Learning Grasp Affordances Through Human Demonstration

Charles de Granville Joshua Southerland Andrew H. Fagg

Symbiotic Computing Laboratory
School of Computer Science
University of Oklahoma

***Abstract*— When presented with an object to be manipulated, a robot must consider the set of actions available for 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 corresponding orientation of the hand. Individual clusters of these three-dimensional orientations are represented using probability density functions that take on Gaussian-like shapes, and thus correspond to variations around canonical approach orientations. 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 (and functional) spatial relationship with the object.** [1]

***Index Terms*— Grasp affordance, learning from demonstration, clustering, mixture models, probabilistic densities of 3D rotations**

I. Introduction

Gibson suggested that objects in the environment can be represented by an agent in terms of the actions that can be made with respect to those objects. Furthermore, he suggested that these representations should be distinct from those that explicitly capture the physical properties or semantics of the objects [9], [10]. This *affordance* representation captures the combination of the relevant physical properties of the object and the capabilities of the agent's body. As a consequence, object properties that are not relevant to action are implicitly lost. A critical advantage of the separation of affordance from semantics is that of generalization: what one learns about interacting with an object can be separated from the specific task in which the interaction is taking place. Hence, when new tasks are presented, the skills of interacting with specific objects can be readily accessed and used. Furthermore, previously unknown objects can still be recognized in terms of the properties that are "of interest" to an affordance (e.g., the affordance might be concerned with objects that exhibit certain shape properties). This provides a way of accessing existing skills for use with the novel object.

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? Coelho, Piater, and Grupen have developed an approach to automatically learn the mapping between constellations of visual features, and hand orientation and configuration in a planar grasping task [5], [12]. A haptically-driven approach was used to explore a given object in order to find a set of finger contact locations that minimized the net force and torque that was applied to the object by the fingers (resulting in high quality grasp configurations) [6]. Given a set of object images and the identified stable grasp configurations, the visual learning algorithm attempted to find sets of geometrically-arranged image features (local edges and textures) that consistently predicted the relative orientation of the hand and the finger configuration. A candidate visual constellation was evaluated with respect to the set of example grasps by identifying clusters of relative hand orientations. When a small number of dense clusters was identified, the visual constellation was considered to be predictive of hand orientation. In novel situations, these affordance maps could then be used to position and configure the hand given the visual input.

Stoytchev recently presented a developmental approach to learning action sequences that results in a physical "binding" between robot and object [15]. The robot was initially endowed with a set of actions (involving rotations and translations of the arm, as well as opening and closing of the gripper) and a set of visual operators that were relativized to various coordinate frames attached to the robot. However, the robot had no explicit representation of hands or grasping. For each of several objects, the robot performed a random sequence of exploratory actions. Subsequences of actions that led to simultaneous movement of the object and a component of the robot were deemed as "interesting." Short subsequences that reliably achieved this interesting bound configuration were identified as viable grasping macro-actions, and were associated with the object through the affordance map. These macro-actions included movements of the arm that brought the hand into alignment with the object, and a

[1]This work is supported in part by NSF/CISE/REU (award #0453545) and by the University of Oklahoma.

subsequent hand motion.

In this paper, we explore the construction of grasp affordance representations based on demonstration by a human teacher. In particular, we focus on the problem of representing the orientation of the hand in three dimensions as it approaches the object. We cluster these points using a mixture distribution-based clustering method. Individual clusters of orientations are represented using probability density functions that have Gaussian-like shapes. Experimental results demonstrate the feasibility of extracting a compact set of canonical grasps from the human demonstration. These extracted grasps can then be used to parameterize controllers that are capable of driving a hand to an appropriate pose for grasping, or interpreting the actions of other agents in the environment.

II. Methods

In our experiment, a single demonstration trial consists of a human teacher haptically exploring an object, pausing briefly in configurations that correspond to quality grasps. During these trials, hand orientation is sampled at $15Hz$. Our goal is to compress this set of observations into a small number of clusters that are meaningful in terms of describing the functionally different ways that the object may be grasped. The first step in this process is to describe individual clusters of orientations.

A. Orientation Models

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 [11], [13]. The probability density function for this distribution is as follows:

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

where $\mathbf{u}$ is a unit quaternion that represents the "mean" rotation, $k \geq 0$ is a concentration parameter; and $F(k)$ is a normalization term. When $k = 0$, the distribution is uniform across all rotations; as k increases, the distribution concentrates about $\mathbf{u}$.

Figure 1, provides a 3D visualization of the Dimroth-Watson distribution, and highlights its Gaussian-like characteristics. Notice that in (a), the points deviate from $\mathbf{u}$ more so than in (b). This shows the effect of the concentration parameter k. As k increases, the data becomes more concentrated about the common axis of rotation. The figure also illustrates the distribution's antipodal symmetry. Points on opposite poles of the sphere are assigned equal densities.

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, resulting in a set of hand orientations that correspond to an arbitrary rotation about a fixed axis. This set of orientations falls on 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 [14]. The probability density function is as follows:

$$\bar{f}\left(\mathbf{x}|\mathbf{u_1},\mathbf{u_2},k\right)=\bar{F}\left(k\right)e^{k\left[\left(\mathbf{x}^{T}\mathbf{u_1}\right)^{2}+\left(\mathbf{x}^{T}\mathbf{u_2}\right)^{2}\right]}, \tag{2}$$

where $\mathbf{u_1}$ and $\mathbf{u_2}$ are orthogonal unit quaternions that determine the great circle, and $\bar{F}(k)$ is the corresponding normalization term.

Figures 1(c) and 1(d) illustrate the girdle distribution on S2 (the 3D sphere) for small and large values of k respectively. All points that fall on the great circle are assigned equal density, with density falling off with larger distances from the great circle. Note that for large values of k, the density becomes more concentrated about the great circle.

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, Λ:

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

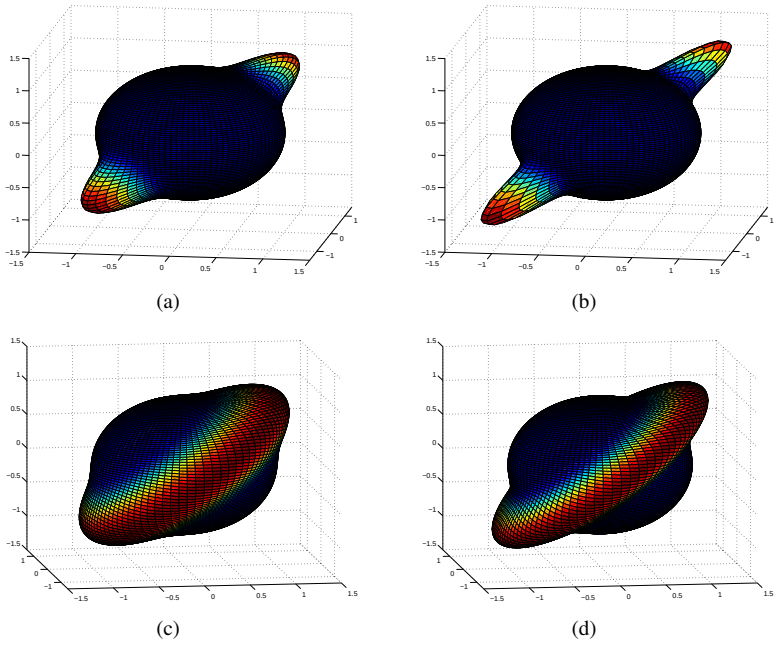
where $\mathbf{x_i}$ is the ith sample, and N is the total number of samples. The MLE of $\mathbf{u}$ is parallel to the first eigenvector of $\mathbf{\Lambda}$ [11], [13]. The orthogonal vectors $\mathbf{u_1}$ and $\mathbf{u_2}$ span the same space as the first and second eigenvectors of $\mathbf{\Lambda}$ [14].

For the Dimroth-Watson distribution, the MLE of the concentration parameter, k, uniquely satisfies the following:

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

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

$$\frac{\bar{F}'\left(k\right)}{\bar{F}\left(k\right)}=-\frac{\sum_{i=1}^{N}\left(\mathbf{x_i}^{T}\mathbf{u_1}\right)^{2}+\left(\mathbf{x_i}^{T}\mathbf{u_2}\right)^{2}}{N}. \tag{5}$$



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

Fig. 1. Three dimensional representations of the Dimroth-Watson and girdle distributions on S2. In all cases, the surface radius is $1 + p$, where p is the probability density at the corresponding orientation. (a) Dimroth-Watson, $k = 10$; (b) Dimroth-Watson, $k = 20$; (c) girdle, $k = 8$; and (d) girdle, $k = 16$.

B. Mixtures of Orientation Models

The Dimroth-Watson and girdle distributions perform well when describing a single cluster of points, but a set of grasps is typically fit best by a set of clusters. We therefore employ a mixture model approach. Here, the composite density function, $g(\mathbf{x})$, is defined as:

$$g(\mathbf{x}|...) = \sum_{j=1}^{M} w_j f_j(\mathbf{x}|...), \tag{6}$$

where

$$\sum_{j=1}^{M} w_j = 1, \tag{7}$$

and where M is the number of component distributions, and $f_j(\mathbf{x}|...)$ is the density function of either a Dimroth-Watson or girdle distribution (each with its own parameter set). Each element of the mixture represents a single cluster of points, and is weighted by w_j. Estimation of the weights is accomplished using the Expectation Maximization algorithm [7], and the parameters of the individual distributions are found by maximum likelihood estimation as described above.

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 fit a set of mixture models and to choose the one that best matches the observations. For this purpose, we make use of the Integrated Completed Likelihood (ICL) criterion [3] 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.

III. Experimental Results

In order to illustrate the capabilities of our orientation clustering approach, we performed multiple grasping experiments for a variety of objects (see Figure 2). Each object has its own unique set of grasps that may be modeled as a mixture of either Dimroth-Watson or girdle distributions (but not a mixture containing both distributions).

A human teacher wears a P5 glove [1] equipped with a Polhemus Patriot [2] (Figure 2(a)). Together, these components continuously capture hand pose and nine degrees-of-freedom of finger flexion at $15Hz$ (although, in this experiment, we do not make use of the flexion information). Each trial consists of approximately 40 seconds of haptic exploration of the (non-moving) object by the teacher. During the trial, the teacher largely maintains contact with the object in configurations that correspond to quality grasps of that object. In addition, the teacher does not approach her own joint limits (where necessary, large orientation changes are achieved by moving the torso or the entire body).

Due to our data collection procedure, some samples fall outside of quality grasps (and instead correspond to transitions between grasps). 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 mixtures models when an individual cluster covers a very small percentage of the

samples. In particular, a model is discarded when:

$$\frac{\max_j(w_j)}{\min_j(w_j)} \geq \lambda, \tag{8}$$

where λ is a threshold (for our experiments, we chose $\lambda = 5$). Of the models that have not been removed by this filter step, the one with the lowest ICL measure is considered to be the best explanation of the observed data set.

Because the EM algorithm is a gradient descent method in an error space containing many local minima, each mixture model was evaluated 80 times. The best performing model (according to ICL) was subsequently evaluated for filtering and comparison with other mixtures.

In the first experiment, we consider the rectangular prism shown in Figure 2(a). The object is approached from above, with the palm parallel to its top face. Figure 3(a) shows the performance of various mixture models created for the rectangular prism. Even though models containing a larger number of clusters result in a lower ICL measure, the two element Dimroth-Watson mixture is chosen due to the filtering process previously described (this choice is indicated by the large circle).

The set of observed orientations and this two-cluster solution is illustrated in figure 3(b). Orientation of the hand is represented as a single point on the surface of the unit sphere: imagine the object 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 representation aliases the set of rotations about the line perpendicular to the palm.

The two Dimroth-Watson clusters that were identified for this object correspond to the two "top" approaches that are possible with the rectangular prism (with the thumb on one side of the box or the other). The centroids of the two clusters are shown as line segments that pierce the surface of the sphere at the corresponding orientations. The geometry of the rectangular prism does not afford grasps with wide rotational variation. Thus, the mixture of two Dimroth-Watson distributions is appropriate.

For the case of a vertically-oriented cylinder (Figure 2(b)), a two-element girdle distribution was favored (Figure 4(a)). This is the case despite the better performance of the largest Dimroth-Watson mixture, which was discarded at the filtering step. The two clusters in the selected solution correspond to a top approach with the palm parallel to the top face of the cylinder, and a side approach with the palm parallel to the lateral surface of the cylinder. These solutions are indicated in Figure 4(d) as the thick circles that are drawn across the sphere (note that on S3, these circles are actually great circles that do not intersect).

For the square prism shown in Figure 2(c), the preferred solution included a total of four clusters, corresponding to four distinct top approaches (Figures 4(b) and 4(e)). These solutions correspond to the thumb being placed on each of the four sides of the prism. Since each grasp did not exhibit wide rotational variation, each cluster is modeled by a Dimroth-Watson distribution.

	DW2	DW3	DW4	DW6	G1	G2	G3
Rectangular prism	**9**				1		
Cylinder						**9**	1
Square prism			**9**		1		
Hexagonal prism				**0**	10		
Cup						**9**	1
Tape dispenser		**10**					

TABLE I

For the hexagonal prism shown in Figure 2(d), our desired outcome was to have a total of six Dimroth-Watson distributions, corresponding to top approaches with the thumb on one of each of the sides (similar to the square prism). However, as shown in Figure 4(c), this solution was removed from consideration in the filtering step. Instead, the preferred solution was a single girdle distribution, shown in Figure 4(f). This solution was selected over the desired solution because the amount of variation in orientation for individual sides of the hexagon approached the difference in orientation from one side to the next.

In order to understand the robustness of our approach, we performed ten replications for each of the objects already discussed and two additional objects: a cup and a tape dispenser (Figure 2). Each replication consisted of an independent demonstration by the teacher. The results are summarized in Table I. For each object, the number of occurrences of the solution preferred by our algorithm is listed (therefore, the sum across the rows is ten). Numbers in bold correspond to the solution "expected" by the authors. Columns labeled *DWi* are solutions involving i Dimroth-Watson distributions; columns labeled *Gi* correspond to girdle solutions.

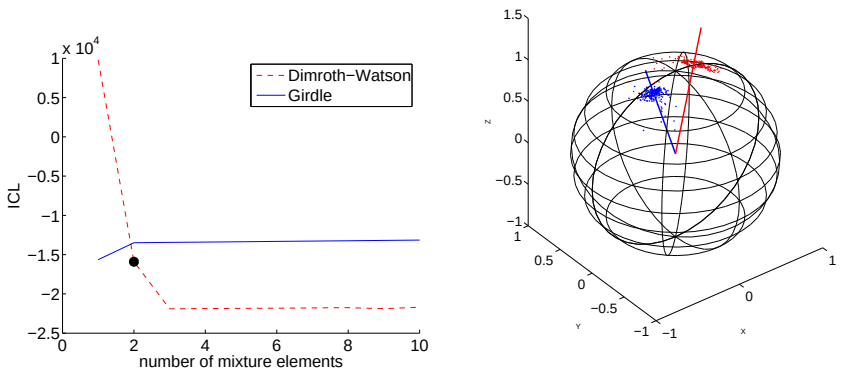
For the rectangular/square prisms, cylinder, and cup, our algorithm identified the expected models on 90% of the trials. In the case of the tape dispenser, the teacher consistently produced top-approach grasps with three distinct orientations. Our algorithm identified three Dimroth-Watson clusters in all ten trials. Finally, in all ten cases of the hexagonal prism, the algorithm preferred a single girdle distribution (as opposed to six Dimroth-Watson distributions).

IV. Discussion

In this paper, we have presented a technique for learning canonical hand orientations for reach-to-grasp actions. Compact representations are constructed from many example grasps made by a human teacher through an orientation clustering process. This represents a key step in learning complete grasp affordances that would also describe the position of the hand and configuration of the fingers. For a

(a) (b) (c) (d) (e) (f)

Fig. 2. Set of objects used in the clustering experiments: (a) a rectangular prism (with modified P5 glove), (b) cylinder, (c) square prism, (d) hexagonal prism, (e) cup, and (f) tape dispenser.



(a)

(b)

Fig. 3. Clustering results for a single rectangular prism demonstration. (a) ICL as a function of mixture size for both Dimroth-Watson (dashed) and girdle (solid) distributions. The best matching mixture distribution is indicated with a black circle. (b) Representation of the clustered hand orientations (individual orientation observations are represented as points on the surface of the unit sphere) and the average orientation for each of the two clusters (line segments emanating from the origin of the sphere). Hand orientation is represented as follows: the object is located at the origin of the sphere, with the palm of the hand tangent to the sphere. In this case, the two clusters correspond to the two possible top approaches to the object.

given object, we want this 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 recognition of motor actions by other agents [4], [8]. In particular, the clusters that have been learned map directly onto resolve-rate controllers that can bring a robot hand to a specific orientation (defined by the mean orientation vector). These controllers can be formulated to handle the don't-care orientation dimension of the girdle distribution (e.g., bringing the hand to a pose that enables a cup to be picked up from the side, but not specifying where along the side). [2]

There are several limitations to our approach that we are addressing in the current work. First, an assumption has been made that the pose of the hand has been captured within an object-centered coordinate frame. Because an individual object did not move during a single data collection trial, it was possible to work equivalently within a global coordinate frame. We are taking steps to track objects as they are moved during the haptic exploration process, and we are investigating techniques for aligning experience across independent trials that may individually be incomplete.

A second limitation is due to our focus on clusters in orientation space. This simplification is feasible when the objects can be approximated by simple geometric primitives. However, interesting objects involve a dependence between orientation and position of the hand. In continuing work, we plan to explicitly capture this dependence with probability distributions over both position and orientation.

Finally, our analysis has focused entirely on the pose of the hand. In future work, we plan to also take into account the positions of the finger tips relative to the hand coordinate frame.

[2]This is accomplished by setting the elements of the appropriate row of the hand Jacobian to zero.

REFERENCES

[1] –. Essential Reality, Inc.
[2] –. Polhemus. Colchester, VT.

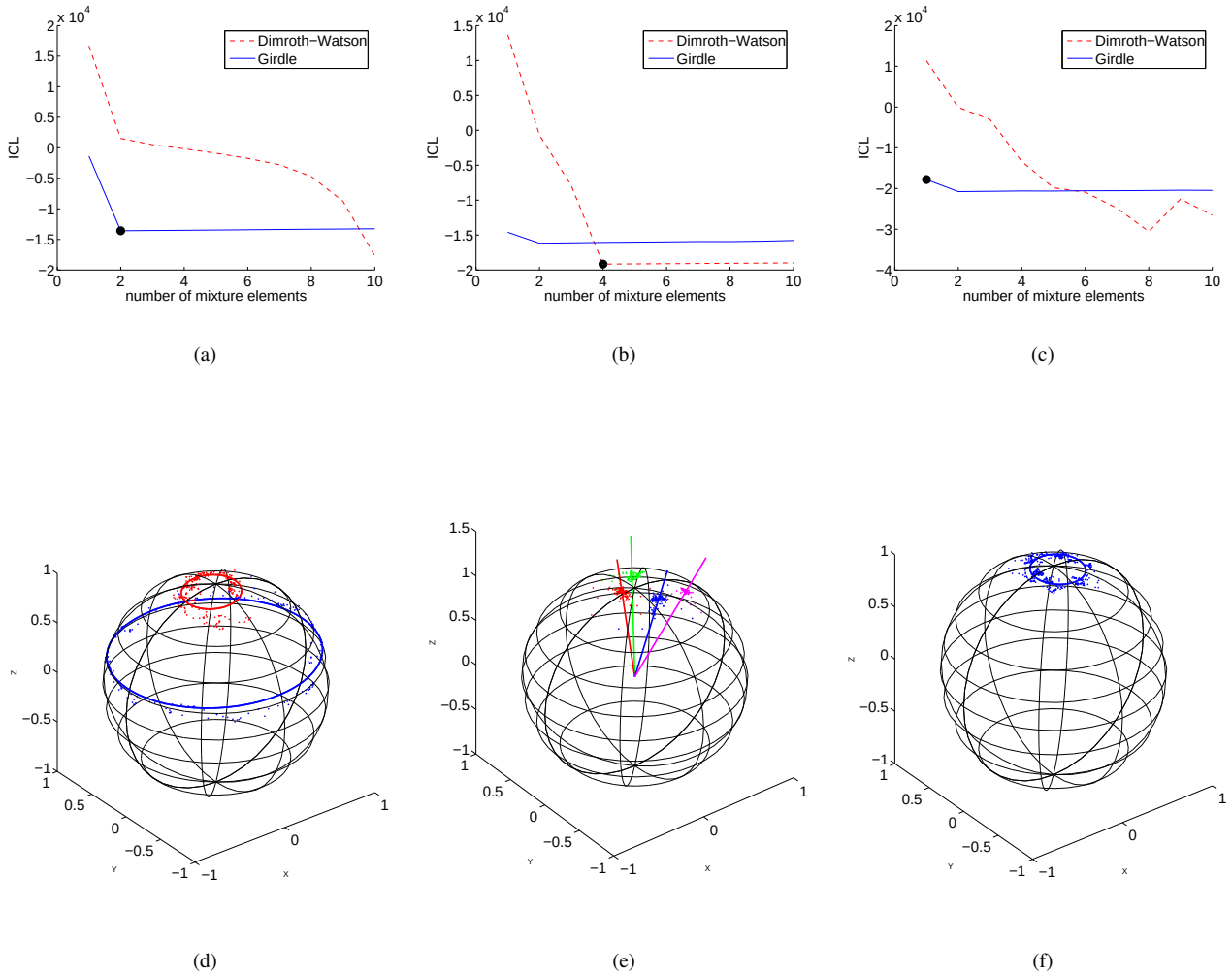


Fig. 4. Mixture model results for (a,d) a vertically-oriented cylinder; (b,e) a square prism; and (c,f) a hexagonal prism. Figure notation is identical to that of Figure 3. The cylinder is allocated two girdle distributions corresponding to a top and a side approach; the square prism is allocated four Dimroth-Watson distributions; and the hexagonal prism is allocated one girdle distribution.

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

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

[5] J. Coelho, Jr., J. Piater, and R. Grupen. 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, November 2000.

[6] Jefferson A. Coelho, Jr. and Roderic A. Grupen. A control basis for learning multifingered grasps. *Journal of Robotic Systems*, 14(7):545–557, 1997.

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

[8] A. H. Fagg, M. T. Rosenstein, R. Platt, Jr., and R. A. Grupen. Extracting user intent in mixed initiative teleoperator control. In *Proceedings of the American Institute of Aeronautics and Astronautics Intelligent Systems Technical Conference*, 2004.

[9] J. J. Gibson. *The Senses Considered as Perceptual Systems*. Allen and Unwin, 1966.

[10] J. J. Gibson. The theory of affordances. In R. E. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing*. Lawrence Erlbaum, Hillsdale, 1977.

[11] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 1999.

[12] J. Piater and R. Grupen. Learning appearance features to support robotic manipulation. In *Proceedings of the Cognitive Vision Workshop*, ETH Zurich, 2002.

[13] D. Rancourt, L.-P. Rivest, and J. Asselin. Using orientation statistics to investigate variations in human kinematics. *Applied Statistics*, 49(1):81–94, 2000.

[14] L.-P. Rivest. A directional model fo the statistical analysis of movement in three dimensions. *Biometrika*, 88(3):779–791, 2001.

[15] A. Stoytchev. Toward learning the binding affordances of objects: A behavior-grounded approach. In *Proceedings of the AAAI Spring Symposium on Developmental Robotics*, 2005.