Learning Visual Features to Recommend Grasp Configurations

Justus H. Piater

CMPSCI Technical Report 2000-40

July 2000

Computer Science Department
Lederle Graduate Research Center
University of Massachusetts
Amherst, MA 01003-4601

{piater|grupen}@cs.umass.edu

This is a revised version of a paper titled "Learning Visual Features to Predict Hand Orientations", presented at the ICML-2000 Workshop on Machine Learning of Spatial Knowledge, Stanford, CA, July 2, 2000.

Learning Visual Features to Recommend Grasp Configurations

Justus H. Piater PIATER@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA

Abstract

This paper is a preliminary account of current work on a visual system that learns to aid in robotic grasping and manipulation tasks. Localized features are learned of the visual scene that correlate reliably with the orientation of a dextrous robotic hand during haptically guided grasps. On the basis of these features, hand configurations are recommended for future gasping operations. The learning process is instance-based, on-line and incremental, and the interaction between visual and haptic systems is loosely anthropomorphic. It is conjectured that critical spatial information can be learned on the basis of features of visual appearance, without explicit geometric representations or planning.

1. Introduction

When a human reaches for an object, the hand is oriented and shaped appropriately in anticipation of the grasp. This anticipatory preconfiguration takes place before contact with the object is made, and is informed by visual cues. For example, when reaching for a vertically oriented rod, during the reach phase the hand forms a vertical opening of a width relating to the perceived diameter of the rod. Once contact is made, haptic feedback dominates the grasping activity, while visual information generally plays a subordinate role.

It is not conclusively known what visual information is extracted and how it is used to inform the reaching process. In most conventional robotic approaches, geometric information is extracted from the scene using cameras, followed by a path/task planning procedure in the modeled environment. More recently, task metrics are computed directly in image space, extracting only the required information without explicit 3-D reconstruction (Jägersand & Nelson, 1995), or in a global appearance space without local feature extraction (Nayar et al., 1994). However, these approaches are not generally practical if the number of degrees-of-freedom is large and the task is complex.

It appears that the inherent problems of multiple degrees of freedom and high task complexity can only be overcome through task decomposition and learning. Specifically, manual grasping seems effortless to humans because we have substantial experience grasping objects. We do not need to plan each grasping process in detail. Rather, a few critical pieces of visual information prime the hand-arm system for an efficient execution. Haptic information along with a wealth of world knowledge permits efficient manipulation with minimal planning or metric data.

This paper describes current research on learning grasping cues purely from aspects of appearance, without any explicit scene reconstruction or geometric reasoning, which increases the efficiency of a haptically-guided grasping process . The idea is to learn localized features in image space that robustly predict relevant grasping parameters resulting in a successful grasp. Learning is on-line and incremental; there is no distinction between learning and performance phases.

2. Scenario

We have a functioning closed-loop haptically-guided grasping system that is able to execute grasps. Using a conventional visual servoing procedure, the hand is lowered down to the object, and the fingers probe the object surface in a systematic way until a stable grasp is formed using two or three fingers, without using any visual or geometric prior information about the object (Coelho & Grupen, 1997; Coelho & Grupen, 2000). This system is implemented on a General Electric P-50 industrial robot arm equipped with a Salisbury dextrous hand (Fig. 1). Because many tactile probes must be executed very carefully to avoid disturbing the object, a single grasp can take many minutes, especially if the initial finger positions are far away from a stable grasp.

This work adds eyes to the grasping system that watch the execution of many grasps. The goal is to learn oriented appearance-based features in image space that robustly correlate with the orientation of the hand during a successful grasp. Then, these features can be used to recommend hand orientations before tactile contact is made, ideally bringing the fingers close to an optimal grasping configuration. The

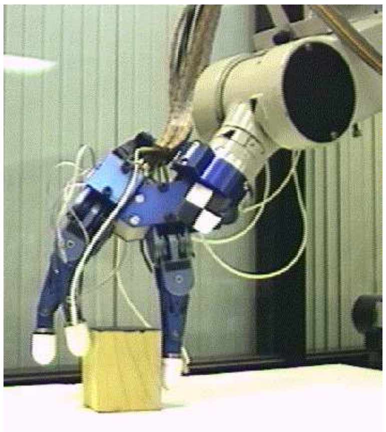

Figure 1. Our haptic grasping system.

orientation of the hand during a given grasp configuration, θ_H, is defined as illustrated in Fig. 2.

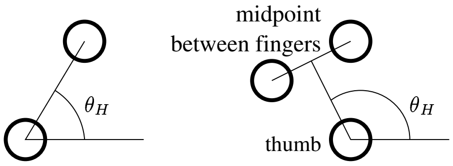

Figure 2. Definition of the hand orientation for two- and three-fingered grasps.

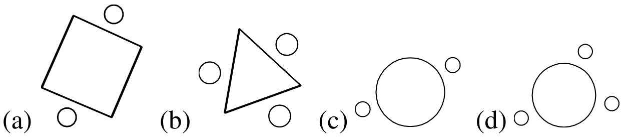

Figure 3. Some objects are better grasped with two fingers (a), some with three (b), and for some this choice is unimportant (c, d).

The robot may encounter a variety of objects that differ in their shapes. Each type of object may require a dedicated feature to recommend a hand orientation. Object identities are not known to the system; the need for dedicated features must be discovered by grasping experience. Moreover, a recommendation is made regarding the number of fingers to use for a grasp (two or three; see Fig. 3).

3. Features

Our earlier work on discrimination learning was based on visual features composed of local appearance descriptors defined by various Gaussian-derivative kernels (Piater & Grupen, 1999; Piater & Grupen, 2000). Here, similar features are used to predict a continuous orientation parameter (as opposed to a categorical class designation). A feature is either a *primitive* or a *compound* feature. A primitive feature is defined by a vector of locally computed Gaussian-derivative filter responses. There are two variants that differ in the number of filters employed: An *edgel* is encoded as a 2-vector containing the filter responses to the two first-derivative basis filters. It represents the magnitude and orientation of a localized spatial intensity gradient. A *texel* is represented as an 18-vector consisting of the responses to the basis filters of the first three derivatives at two scales. This represents a local texture signature. Like edgels, texels have an associated orientation θ that is defined by the first derivatives.

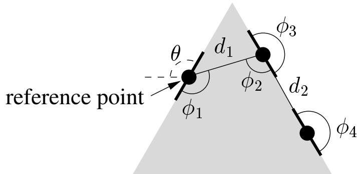

Figure 4. A geometric compound feature of order 3, composed of three primitives. The feature is defined by the angles ϕ and the distances d, and the orientation of this specific instance is denoted by θ. Each primitive is either an edgel or a texel.

Primitive features cannot generally be expected to correlate robustly with object orientation because they are likely to respond strongly to irrelevant parts of a scene. Compound features consisting of several primitives can potentially be much more specific to relevant object parts. Here, a compound is defined as illustrated in Fig. 4.

Each feature $\mathbf{f}$ is present at a pixel location l to a degree $s_{\mathbf{f}}(l) \in [0, 1]$, which is the normalized inner product of the vector of applicable filter responses at l with the pattern vector defining $\mathbf{f}$:

$$s_{\mathbf{f}}(l) = \max \left\{ 0, \frac{\mathbf{f}_l \mathbf{f}_p^T}{||\mathbf{f}_l||\ ||\mathbf{f}_p||} \right\}$$

The normalized inner product is pleasing in that it returns the cosine of the angle between the vectors in question, such that $s_{\mathbf{f}}(l) = \max\{0, \cos(\theta_l - \theta_p)\}$ for edgels $\mathbf{f}$. A feature is present in an image I to the degree $s_{\mathbf{f}} = \max_{l \in I} s_{\mathbf{f}}(l)$. For more detail on these features, see our earlier work (Piater & Grupen, 1999).

4. Using Features to Recommend Grasp Parameters

As the camera observes the objects and associated hand orientations, features are sampled from images taken by an overhead camera. Assuming that these features respond to the object itself, their image-plane orientation $\theta_{\mathbf{f}}$ should be related to the robotic hand orientation θ_h by a constant additive offset $\Delta\theta$. A given feature, measured during many grasping tasks, hence generates data points that lie on straight lines in the toroidal space spanned by the hand and feature orientations (Fig. 5). Here, $\Delta\theta = \theta_{\mathbf{f}} + \theta_h$. There may be more than one straight line because a given visual feature may respond to more than one specific object location (e.g., due to object symmetries), or to several distinct objects that differ in shape. Given $\Delta\theta$, one can then infer appropriate hand orientations as a function of observed feature orientations:

$$\theta_h = \Delta\theta - \theta_{\mathbf{f}} \tag{1}$$

The remaining problem is to find the offsets $\Delta\theta$. This is an instance of a K-Means problem in one-dimensional circular (angular) space, with K unknown.

4.1 Fitting a Parametric Orientation Model

Assume the $\Delta\theta$ are drawn independently from a mixture of *von Mises* distributions. The circular von Mises distribution can be regarded as corresponding to the linear Gaussian distribution, and has the probability density function (Fisher, 1993)

$$f_{\mathrm{vM}}(\theta|\mu,\kappa) = \frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)} \tag{2}$$

where $0 \leq \mu < 2\pi$, $0 \leq \kappa < \infty$, and

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos(\phi-\mu)}\,\mathrm{d}\phi$$

is the modified Bessel function of order zero. The mean direction of the distribution is given by μ, and κ is a concentration parameter with $\kappa = 0$ corresponding to a circular uniform distribution, and $\kappa \to \infty$ to a point distribution. The mixture distribution (see Fig. 5) is defined by its density function

$$f_{\mathrm{mix}}(\theta) = \sum_{k=1}^{K} p_k f_{\mathrm{vM}}(\theta|\mu_k,\kappa_k) \tag{3}$$

with mixture proportions $0 < p_k < 1$, $\sum_k p_k = 1$.

For all plausible numbers of clusters K, a $(3K-1)$-dimensional non-linear optimization problem is solved to find the μ_k, κ_k and p_k. The objective function to be maximized is the log-likelihood of the observed data Θ given a parameterization $\mathbf{a}$ consisting of the μ_k, κ_k and p_k:

$$\log P(\Theta|\mathbf{a}) = \sum_i \log \sum_{k=1}^{K} p_k f_{\mathrm{vM}}(\theta_i|\mu_k,\kappa_k) \tag{4}$$

The most probable model can then be found using Bayes' Rule:

$$P(\mathbf{a}|\Theta) = \frac{P(\Theta|\mathbf{a})P(\mathbf{a})}{\sum_m P(\Theta|\mathbf{a}_m)P(\mathbf{a}_m)}$$

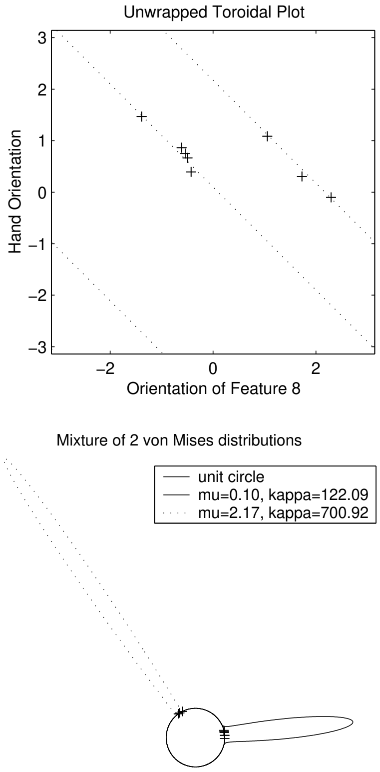

Figure 5. Top: Data points induced by a given feature on various images of an object form straight lines on a torus (two in this case). Bottom: A mixture of two von Mises distributions was fit to these data. The probability density at an angle is visualized by the distance of the line from the unit circle.

In the case of uniform prior probabilities over all possible model parameterizations $\mathbf{a}_m$, the model $\mathbf{a}$ maximizing $P(\mathbf{a}|\Theta)$ is simply the one that maximizes $\log P(\Theta|\mathbf{a})$ (Eqn. 4).

The appropriate number of clusters K is determined according to the Integrated Completed Likelihood criterion (Biernacki et al., to appear), an adaptation to clustering problems of the more well-known Bayesian Information Criterion (Schwarz, 1978).

While the system learns and performs, all features are evaluated on all images. The response strengths of all features, their orientations, the actual hand orientations θ_H (the training signal), and the prediction errors $e_{\mathbf{f}} = |\theta_H - \theta_h|$ produced by each feature are stored in an *instance list.* To compute the mixture model for each feature, this feature's data points (such as those shown in Fig. 5) are taken from this instance list.

If different types of objects are encountered, dedicated features may have to be learned. Without a supervisor providing object identities, the data collections (Fig. 5) will be cluttered up with inappropriate feature responses, and any reliable patterns will be obscured. The key to learning dedicated features is to ignore data points corresponding to weak feature responses. This permits features to emerge that respond strongly only to specific, highly characteristic object parts, but that respond weakly in any image that does not contain such an object. These weak responses will be ignored, and reliable models of $\Delta\theta$ can be fitted to the strong responses.

Deciding whether a given point is "strong" in this sense involves a threshold. Such thresholds $t_{\mathbf{f}}$, specific to each feature, can be determined optimally in the Bayesian sense that the number of poor recommendations made by the resulting model is minimized. To do this for a given feature $\mathbf{f}$, the strengths $s_{\mathbf{f}}$ of experienced responses and the associated prediction errors $e_{\mathbf{f}}$ are analyzed in order to find a threshold $t_{\mathbf{f}}$ such that most predictions for cases with $s_{\mathbf{f}} > t_{\mathbf{f}}$ are correct. To formalize this intuitive notion, we introduce a threshold t_e, meaning that a prediction with $e_{\mathbf{f}} < t_e$ is correct, and false otherwise. We can then define the optimal threshold $t_{\mathbf{f}}$ as a value that maximizes the Kolmogorov-Smirnoff distance[1] $\mathrm{KSD}_{\mathbf{f}}$ between the two conditional distributions of $s_{\mathbf{f}}$ under the two conditions that the associated predictions are correct and false, respectively (Fig. 6).

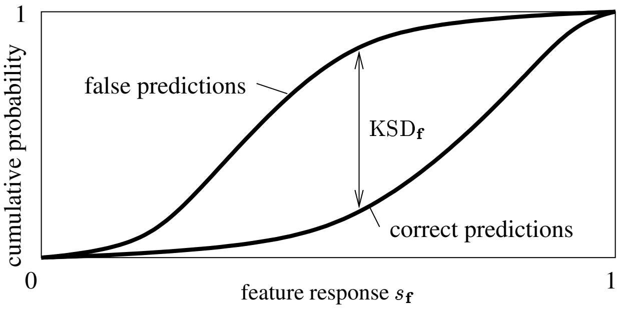

Figure 6. Kolmogorov-Smirnoff distance $\mathrm{KSD}_{\mathbf{f}}$ between response strength probabilities given correct and false predictions.

4.2 Operation

The system operates by recommending hand orientations and observing the outcomes of grasping procedures. Before a grasp, the responses of all features are measured in an overhead image of the object on a table. The system then considers all features $\mathbf{f}$ with $s_{\mathbf{f}} > t_{\mathbf{f}}$. From these, the feature with highest prediction potential $\mathrm{KSD}_{\mathbf{f}}$ is chosen. This is the feature that maximizes the expected correct prediction rate, based on all experience recorded in the instance list. If the mixture model corresponding to this feature has more than one mode k that is supported by at least three data points, the mode with maximal κ_k is selected. A hand orientation is computed according to Eqn. 1, using the $\Delta\theta$ corresponding to the selected model k.

The robot subsequently executes the grasp, starting with the recommended hand orientation. If the hand orientation turns out not to be appropriate, i.e. it needs to be corrected by more than t_e, then all $\mathrm{KSD}_{\mathbf{f}}$ are recomputed, and all mixture models are re-estimated based on the cases recorded in the instance list of previous experiences. A new prediction is made based on the new models. If this new prediction is still wrong, then two new features are generated: One primitive feature is randomly sampled from the image, and one compound feature is generated by randomly expanding an existing feature by adding a new point as illustrated in Fig. 4.

Many of the randomly sampled features will perform poorly, e.g. because they respond to parts of the scene unrelated to the object to be grasped. Such features will develop a poor KSD, as there is no systematic association between their response strengths and their prediction accuracies. Due to their low KSD, such features will cease to be used at all. Unused features are discarded periodically. On the other hand, if a feature performs well, its KSD will be increased, and it will be more likely employed. Moreover, since only super-threshold features are consulted, features can be learned that selectively respond to different object shapes requiring different offsets $\Delta\theta$.

This search for good predictive models is an instance of the Expectation-Maximization (EM) algorithm. The parametric model to be optimized is the collection of all feature-specific models that are used by the angular recommendation process. The hidden parameters of the model determine which recorded data points should participate in the feature-specific models. At the Expectation step, these hidden parameters are estimated by computing the $\mathrm{KSD}_{\mathbf{f}}$ such that the probability of making the right choice for each data point is maximized, given the current model. At the Maximization step, the probability of the model given the participating data points is maximized by optimizing the model parameters according to Eqn. 4. As the system operates, these two steps alternate.

This instance of the EM algorithm is non-standard as the Expectation step is not executed using all available current data. Instead, the instance list of past experience is consulted for previous prediction results, which were generated by models derived from all data available *at that time*. Taking the correct expectation using the most recent model would involve revisiting all previously seen images at each

[1] The Kolmogorov-Smirnoff distance between two distributions of a variable is the maximum difference between the two cumulative distributions, which occurs at some threshold value t of that variable. Given a quantity that was drawn from one of the two distributions, and one has to guess the correct distribution on the basis of this quantity, then guessing on the basis of t maximizes the probability of guessing correctly.

execution of the Expectation step, which is clearly impractical. Nevertheless, the convergence properties of the EM algorithm are unaffected. As data accumulate, the accuracy of recent expectations can only increase, and the influence of possibly inaccurate data from early history diminishes.

4.3 Predicting the Quality of a Grasp

The quality of a grasp is defined by the minimum friction coefficient μ_0 required to execute the grasp using a given finger configuration. The lower (closer to zero) a value of μ_0, the better the grasp. It is not possible to separate good from poor grasp configurations based on a generic threshold on μ_0 because the best achievable grasp depends on the object properties and the number of fingers. For example, the best achievable grasp of a triangular prism using two fingers is far worse than if three fingers are used. Cubes are best grasped with two fingers because of their parallel opposing surfaces.

It is possible to recommend the number of fingers to be used in a grasp, based on the expected μ_0 associated with a feature. To do this, separate models are learned as described above for each available number of fingers. The actually experienced value of μ_0 is stored in the instance list along with each executed grasp. These μ_0 values are regarded as samples of a continuous random variable M_0 with probability density function $f(\mu_0)$ and expected value

$$E[M_0] = \int_{-\infty}^{\infty} \mu_0 f(\mu_0)\, \mathrm{d}\mu_0.$$

Observe that each sample μ_{0i} was generated by a specific grasp with hand orientation θ_i. Therefore, the μ_{0i} are governed by the same probability density function as the θ_i, i.e., $f(\mu_{0i}) = f_{\mathrm{vM}}(\theta_i)$. Then, a sample estimate of $E[M_0]$ for a given feature $\mathbf{f}$ can be computed as

$$\hat{E}[M_0] = \frac{\sum_i \mu_{0i} f_{\mathrm{vM}}(\theta_i)}{\sum_i f_{\mathrm{vM}}(\theta_i)} \tag{5}$$

where the corresponding pairs of μ_{0i} and θ_i are taken from the instance list, using only instances corresponding to super-threshold occurrences of $\mathbf{f}$. When recommending grasp parameters, the system then derives an expected hand orientations for each available number of fingers, and recommends that with the lowest associated value of $\hat{E}[M_0]$.

5. Experiments

A series of pilot experiments was performed in simulation, using data generated by the real grasping system and by a detailed simulator, using photo-realistically rendered and noise-degraded images. Three object types were used (Fig. 7). Lacking the ability to perform large numbers of grasps on the real robot, the recommended grasps were

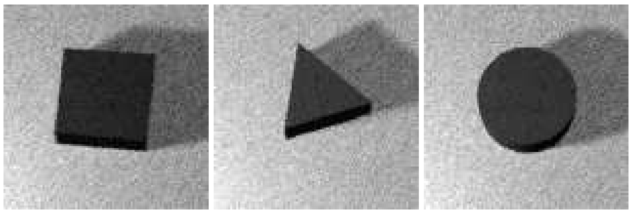

Figure 7. Example views of objects used to test the system.

simulated by comparing the recommended hand orientation with the previously executed hand orientation associated with the training image, modulo the known rotational symmetry properties of the object. Since cylinders have infinite-fold rotational symmetry, no features were ever learned for cylinders. All results reported below were computed in two-fold cross-validation.

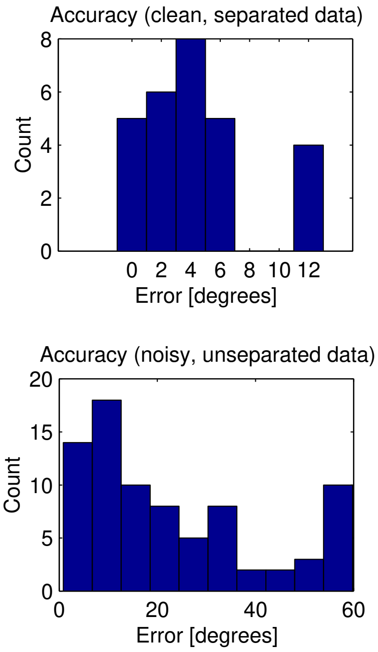

Figure 8. Quantitative results of hand orientation prediction.

Our pilot studies indicate that the system learns to make useful recommendations (Figure 8). If the training set contains a single object class and little noise in the training signal (the actual hand orientation θ_H during the grasp), the training set is typically learned during a single iteration. Performance on an independent test set is almost always excellent, with prediction error magnitudes on the order of the variation in the training signal.

If the training set contains outliers, i.e. hand orientations that produced a poor grasp, then the training set is harder to learn because the system expends a lot of effort trying to learn these outliers. However, performance degrades

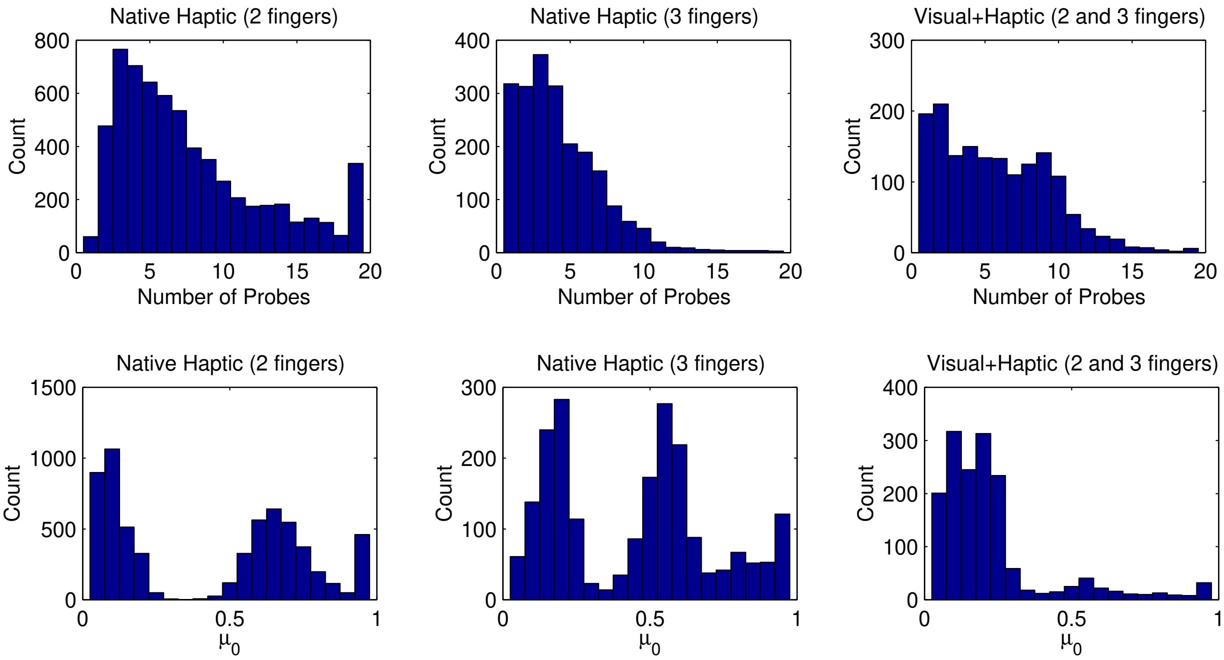

Figure 9. Utility of the learned visual context to the haptic system when grasping rectangular and triangular prisms (see text).

gracefully because features are selected by Kolmogorov-Smirnoff distance, which prefers generic features that work well for the majority of useful training examples. On a noisy test set, most poor recommendations occur on outliers. Notably, two-fingered grasps of the triangular object are inherently unstable and unpredictable. Here, prediction errors produced by the trained system depend on the error threshold that divides "good" from "poor" predictions during training. Choosing a low threshold generally produces more accurate predictions on a test set, as long as this threshold is larger than the variation contained in the majority of the training data.

Figure 9 demonstrates the utility of the learned visual context to the haptic grasping system when grasping rectangular and triangular prisms. The two columns on the left show the performance of the two- and three-fingered "native" controllers without a visual component. The rightmost column shows the performance achieved if the visual system determines the initial hand orientation, and which of the two native controllers to employ.

The bottom row illustrates that neither two- nor three-fingered native controllers alone are sufficient to execute high-quality grasps reliably. The two-fingered native controller works well on rectangular but poorly on triangular prisms; for the three-fingered controller the opposite is true. If the recommendation of the visual system is followed, the achieved grasp quality is consistently high – almost all of the μ_0 values cluster around the lower end of the range. Moreover, the proportion of extremely fast single-probe grasps increases drastically, and very long trials (more than about 20 probes) are practically eliminated (cf. the two-fingered native controller on the left).

6. Discussion

This paper describes a system for learning to recommend hand orientations and finger configurations to a haptically-guided grasping system. Localized appearance-based features of the visual scene are are learned that correlate reliably with observed hand orientations. In this way, visual guidance takes place without prior knowledge, explicit segmentation or geometric reconstruction of the scene. The interaction between haptic and visual system is a plausible model of human grasping behavior. Learning is on-line and incremental; there is no distinction between learning and execution phases. Pilot results were presented that demonstrate the operation of the system. More rigorous evaluation is currently underway.

In the future, this method will be extended to learning spatial knowledge in more degrees of freedom. Learned spatial knowledge about objects and environments is critical to the efficient sensorimotor activity of humans. It is known that much of the spatial information extracted by the human visual system is more qualitative than geometric. Future research will further explore the promise of feature-based learning of spatial information in more elaborate scenarios.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CISE/CDA-9703217, IRI-9704530 and IRI-9503687, and by the Air Force Research Labs, IFTD (via DARPA) under grant F30602-97-2-0032.

References

Biernacki, C., Celeux, G., & Govaert, G. (to appear). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Coelho, Jr., J. A., & Grupen, R. A. (1997). A control basis for learning multifingered grasps. *Journal of Robotic Systems*, *14*, 545–557.

Coelho, Jr., J. A., & Grupen, R. A. (2000). Learning in nonstationary conditions: A control theoretic approach. *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press.

Jägersand, M., & Nelson, R. (1995). Visual space task specification, planning and control. *Proc. IEEE Symposium on Computer Vision* (pp. 521–526).

Nayar, S. K., Murase, H., & Nene, S. A. (1994). Learning, positioning, and tracking visual appearance. *Proceedings of the International Conference on Robotics and Automation*. San Diego, CA.

Piater, J. H., & Grupen, R. A. (1999). Toward learning visual discrimination strategies. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 410–415).

Piater, J. H., & Grupen, R. A. (2000). Constructive feature learning and the development of visual expertise. *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*, 461–464.