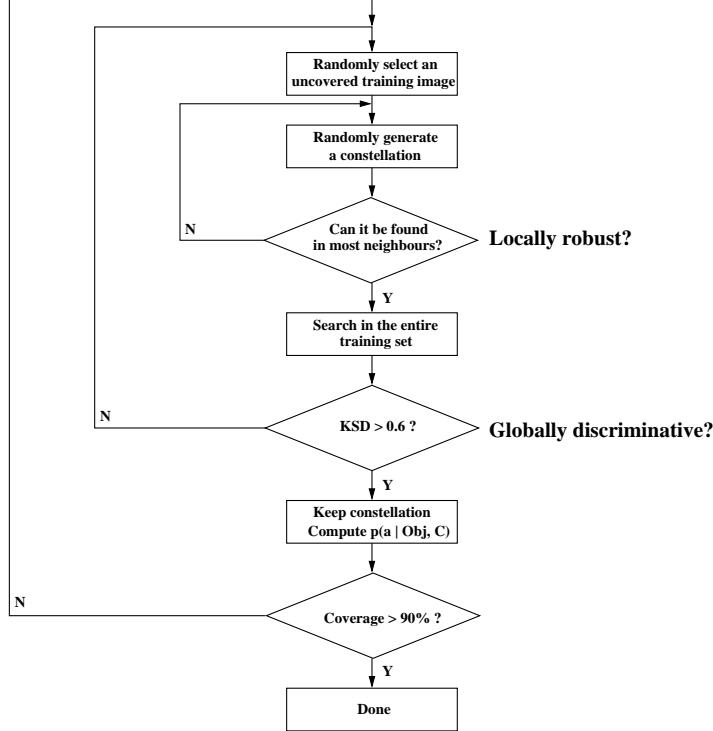


Fig. 5 Overall structure of the aspect recognition algorithm.

2.4 Data Collection and Preprocessing

In our experiments, each element in the data set is a *tuple* of image and object pose. A Polhemus Patriot (Colchester, VT) is attached to the object so that the 3D position and orientation of the object can be paired with each image. Tuples are gathered continuously as the object is rotated in front of the camera. In all, a data set will contain about 2000 such tuples. We employ an image preprocessing step that identifies a region of interest (ROI) that ideally contains only the object. The stationary background and skin-colored objects are first subtracted from the image. The ROI is then selected to encompass a large, contiguous set of the remaining pixels within the middle of the image frame. In practice, the ROI contains the object in excess of 99% of the images.

Fig. 6 illustrates the recognition process. An independent testing image of the cup is shown in Fig. 6(a) and the corresponding (true) aspect is shown on the aspect sphere in Fig. 6(b). For this particular image, two constellations are observed (Fig. 6(c) and Fig. 6(e)). The density functions corresponding to these two constellations are shown in Fig. 6(d) and Fig. 6(f). The combined density function (Eq. 2)

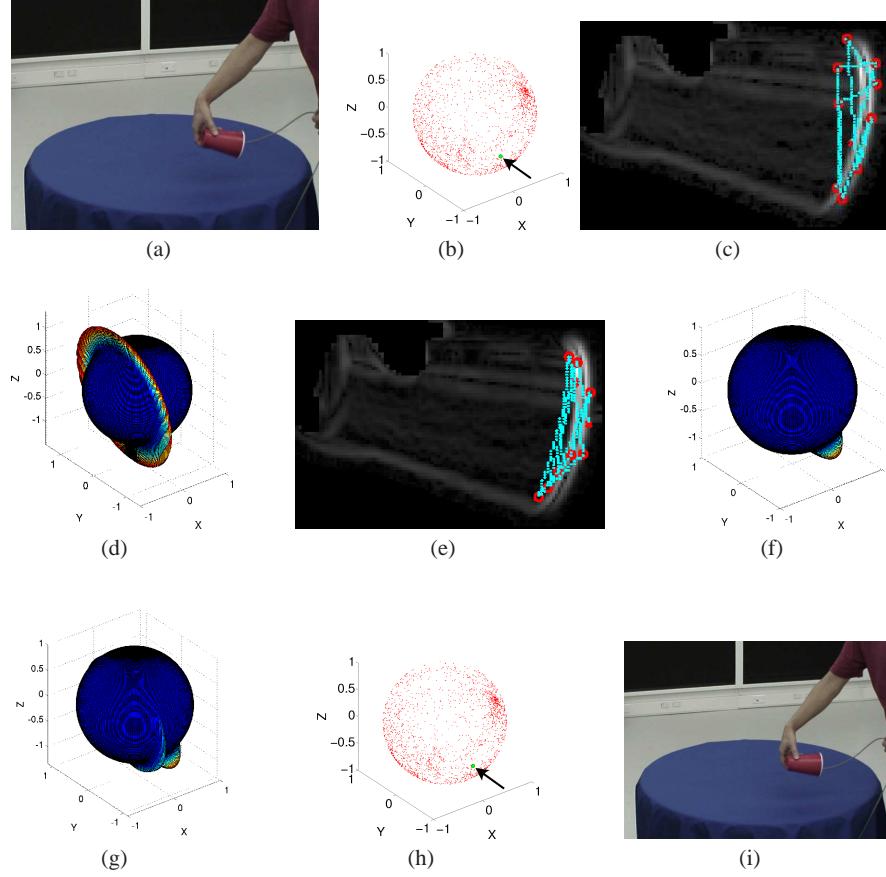


Fig. 6 Example aspect recognition. (a) a testing image of the cup, (b) the true aspect from which the object is observed, (c) match of constellation 1 to the test image, and (d) $p(\mathbf{a}|cup, C_1)$. (e) match of constellation 2 to the test image, (f) $p(\mathbf{a}|cup, C_2)$, (g) $p(\mathbf{a}|cup, C_1, C_2)$, (h) the maximum likelihood aspect, and (i) the nearest training image that corresponds to this aspect.

is shown in Fig. 6(g). The maximum likelihood aspect is 0.31° from the true aspect (Fig. 6h).

2.5 Experimental Results

Both symmetric (a cup and a block) and asymmetric (a mug and a spray bottle) objects are used in the experiment. For each object, about 2000 sample image/aspect tuples are taken uniformly in order to cover the aspect sphere as well as possible. For each object, we performed 10 independent experiments. For each experiment, a dif-

ferent set of 100 samples are randomly selected and reserved as the test data set; the remaining samples are used as training data. Error is measured for each test image as the angle between the estimated and true aspects, down to the symmetry of the object. When there are multiple estimated aspects, the mean of error is calculated for a single test image. We report the mean error over 100 images and 10 experiments. We compare the proposed approach with one in which no filtering is performed (the “unfiltered method”) and with a method that guesses aspects randomly.

Both the filtered and unfiltered methods cover 3556 out of 4000 testing images (of 4 objects, 100 test images and 10 experiments). No constellations are found in the remaining test images. The aspect estimate error histogram for the three methods is shown in Figure 7a. These histograms include errors from all ten experiments and four objects. The filtered and unfiltered methods decrease exponentially with increasing error. However, the filtered method is biased more toward lower errors. The mean error for the random method is substantially higher than either of the other two methods.

The mean errors and standard deviations for each object are shown in Fig. 7b. For both methods, we can see that the errors for the spray bottle are relatively large compared to that for the other objects. The reason is that the shape and texture of the spray bottle are more complex than the other objects. As a result, many constellations often match to a high degree with the texture of the labels, even though they are not originally generated from those regions.

We can also see that filtering is a benefit, especially for the more complicated objects. For the simple objects, sampled constellations for a particular aspect are often very similar to each other. Hence, the filtering step does not make any practical distinctions between different constellations. As the objects become more complicated, such as with the spray bottle, a particular aspect will give rise to a set of rather different constellations. Hence, the filtering step is able to make meaningful distinctions between these constellations.

The performance difference between the filtered and unfiltered methods is significant for all four objects by a two-tail, paired t-test (block: $p < 0.05$; cup: $p < 10^{-3}$; mug: $p < 10^{-4}$; spray bottle: $p < 10^{-4}$). We should also note that the random guess method does not perform as poorly as one might expect. This is because these errors have also been adjusted according to the symmetric properties of the objects.

3 Learning Canonical Grasps for Objects

Once an object has been visually identified and localized in space, how can an agent describe the set of possible grasping actions that may be taken? Because the set of possible actions will ultimately be used for planning and for exploratory learning, we are motivated to make this set as small as possible so as to reduce the complexity of either search process. One approach to constructing this representation is to begin with a large set of successful example grasps and then to cluster them into a small set of canonical grasps. This set of examples could be derived from manipulation

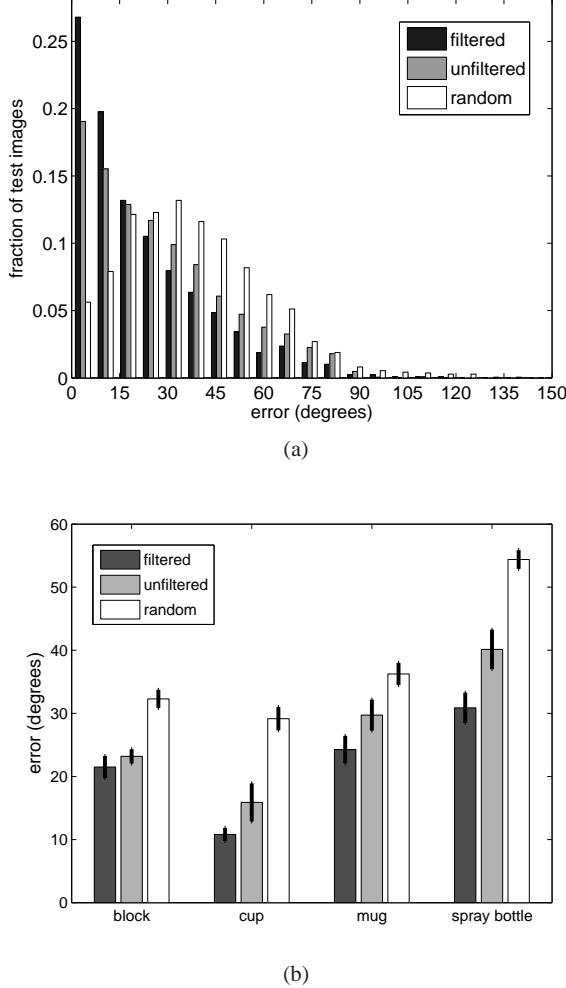


Fig. 7 (a) Histogram of aspect estimate errors made by the three methods (all objects), and (b) aspect estimation errors for each method and object.

sequences produced by the agent itself, or by a human acting directly on the manipulated object or acting through the agent via teleoperation. Our focus to date has been on these two human-driven methods.

We describe each example grasp with the following: 1) position of the hand in an object-centered coordinate frame, 2) orientation of the hand, and (in some cases) 3) the joint angles of the fingers of the hand. Clustering is performed using a *mixture of probability density functions* approach, in which each cluster corresponds to a canonical hand configuration that describes all three of these components (de

Granville et al., 2006, 2009, and submitted; de Granville, 2008). Below, we detail each of these steps and then show that this method can successfully identify meaningful clusters from teleoperation experiments performed using NASA’s humanoid robot *Robonaut*.

3.1 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:

$$f(\mathbf{q}|\mathbf{u}, k) = F(k) e^{k(\mathbf{q}^T \mathbf{u})^2}, \quad (3)$$

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. 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} . Fig. 8(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:

$$\bar{f}(\mathbf{q}|\mathbf{u}_1, \mathbf{u}_2, k) = \bar{F}(k) e^{k[(\mathbf{q}^T \mathbf{u}_1)^2 + (\mathbf{q}^T \mathbf{u}_2)^2]}, \quad (4)$$

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. Fig. 8(b) illustrates the girdle distribution on 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 $(\mathbf{q}^T \mathbf{u}_1)^2 + (\mathbf{q}^T \mathbf{u}_2)^2 = 1$. However, as the angle between \mathbf{q} and the closest point on the circle increases, the density decreases exponentially.

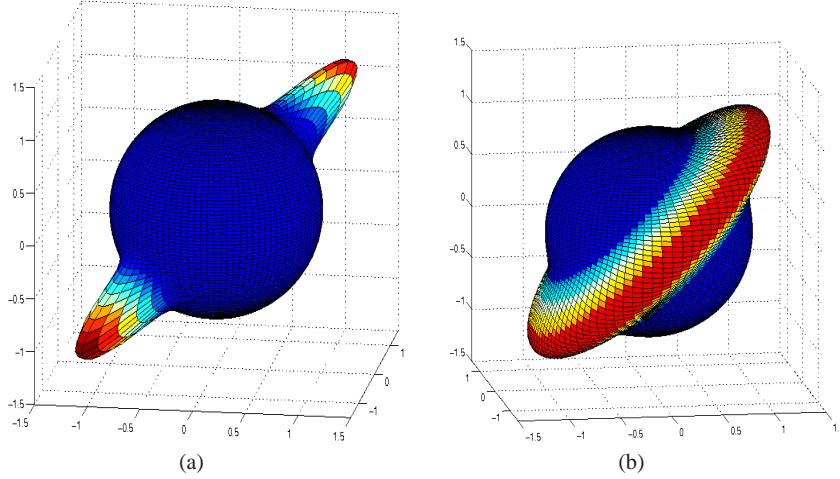


Fig. 8 Three dimensional representations of the Dimroth-Watson (a) and girdle (b) distributions on S^2 . In both cases, the surface radius is $1 + p$, where p is the probability density at the corresponding orientation

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, $\Lambda \in \mathbb{R}^{4 \times 4}$:

$$\Lambda = \frac{\sum_{i=1}^N \mathbf{q}_i \mathbf{q}_i^T}{N}, \quad (5)$$

where \mathbf{q}_i is the orientation of the i th sample, and N is the total number of samples. The MLE of \mathbf{u} is parallel to the first eigenvector of Λ (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 Λ (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):

$$G(k) \equiv \frac{F'(k)}{F(k)} = -\frac{\sum_{i=1}^N (\mathbf{q}_i^T \mathbf{u})^2}{N}. \quad (6)$$

In the case of the girdle distribution, the MLE of k uniquely satisfies:

$$\bar{G}(k) \equiv \frac{\bar{F}'(k)}{\bar{F}(k)} = -\frac{\sum_{i=1}^N [(\mathbf{q}_i^T \mathbf{u}_1)^2 + (\mathbf{q}_i^T \mathbf{u}_2)^2]}{N}. \quad (7)$$

For computational efficiency, we approximate $G^{-1}()$ and $\bar{G}^{-1}()$ when solving for k . This approximation is discussed in detail by de Granville (2008).

3.2 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:

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}. \quad (8)$$

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

3.3 Modeling Finger Posture

Humanoid robots such as Robonaut typically have many degrees of freedom (DOF) available to perform manipulation tasks. For example, each of Robonaut's hands has 12 DOF: three for the thumb, index, and middle fingers; one for the ring and pinkie fingers; and one for the palm (Ambrose et al, 2000). Incorporating finger configurations into our clustering algorithm is a key step to constructing more complete grasp affordance representations. One possible approach to this problem is to learn clusters using the full dimensionality of the robot's end-effector. However, hands with a large number of joints can be difficult to model because an increasingly large number of training examples is needed to adequately sample a space as more and more dimensions are added to it.

One question is whether or not all of the available DOFs of the hand are even necessary to accurately model the finger configurations used for grasping. For example, when executing a power grasp, the fingers tend to flex in unison. This means that there is a strong correlation between the distal and proximal joints of each finger, as well as a correlation across fingers. Santello et al (1998) and Ciocarlie et al (2007) present an approach that takes advantage of such correlations through the notion of an *eigengrasp*. The eigengrasps of a hand comprise a set of basis vectors in the joint space of the hand. Linear combinations of a small number of these eigengrasps can be used to approximate the finger configurations used when grasping.

More formally, let $\mathbf{p} \in \mathbb{R}^d$ be a column vector of joint angles describing the finger configuration of a robot's end-effector, and $\mathbf{V} \in \mathbb{R}^{d \times d}$ constitute a basis for the vector space of which \mathbf{p} is a member. The columns of \mathbf{V} represent directions in the joint space of the hand (the eigengrasps), and are ordered from those that capture the most

variance in finger configuration to the smallest (i.e., from the largest corresponding eigenvalue to the smallest). Linear combinations of the columns of \mathbf{V} can be used to represent any possible pose for the fingers of the robot's hand:

$$\mathbf{p} = \sum_{i=1}^d a_i \mathbf{v}_i = \mathbf{Va}. \quad (9)$$

Here, $\mathbf{v}_i \in \mathbb{R}^d$ is the i 'th column of \mathbf{V} , $\mathbf{a} \in \mathbb{R}^d$ is a column vector of coefficients, and $a_i \in \mathbb{R}$ is element i of the vector \mathbf{a} .

Because there may be a large number of joints in the robot's hand, the configuration of the fingers may be approximated by using a small number (K) of eigengrasps:

$$\mathbf{p} = \sum_{i=1}^K a_i \mathbf{v}_i = \hat{\mathbf{V}}\hat{\mathbf{a}}, \quad (10)$$

where $\hat{\mathbf{V}} = [\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_K]$, and $\hat{\mathbf{a}} = [a_1 a_2 \dots a_K]^T$. Given a finger configuration \mathbf{p} and a subset of the eigengrasps $\hat{\mathbf{V}}$, a low dimensional representation of \mathbf{p} is obtained by solving the system of linear equations in 10 for $\hat{\mathbf{a}}$.

We compute the set of eigengrasps using samples of the joint angle vector as a teleoperator grasps a set of objects. Let $\mathbf{P} \in \mathbb{R}^{d \times N}$ be the set of finger configurations resulting from the human demonstration, where N denotes the number of examples. The eigengrasps are determined by computing the eigenvectors of \mathbf{P} 's covariance matrix (Hand et al, 2001).

3.4 Modeling Mixtures of Hand Postures

We model a grasp using a joint distribution defined over hand pose and the finger posture. Specifically:

$$g(\mathbf{x}, \mathbf{q}, \hat{\mathbf{a}} | \theta) = p(\mathbf{x} | \theta_p) f(\mathbf{q} | \theta_f) p(\hat{\mathbf{a}} | \theta_h), \quad (11)$$

and

$$\bar{g}(\mathbf{x}, \mathbf{q}, \hat{\mathbf{a}} | \bar{\theta}) = p(\mathbf{x} | \theta_p) \bar{f}(\mathbf{q} | \theta_{\bar{f}}) p(\hat{\mathbf{a}} | \theta_h). \quad (12)$$

Here, $p(\hat{\mathbf{a}} | \theta_h)$ is a multivariate Gaussian distribution over K dimensions. We assume that hand position, hand orientation, and finger configuration are conditionally independent given a cluster.

An individual hand posture 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, $h()$, is defined as:

$$h(\mathbf{x}, \mathbf{q} | \Psi) = \sum_{j=1}^M w_j c_j(\mathbf{x}, \mathbf{q} | \theta_j), \quad (13)$$

$$\Psi = (w_1, \dots, w_M, \theta_1, \dots, \theta_M), \quad (14)$$

and

$$\sum_{j=1}^M w_j = 1, \quad (15)$$

where M denotes the number of component densities, and c_j is one of the two density functions describing hand pose ($g()$ or $\bar{g}()$). 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 (EM) algorithm (Dempster et al, 1977).

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:

$$\frac{\max_j(w_j)}{\min_j(w_j)} \geq \lambda, \quad (16)$$

where λ is a threshold. For our experiments, we choose $\lambda = 5$ 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.

3.5 Data Collection

The human teleoperator is able to control Robonaut's many degrees of freedom with a virtual reality-like helmet and a data glove equipped with a Polhemus sensor (Ambrose et al, 2000). In addition to articulating Robonaut's neck, the helmet provides visual feedback from the environment to the teleoperator. The arms and hands of the robot are commanded by tracking the movements of the human's wrists and fingers, and performing a mapping from human motion to robot motion.

Each trial consists of the human teacher haptically exploring an object for approximately 15 minutes. The object is located in a fixed pose relative to the robot. To maximize the number of quality samples collected, different grasping strategies may be employed by the teleoperator based on the local geometry of the object. For example, when grasping larger surfaces, a sliding motion in conjunction with a fixed finger configuration is used. This ensures that the feasible positions and orientations of the hand are collected in a timely manner. In contrast, the teleoperator repeatedly opens and closes the robot's hand when grasping small surfaces. This strategy forces hand pose to vary even though the hand may not be able to slide along the local surface.

When compared with the data collected during direct observation of a human performing grasping actions (de Granville, 2008; de Granville and Fagg, submitted), the robot teleoperation experience tends to contain larger amounts of noise. Robonaut's arm motions are slower and less fluid under human control. Hence, the hand posture samples contain a large number of cases in which the hand is not in contact with the object. To alleviate this problem the transitions are removed manually by identifying the time intervals in which they occur.

3.6 Experimental Results

To demonstrate the effects of incorporating finger configuration into the grasp learning process, a number of experiments are performed. First, the eigengraps are learned based on experience that is generated by the human teleoperator. A number of different objects are used to ensure a reasonable sampling of the finger configurations. Due to invalid sensor data, seven of the finger joints are ignored. This means that the number of effective degrees of freedom in Robonaut's hand has been reduced from twelve to five. Of these remaining five degrees of freedom, approximately 98% of the variance can be explained by the first three principal components. This, in conjunction with the ability to visualize the resulting low dimensional rep-

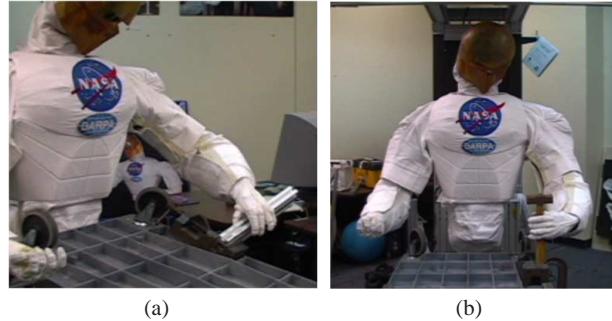


Fig. 9 The set of objects used in the Robonaut clustering experiments. (a) Handrail; (b) Hammer.

resentation of finger configuration, led to the use of only the first three eigengrasps (i.e., $K = 3$).

3.6.1 Handrail

Figs. 10(a,c,e) show the training examples for the handrail object. Panel (a) shows the 3D position of the hand throughout the course of the experiment, while panel (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. In both panels (a) and (c), the major axis of the handrail is located along the X axis, with the grasped end at $X = -60$ in the position space and at $X = -1$ in the orientation space. Finally, panel (c) shows the finger configurations projected into the eigengrasp space.

A total of five clusters were learned for the handrail object: two that correspond to an overhand reach in which the handrail is approached from the top, two for the underhand configuration, and one for the side approach. The learned position clusters are shown in Fig. 10(b) as first standard deviation ellipsoids of the Gaussian distribution. The orientation component of these clusters is represented using a Dimroth-Watson distribution and is shown in panel (d). The mean orientation is indicated using the line segment emanating from the center of the sphere. Clusters 1 and 4 correspond to the top approach and show an elongation in the position component along the X axis. This elongation encodes the fact that the top approach results in grasps at many points along the length of the handrail. Likewise, clusters 2 and 5 correspond to the underhand approach (with the palm up) and are also elongated along the handrail. Cluster 3 corresponds to the side approach to the handrail. The

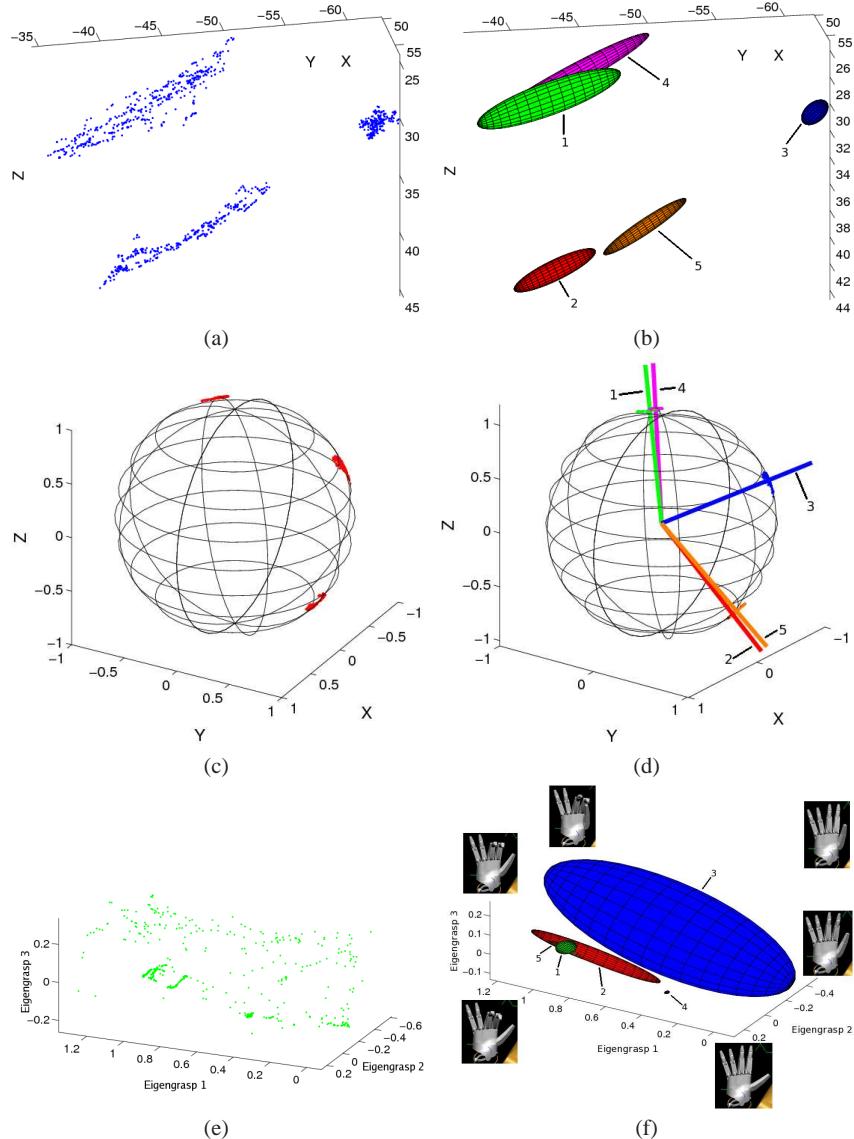


Fig. 10 The training examples and learned affordance model for the handrail. (a) The position of the hand; (b) The position component of the learned affordance model; (c) The orientation of the hand; (d) The orientation component of the learned affordance model; (e) The finger configuration of the hand; (f) The finger configuration component of the learned affordance model.

demonstrated variation in hand position was very small, and is reflected in the size of the ellipsoid.

Panel (f) shows the learned eigengrasp clusters. Each corner of the bounding box provides a visualization of the mapping that occurs between the low dimensional representation of finger configuration and each joint of the robot's hand. Notice that variation along the first eigengrasp corresponds to flexion of the index and middle fingers, while variation along the second eigengrasp causes adduction and abduction of the thumb. However, variation along the third eigengrasp does not affect the configuration of the fingers significantly, only affecting the flexion of the most distal joints of the index and middle fingers. Also, note that the ring and pinkie fingers remain in their extended configurations. (these are among the degrees of freedom for which no data were recorded).

Turning to the learned eigengrasp clusters, notice that the ellipsoids 1 and 2 are in the same region of the finger configuration space even though they correspond to grasp approaches from above and below the handrail. Because the same sliding technique was employed by the teleoperator when demonstrating these grasps, the hand had a similar shape for each approach. However, for cluster 2 there is more variation in finger configuration, which is indicated by the elongation of ellipsoid 2. In contrast, the hand was continually opened and closed when the side approach was used to grasp the handrail. This is evident by comparing the hand shapes that correspond to points on opposite ends of ellipsoid 3's major axis. On the right end of the figure, the hand is in an open configuration, but on the left end the middle and index fingers are flexed considerably. Also, notice that ellipsoid 3 is separated from the other eigengrasp clusters, which highlights the different hand shapes used when grasping the handrail from above and below versus from the side.

3.6.2 Hammer

The example grasps demonstrated by the human teleoperator and the learned grasp affordance model for the hammer are shown in Fig. 11. In this case five clusters were learned: cluster 1 represents grasps when approaching from above the hammer's head. The orientation of this cluster is represented using a girdle distribution, as indicated by the circle on the surface of the sphere in panel (d). In our visualization, the points along the circle correspond to the orientations of maximum density. For the case of cluster 1, we would have expected the use of a Dimroth-Watson distribution because there is little variation about the points corresponding to the top approach. However, what variation there is falls along a narrow arc that is best captured by the girdle.

The remaining clusters (2–5) capture grasps along the handle of the hammer when approaching from the side. Girdle distributions were selected to model the orientation of the hand for the side approach. While this is encouraging, the algorithm learned four clusters instead of one. This is most likely due to the spatially distinct hand positions used to grasp the hammer's handle.

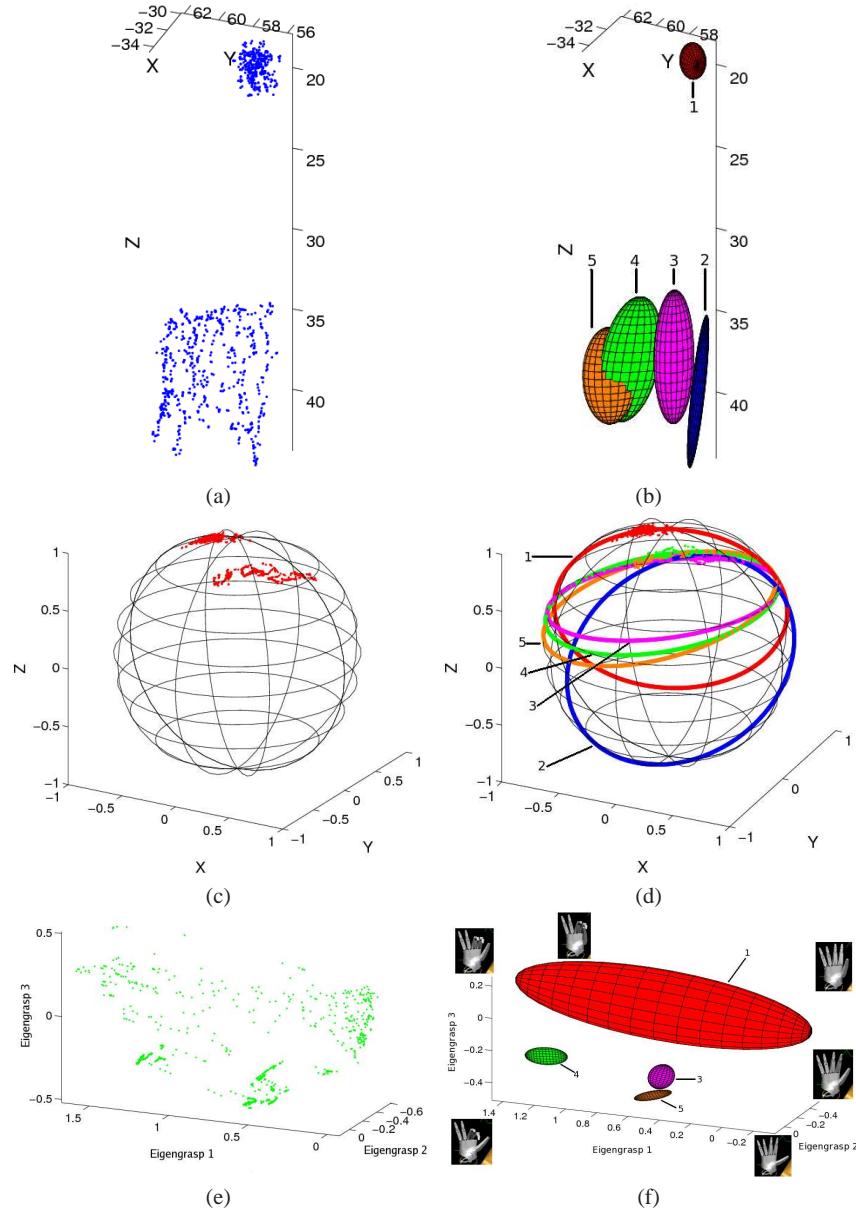


Fig. 11 The training examples and learned affordance model for the hammer. (a) The position of the hand; (b) The position component of the learned affordance model; (c) The orientation of the hand; (d) The orientation component of the learned affordance model; (e) The finger configuration of the hand; (f) The finger configuration component of the learned affordance model.

The finger configurations used by the teleoperator to grasp the hammer's handle were very different than those used to grasp the hammer's head. When approaching the object from the side, power grasps that maximized the contact surface area between the hand and the handle were more likely to be used. Conversely, precision grasps that mainly used the finger tips were employed when grasping the head of the hammer. These differences in hand shape can be seen in Fig. 11(f). Ellipsoid 1 represents the finger configurations used to grasp the hammer from above. The large volume of the ellipsoid is due to the exploration strategy employed by the teleoperator: the hand was continually opened and closed on this portion of the object. Hence, there was a large variance in finger configuration. Also, notice that ellipsoids 2–5 are spatially distinct from eigengrasp cluster 1.

4 Discussion

In this chapter, we presented several steps toward robot learning of affordance representations in support of grasping activities. Affordances provide a means of mapping sensory information, including vision, into a small set of grasping actions that are possible with the object being viewed. Key to this representation is the fact that it captures the specific interaction between the object and the agent. The ability to learn these representations automatically will be important as we begin to field robots in unstructured environments and expect them to perform a wide range of manipulation tasks.

Given a sequence of tuples consisting of an image and an object pose, our algorithm learns 3D appearance models for objects. In particular, the algorithm identifies visual operators that are robust to spurious image features such as object occlusions and shadows. Visual operators are implemented as edgel constellations that describe a specific geometrical relationship between a set of small edges. The 3D appearance of an object is captured by compactly describing the set of viewing angles from which each image feature is viewable. When a novel image is presented, the set of observed features can then be used to estimate the most likely viewing angle of the object. Ultimately, we will estimate the complete pose of the object, which, in turn, can be used for planning and executing grasping actions.

In more recent work, we have begun to make use of scale-invariant feature transform (SIFT) image features in place of edgels (Lowe, 2004). This method is showing promise in addressing image scale issues, improving the computational efficiency of identifying features, and increasing the accuracy of the viewing angle estimates. One of the challenges in using such an approach is that of pruning the set of primitive features that arise from such a large database of images that exhibit very similar appearance. In addition, we are now making use of a particle-based approach for describing the density functions. This approach is helping to address the overfitting issues that can arise with mixture-type models and allows us to capture irregular shapes in the density functions.

The second component of our algorithm uses tuples of object pose and hand posture to construct a small menu of grasps that are possible with the object. These compact representations are constructed from many example grasps made by clustering the hand posture examples. This property enables the use of the affordance as a way to access “primitives” in higher-level activities, including planning, learning, and the recognition of motor actions by other agents (Fagg et al, 2004; Brock et al, 2005). In particular, the hand posture 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 assumes that haptic exploration methods are available to refine the grasps once the hand has approached the object (Coelho and Grupen 1997; Platt et al., 2002, 2003; Platt 2006; Wang et al., 2007).

Our approach to date has assumed an intermediate representation between vision and grasp that is rooted in the individual objects. However, we would ultimately like for this representation to be able to generalize across objects. This step will be important as the robotic agent is faced with objects with which it has little to no prior experience. Our approach is to identify canonical grasps that routinely co-occur with particular visual features. When two or more objects have components that share a common shape, and hence common visual features, it is likely that similar hand postures will be used to grasp these components.

The 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 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. We also thank Brian Watson for his contributions to this work.

References

- Ambrose RO, Aldridge H, Askew RS, Burridge RR, Bluethmann W, Diftler M, Lovchik C, Magruder D, Rehnmark F (2000) Robonaut: Nasa’s space humanoid. *IEEE Intelligent Systems* 15(4):57–63
- 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
- Bingham C, Mardia KV (1978) A small circle distribution on the sphere. *Biometrika* 65(2):379–389
- 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, Fagg AH (submitted) Learning grasp affordances through human demonstration
- 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
- Hand DJ, Smyth P, Mannila H (2001) *Principles of Data Mining*. MIT Press, Cambridge, MA, USA
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110
- Mardia KV, Jupp PE (1999) *Directional Statistics*. Wiley Series in Probability and Statistics, Wiley, Chichester, West Sussex, England
- Piater JH, Grupen RA (2000) Feature learning for recognition with Bayesian networks. In: *Proceedings of the Fifteenth International Conference on Pattern Recognition*, Barcelona, Spain
- Piater JH, Grupen RA (2002) Learning appearance features to support robotic manipulation. In: *Proceedings of the Cognitive Vision Workshop*, electronically published

- Platt, Jr R (2006) Learning and generalizing control-based grasping and manipulation skills. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst
- 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
- Platt, Jr R, Fagg AH, Grupen RA (2003) Whole body grasping. In: Proceedings of International Conference on Robotics and Automation (ICRA'03)
- 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
- Wang D (2007) A 3D feature-based object recognition system for grasping. Master's thesis, School of Computer Science, University of Oklahoma, Norman, OK
- Wang D, Watson BT, Fagg AH (2007) A switching control approach to haptic exploration for quality grasps. In: Proceedings of the Workshop on Robot Manipulation: Sensing and Adapting to the Real World at the 2007 Robotics: Science and Systems Conference, electronically published