Acquiring State from Control Dynamics to Learn Grasping Policies for Robot Hands

R.A. Grupen **J.A. Coelho, Jr.**
Department of Computer Science – Laboratory for Perceptual Robotics
140 Governors Dr., Amherst, MA, USA 01003
{grupen, coelho}@cs.umass.edu

submitted as: REGULAR PAPER (#01007)

Abstract

A prominent emerging theory of sensorimotor development in biological systems proposes that control knowledge is encoded in the dynamics of physical interaction with the world. From this perspective, the musculoskeletal system is coupled through sensor feedback and neurological structure to a nonstationary world. Control is derived by reinforcing and learning to predict constructive patterns of interaction. We have adopted the traditions of dynamic pattern theory in which behavior is an artifact of coupled dynamical systems with a number of controllable degrees of freedom. For grasping and manipulation, we propose a closed-loop control process that is parametric in the number and identity of contact resources. We have shown previously that this controller generates a necessary condition for force closure grasps. In this paper, we will show how control decisions can be made by estimating patterns of membership in a family of prototypical dynamic models. A grasp controller can thus be tuned on-line to optimize performance over a variety of object geometries. This same process can be used to match the dynamics to several previously acquired haptic categories. We will illustrate how a grasping policy can be acquired that is incrementally optimal for several objects using our Salisbury hand with tactile sensor feedback.

1 Introduction

The human hand has often been cited as an important factor in the development of the ability of the human brain to form critical categories in sensorimotor experience. Many are of the opinion that this faculty for building predictive models underlies much of what we recognize as human-level cognitive ability. While experts disagree on cause and effect, it is clear that the mechanical dexterity and redundancy afforded in the hand requires a neural architecture capable of modeling a huge variety of interactions with the world. We propose a representation couched in models of interaction dynamics and focus on finding features in this representation that predict important aspects of the control state. This kind of sensor and motor abstraction supports *prediction* in a control framework and may apply equally well to more general problem solving activities.

We pose the development of robot programs as an incremental search for strategies that exploit the intrinsic

dynamics of the robot/world interaction. "Intrinsic dynamics" is interpreted fairly broadly as any kinematic, dynamic, perceptual, or motor synergy that produces characteristic and invariant temporal sequences of observable features. The range of interaction possible and the kinds of perceptual distinctions required challenge commonly used methodologies for control and programming. In this paper, we provide a biologically inspired account of how sensorimotor strategies are formed. The process is constrained to employ a compact instruction set in the form of a closed-loop control basis. A Partially Observable Markov Decision Problem (POMDP) is designed to select a policy from the set of possible motor programs. The state of the POMDP is a boolean pattern describing the probabilistic membership of the last n-observations to a set of prototypical dynamic models. Dynamic models grow as necessary through long-term interaction with the domain and involve rich associations across sensor modalities [28, 11, 46, 19]. These generative models allow the agent to consider all possible outcomes and select the control sequence that yields the largest future reward. The representation proposed is applied to the multifingered grasping problem in which the agent must grasp objects with unknown geometries, independent of the initial relative orientation between hand and object. The results obtained show that the representation proposed allows the grasp policy to act as both a run-time state estimator and a grasp planner that uses the run-time feedback.

2 Relationship to Other Work

This work gathers insight from developmental psychology, control theory, reinforcement learning, robotics and artificial intelligence to look for mechanisms with which to program robot hands automatically by direct haptic interaction with manipulation tasks. Several research threads are immediately relevant to the proposed project. The first is the growing interest among behavioral scientists and roboticists regarding representations couched in patterns of interaction. The second is a body of results in the robot grasping community related to the mechanics of grasping, and finally, we will review some of the relevant literature regarding the application of reinforcement learning techniques to robot control.

2.1 Behavioral Models of Interaction

Over the past several years, developmental psychologists and roboticists are proposing similar theories of motor control — namely, that latent aptitudes expressed by virtue of the kinematic, dynamic, and "neurological" properties of a robot are exploited to simplify and structure motor programs. Furthermore, learning in these systems in achieved in the context of an on-going interaction with the world. An *interaction-based* representation

grounds knowledge in *activity*. From this perspective, "... motor timing in skilled actions is discovered ... through perceptual exploration of the body's intrinsic (or autonomous) dynamics within a changing task and physical space [55]." It is the potential for rich and varied interaction with the world that, we contend, leads to cognitive organization and development in humans — in our view, an important issue that has not received due attention.

During the first several months in an infant's life, reflexive responses are organized into coherent motor strategies, sensory modalities are coordinated and attentional mechanisms begin to emerge. **Native reflexive responses** like the primary walking reflex and the palmar grasp reflex [2] provide primitive behavior that accomplish sensor-driven work in the world. Bruner [6] refers to these types of behaviors as "preadaptation" primitives for learning skillful motor policies. Subsequently, policies for coordinating multiple sensory and motor modalities appear as primary circular reactions [45] which are practiced until the infant finds it possible to prolong certain interactions with the world.

This perspective is consistent with recent trends in robotics, where researchers are developing artificial muscles and neural oscillator models [25, 54, 58] to exploit the intrinsic dynamics of the control process. For example, series elastic actuators[47] are designed to provide an appropriate passive behavior to the limb. These approaches minimize muscular effort by looking for synergistic kinematic and dynamic coupling between the robot and the world. Moreover, these methods are generally robust to changing environments and other run-time perturbations.

In 1989, Koditschek *et al.* argued that motor planners should focus on finding controllers with inherent dynamical properties that produce valuable artifacts in the world rather than computing the artifacts directly [49, 33, 50, 34]. Such systems can be significantly more robust by virtue of error suppression in closed-loop actions. As result, one can avoid precise pre-modeling of the task and rely on robustness to accommodate the details of a particular run-time environment. The dynamics of the coupled system can be used to form a feature space for such systems as in the attractor landscape proposed by Huber *et al.* [28] or the limit cycles proposed by Schaal *et al.* [52]. Raibert's hopping platforms are representative of this class of approaches [48], in which control consists of a finite state supervisor that switches discrete controllers on and off as a function of observable events. This work presented in this paper is inspired by these previously published results.

2.2 Grasp Mechanics

A great deal of progress has been made in the analysis of phenomena associated with grasping and manipulation tasks [42]. There exist standard models of contact types, for example, including point contacts with and without friction, and soft-fingers that can produce torque around the contact normal [38, 14]. To control contact position during manipulation, techniques exist to use controllable slippage [16, 5, 12, 57, 31], and rolling contact geometries [41, 24, 7, 42]. Some of the most widely read literature on grasping concerns the conditions under which a grasp can restrain an object. Motivated by fixturing problems in machining tasks, a *form closure* grasp typically considers the placement of frictionless point contacts so as to fully restrain an object [37]. *Force closure* properties speak to the ability of a grasp to reject disturbance forces and usually considers frictional forces [17, 44, 15]. We have adopted insights from these results.

The stability of a grasp can be described in terms of the depth and steepness of the potential well determined by a grasp configuration. These properties define control forces that tend to restore the object to an equilibrium position when perturbed [56, 27]. It is noteworthy to mention that humans use many grasps in everyday life that are technically unstable by this account.

Our goal is to achieve flexible and adaptive control systems that acquire reusable control knowledge for application in "open" domains. The approaches cited above all rely on prior models of the object geometry to varying degrees of completeness. We are focused, as well, on where the models underlying grasp control come from. Despite the significant theoretical impact of this literature, we have not yet developed an adequate model of the sensory and motor *process* of grasping and manipulation. This process moves fluidly through multiple contact regimes and can trade stability margins early in a manipulation process for constructive interactions later, e.g. as in pick-and-place tasks [30]. Finally, we feel that the real challenge and opportunity afforded by multifingered hands is the automatic modeling of complex and non-stationary modes of interaction between a robot and the world. Modeling end-to-end manipulation sequences leads immediately to issues of representation and learning — issues that have been largely ignored by the grasping community to date.

2.3 Learning Control

Currently, there is a great deal of interest in adaptive control architectures for non-stationary, nonlinear processes [43, 22, 50, 48, 4]. These approaches postulate a family of local models that can be used to approximate the optimal, global control surface [40]. By switching controllers, or by reformulating local models, a linear

control substrate can be applied more generally to nonlinear and/or non-stationary tasks [48, 1]. Some of these approaches incorporate learning methods for control law synthesis [40, 1, 29, 39]. If local control models are stable (in the sense of Lyapunov), then they can actively confine the state to a neighborhood surrounding the attractor, thus approximately preserving some property in the system until conditions permit a transition to another attractor [50, 29]. We will employ this approach to express grasping behavior on multiple object types and will learn grasping policies that switch between closed-loop grasp controllers.

We propose to use dynamic programming-based Reinforcement Learning (RL) to learn policies that exploit control dynamics toward optimal grasping and manipulation policies. RL involves a stochastic search for delayed rewards and is widely acknowledged to be subject to the "curse of dimensionality." In high-dimensional, continuous domains, it may be quite improbable that a solution is found by stochastic search alone. Our goal is not to advance learning theory *per se*, but instead to provide representation and structure to learning methods that depend on exploration. A number of control systems employing reinforcement learning techniques have been designed to acquire policies off-line using simulated experience [3, 13], or on-line using robot platforms [23]. To address on-line learning in systems of moderate-to-high complexity, behavior-based techniques have been used to manage the size of the search space. Other approaches advocate learning pre-requisite skills that solve pre-defined subproblems and then combine them in a subsumption or voting framework [36, 26], or conversely, use previously designed behaviors as primitives within the learning framework [35]. The approach presented here falls into this general category.

3 Multifingered Grasp Synthesis

Grasp synthesis requires the determination of the grasp configuration parameters (contact normals and relative contact positions with respect to a task frame) that satisfy requirements determined by the task, sometimes indirectly. While most researchers describe grasp synthesis as an optimization problem, we proposed that it is best characterized as a robust control problem. In this framework, the robot uses tactile feedback to compute incremental contact displacements, performed in order to optimize the grasp metric.

We begin by defining a heuristic control basis. Consider an artificial potential derived as the squared resultant wrench imposed by unit magnitude frictionless point contacts on the surface of the object. The metric is observed by establishing a contact configuration and measuring contact normals with a tactile sensor. In experimental implementations, this required a finite state supervisor to manage the contact loads to be of sufficient magnitude to produce useful position and normal estimates, but small enough to avoid saturating the Brock tactile sensors.

This supervisor limits contact loads to between a minimum of 0.05 N and a maximum of 0.1 N, which provides a good quality contact measurement and makes it possible to minimize disturbances to the object pose during grasp formation [21]. In the worst case for three fingers, this yields a total force of 0.3 N. In practice, we were able to probe an object with mass as little as approximately 50 g (1.6 oz) without significant object displacements.

Each tactile observation is used to estimate the contact wrench derived from a frictionless point contact with the observed position and normal. Given the wrench residual vector for n contacts

$$\rho = \sum_{i=1}^{n} \left[f_x^i \;\; f_y^i \;\; f_z^i \;\; \tau_x^i \;\; \tau_y^i \;\; \tau_z^i \right]^T,$$

then the squared wrench residual is defined by

$$\epsilon = \rho^T \rho. \tag{1}$$

Control gradients on this metric produce differential displacements of the contact configuration by assuming either planar or constant curvature surface types [8, 9, 10]. Fixed points in these heuristic potential functions minimize the force and moment residuals generated by normal contact forces and, thus, provide a necessary but not sufficient condition for *force closure* grasps. In previous work, we defined a sequence of these controllers that produces optimal (minimum friction coefficient) grasps for regular convex objects [11]. For planar objects in this class, this approach produces solutions in the set of those obtained using Nguyen's geometric algorithm for force closure grasps in the plane [44], however, it also can produce non-force closure grasps and requires further management to avoid inadequate solutions.

The controller π_c computes the residual associated with contact set, c, and modifies the configuration of set c by descending the residual gradient until a local minimum for ϵ is reached. The minimization of ϵ is closely related to the grasp performance since it implies the existence of a null space of non-zero rank in the grip Jacobian [51] and therefore provides a necessary condition for force closure [11]. The subset of contacts c specifies which fingers and surfaces are enlisted in the grasp task. The Stanford/JPL hand has 3 fingers labeled $\{T, 1, 2\}$ that leads to 4 distinct contact subsets, assuming that two or more fingers are required to grasp the object:

$$\mathcal{C} = \{(T,1),\;\; (T,2),\;\; (1,2),\;\; (T,1,2)\}. \tag{2}$$

Each instance of $c \in \mathcal{C}$ defines a new control law, affecting the outcome of the grasp process directly. The controller π_c is better characterized as an element in a family of grasp controllers, referred to as $\Pi = \{\pi_c |\; c \in \mathcal{C}\}$; the control basis Π represents the robot's native reflexes in response to tactile stimuli.

The stability of controller π_c ensures that the system will achieve a configuration such that $|\dot{\epsilon}| < \delta$, where

$\delta > 0$ is a threshold for equilibrium. This criterion determines the termination of π_c, segmenting the interaction between the robot and the environment into trials.

The control actions of π_c depend on local tactile feedback exclusively. The convergent configurations for π_c correspond to local minima of ϵ_c. From a given contact configuration, each choice of control law $\pi_c \in \Pi$ will lead to a distinct candidate grasp configuration. One can imagine a large set of such equilibrium configurations resulting from all the fixed points reachable using $\pi_c : c \in \mathcal{C}$ from all initial configurations. In the following, we will introduce an interaction-based state estimator on which control decisions can be based to avoid non-force closure equilibria and to produce adaptive optimal grasp policies when embedded in manipulation tasks.

4 Constructing Models of Control Dynamics

The use of parametric models to represent a sequence of observations is common practice in the dynamical systems literature (e.g. see Fraser [18]), especially when insight about their structure is available. We define an observation, $\mathbf{o} = [\epsilon \;\; \dot{\epsilon}]^T$ where ϵ and $\dot{\epsilon}$ are the squared residual and its time rate of change, respectively. We assume that the observation will evolve along a piecewise continuous contour in the "residual" phase portrait to an equilibrium configuration, $[\epsilon_0 \;\; 0]^T$. A particular parametric model, M_{π_i}, describing controller π_i, predicts observation $\tilde{\mathbf{o}}$ given the observed residual squared error ϵ. Observations are noisy so that membership of observation $\mathbf{o}$ in M_{π_i} is estimated probabilistically.

The complete representation of system dynamics under policy π_i requires a set of m observation models, expressed as $\mathcal{M}(\pi_i) = \bigcup_{k=1}^{m} M_{\pi_i}(\theta_k, \mathbf{o})$. The set $\mathcal{M}(\pi_i)$ expresses empirical knowledge acquired by the agent during the execution of policy π_i. The derivation of $\mathcal{M}(\pi_i)$ involves sampling system dynamics for a predetermined number of epochs τ, while recording the data $\mathcal{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n\}$ observed as the control error evolves toward equilibrium. Any single object will present several distinct models and models are not exclusive to single objects. Once $\mathcal{M}(\pi_i)$ is available, Bayesian estimation is used to identify the subset $\mathbf{q} \subset \mathcal{M}(\pi_i)$ of models compatible with a sequence of run-time observations. The **state** of the system is defined as the concatenation of the control law and the membership pattern, $(\pi_i, \mathbf{q})$.

Figure 1 depicts a set of hypothetical models corresponding to policy π_i. If each model is given a discrete label (A, B, C, D, E), one can describe the transitions between subsets of models in terms of a discrete graph, shown in the right panel of Figure 1. The regions in phase space where two or more models overlap are identified with the labels corresponding to the overlapping models; (B, C) is one example. The resulting representation defines

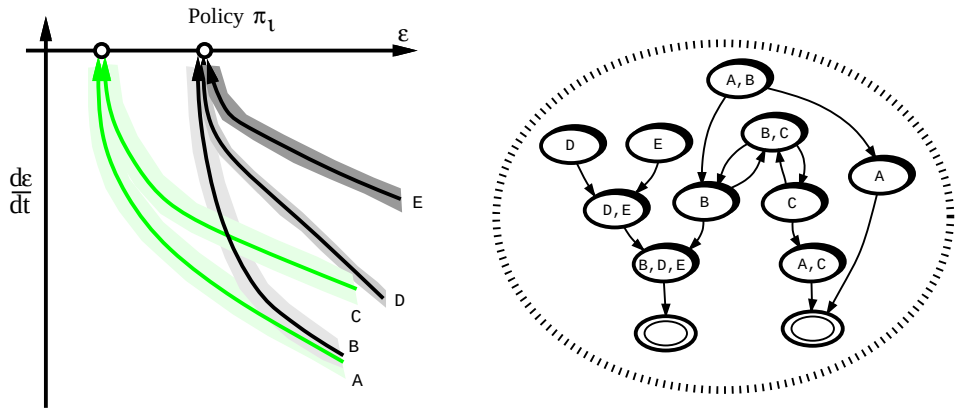

Figure 1: Diagrams depict the phase portrait $[\epsilon \ \dot{\epsilon}]$ for policy π_i (left) and all possible context transitions (right).

a discrete state space that describes the evolution of information in this grasping process.

The transformation from a set of continuous models to a graph of discrete states is to be carried out for each control policy $\pi_i \in \Pi$. The discrete state space allows the system to experiment with sequences of control policies using the reinforcement learning framework. Figure 2 illustrates how policy switching may lead to improved performance. Switching policies, as in this example, forms finger gaits toward optimal contact configurations.

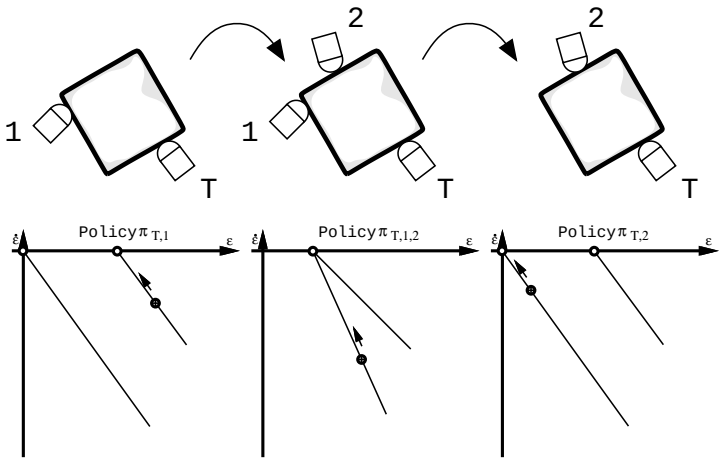

Figure 2: A hypothetical context-dependent grasp of a cube (top view). Policy sequence $\pi_{T,1}$, $\pi_{T,1,2}$, $\pi_{T,2}$ might accomplish the best two-fingered grasp configuration by using an intermediate three-fingered control context.

5 Context-Dependent Grasp Policies

The problem consists of identifying a sequence of grasp controllers that acquire adequate state information and that achieve optimal quality grasps. Optimality is expressed as the grasp configuration that minimizes the coefficient of friction between fingers and object surfaces required for stability. The object's identity, geometry, and pose with respect to the hand are unknown.

Determining an optimal switching policy is a sequential decision problem which can be solved within the POMDP framework, provided one can describe the control sequence in the form of transitions between discrete states in a Markov chain. The next subsection describes a representation in which state information is derived from a mixture of generative models describing the grasp dynamics in the phase space defined by $(\epsilon, \dot{\epsilon})$. It is assumed that the resulting representation is approximately Markovian; this assumption allows us to use the Q-learning algorithm to approximate the optimal switching policy. A description of the experimental setup and parameters follows.

5.1 Experimental Setup

In [21], we constructed models and policies that proved the feasibility of the proposed approach on our Stanford/JPL equipped with Brock tactile sensors. Figure 3 depicts the setup used in these pilot studies. In addition

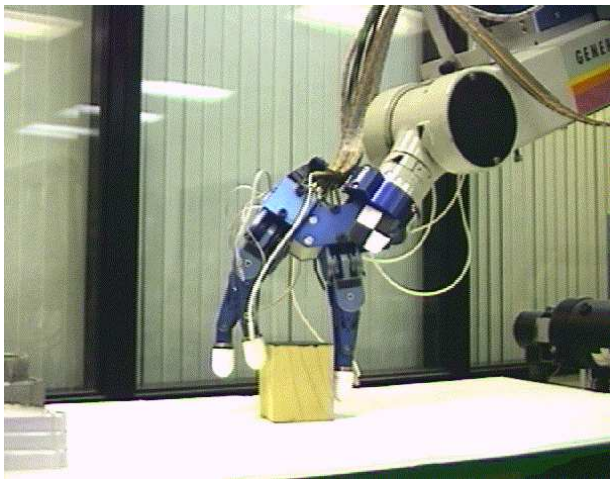

Figure 3: The setup for the grasp task: the hand is positioned so that all fingers can reach the object.

to acquiring empirical models of the grasp dynamics for 3 object types and using these models to generate state

estimates and robust grasping policies, we were able to show that given several grasps on an object, we can often recognize the object type by virtue of the models visited while grasping.

To execute a more thorough evaluation of the statistical behavior of the grasp policies, we used a kinematic simulation of the Stanford/JPL hand to run significantly more trials than was practical with the robot. In each trial, the object was selected from three possible object types: cylinders, cubes, and triangular prisms. The geometric parameters of the objects were drawn from a probability distribution; as a result, no two grasp trials were performed on identical objects.

5.2 Sequential Decision Problem

Initially, the hand is positioned next to the object in a configuration such that a portion of the object surface is within the workspace of each finger. Tactile feedback is simulated using the geometrical object description with superimposed noise processes consistent with that observed in our Brock sensors.

5.2.1 State Representation

The state representation proposed is based on two discrete sets: the control basis $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, and the membership pattern vector $\mathbf{q} = [p_1 \quad p_2 \quad \ldots \quad p_m]^T$, that indicates which m generative models are consistent ($p_k = 1$) with the last observation and which are not ($p_k = 0$).

Given the observation vector $\mathbf{o}$ collected as the agent executes the primitive π_i, the corresponding system state is denoted by the tuple $(\pi_i, \mathbf{q})$. The vector $\mathbf{q}$ expresses the models (or operational contexts) that are compatible with the observation $\mathbf{o} = (\epsilon, \dot{\epsilon})$. It conveys more information than individual observations, as it correlates $\mathbf{o}$ with the information derived from the agent's past experiences. The resulting state representation is intrinsically situated, because it is grounded in autonomous interaction with the environment.

A generative model describes how observation sequences can be "generated" within a particular operational context. As the agent executes the controller π_i, it can discover a variety of operating contexts characterized by distinct dynamics. If an exhaustive set of generative models is available, the robot may infer which models are compatible with a given sequence of observations by computing and updating the likelihood that each model has generated the data observed. If $P(M_k)$ is the likelihood that model k generates the sequence of observations, and $P(\mathbf{o}|M_k)$ the likelihood of observing $\mathbf{o}$ given that it is drawn from the distribution expressed by model k,

then $P(M_k|\mathbf{o})$ (the likelihood that model k explains the data observed given observation $\mathbf{o}$) can be computed as

$$P(M_k|\mathbf{o}) = \frac{P(\mathbf{o}|M_k)P(M_k)}{\sum_{i=1}^{m} P(\mathbf{o}|M_i)P(M_i)},$$

assuming the existence of m generative models.

In this work, the probability $P(\mathbf{o}|M_k)$ is computed by the generative models, and the corresponding conditional probability distributions are represented by parametric models. The cost of acquiring data in real robots favors the choice of parametric models, whose derivation typically require less data than for non-parametric models. The following parameterization was used to represent the generative models:

$$\begin{aligned} \bar{\mathbf{o}} &= [\epsilon \quad -K\epsilon + \epsilon_0]^T \\ M_{\pi_i}(\theta, \mathbf{o}) &= \frac{1}{L}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\mathbf{o}-\bar{\mathbf{o}})^T(\mathbf{o}-\bar{\mathbf{o}})}{2\sigma^2}\right). \end{aligned} \tag{3}$$

The observation vector $\mathbf{o} = [\epsilon \quad \dot{\epsilon}]^T$ consists of the squared wrench residual ϵ (defined by Equation 1) and its temporal rate of change $\dot{\epsilon}$. The parameter vector $\theta = [\sigma^2 \quad K \quad \epsilon_0]^T$ contains the parameters of the normal distribution, characterized by the variance σ^2, and the parameters K and ϵ_0 used to compute an estimate of $\dot{\epsilon}$. The constant L normalizes the distribution.

The representation of the dynamics elicited by policy π_i requires a set of m generative models, expressed as $\mathcal{M}(\pi_i) = \bigcup_{k=1}^{m} M_{\pi_i}(\theta_k, \mathbf{o})$. The parameter vector θ_k completely specifies the distribution $M_{\pi_i}(\theta_k, \mathbf{o}) = P(\mathbf{o}|M_k)$.

Each model $M_{\pi_i}(\theta_k, \mathbf{o}) \in \mathcal{M}(\pi_i)$ is constructed using data obtained empirically as the agent executes the controller π_i. It records a sequence of n observations $\mathcal{O}= \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n\}$ gathered until convergence of π_i. The derivation of model $M_{\pi_i}(\theta_k, \mathbf{o})$ consists in determining the parameter vector θ_k^* that maximizes the likelihood that $M_{\pi_i}(\theta_k, \mathbf{o})$ "generates" $\mathcal{O}$.

The log-likelihood associated with sequence $\mathcal{O}$ is defined as $L(\mathcal{O}, \theta_k) = \sum_{j=1}^{n} \log M_{\pi_i}(\theta_k, \mathbf{o}_j)$, where n is the length of the sequence of observations $\mathcal{O}$, and $\mathbf{o}_j$ denotes the j^{th} element of $\mathcal{O}$. Many optimization procedures can be used to derive the parameter vector θ_k. In the special case in which $M_{\pi_i}(\theta_k, \mathbf{o})$ is a Gaussian distribution with an average that is a linear function of the parameter vector θ_k (as in Equation 3), there exists a closed-form solution for θ_k^*.

In this paper, the models were constructed based on data collected over the course of 35 grasp trials for each family of objects. Considering the four possible choices for π_i (see Equation 2), a total of $4 * 3 * 35 = 420$ models resulted; after the elimination of the redundant models, 61 models were identified.

Notice that the cardinality of the state space is given by the product $|\mathcal{C}| \times |\mathcal{Q}|$, where $\mathcal{Q}$ is the set of all distinct membership pattern vectors $\mathbf{q}$. It can be shown under reasonable assumptions that $|\mathcal{Q}|$ is proportional to the square of the number of models; in this case, the estimate corresponds to $\approx 61^2 = 3721$ states.

5.2.2 Reward Structure

Grasps are evaluated by computing the minimum friction coefficient required to produce a zero net force and moment through the contact configuration. The computation of μ_0 requires the resolution of a mathematical programming problem subject to the following constraints on normal and tangential forces [32, 20]:

1. $|\mathbf{f}_n| \geq 1$ — normal forces are compressive with magnitude greater than or equal to 1,
2. $|\mathbf{f}_t| \leq \mu |\mathbf{f}_n|$ — the net contact force must remain within the friction cone, and
3. $\sum_{i=1}^{k} \mathbf{f}_i = \sum_{i=1}^{k} (\mathbf{r}_i \times \mathbf{f}_i) = \mathbf{0}$.

In general, infinitely many force distributions satisfy these constraints. The solution selected corresponds to the one with the minimum combined magnitude $M = \sum_{i=1}^{k} |\mathbf{f}_{t_i}| + |\mathbf{f}_{n_i}|$. The solution for μ_0 is computed from the resulting force distribution. A more efficient algorithm for computing μ_0 is available [53]; the approach reported here was chosen for its simplicity.

A convergent grasp configuration receives a score of $1 - \mu_0$, where μ_0 is the minimum friction coefficient required for a *stabilizable* grasp; a score of 1 ($\mu_0 = 0$) is best; no other costs or rewards are present. It is important to note that there is nothing special about the choice of μ_0 - virtually any quality metric could be used to provide a partial order to the fixed points in the controllers π_c.

5.2.3 Transition Semantics

The system evaluates the utility of pursuing a different control law whenever a change is detected in the membership pattern $\mathbf{q}$. Changes in $\mathbf{q}$ are special events, as they signal the fact that extra information has been acquired by the system.

After the convergence of the current control law, the system has the choice of terminating the trial or invoking a different controller. Including choice points after convergence of the current control law is necessary for two reasons: (1) it allows the system to use suboptimal convergent configurations as way points to the optimal

configurations, and (2) the system's initial configuration may be at, or very close to, a suboptimal grasp configuration. The choice point allows the system to recover from these undesirable (albeit unlikely) situations.

6 Results

A total of 1600 simulated trials were performed. In each trial, a random object with random parameters was generated. The first action, $(\pi_c)_0$, is chosen at random, determining the contacts (fingers) first employed to probe the object. Up to 50 probes are used in each grasp trial. After 50 probes, the trial is interrupted and the current grasp configuration scored as if it were a convergent configuration. Exploration was regulated by a Boltzmann distribution.

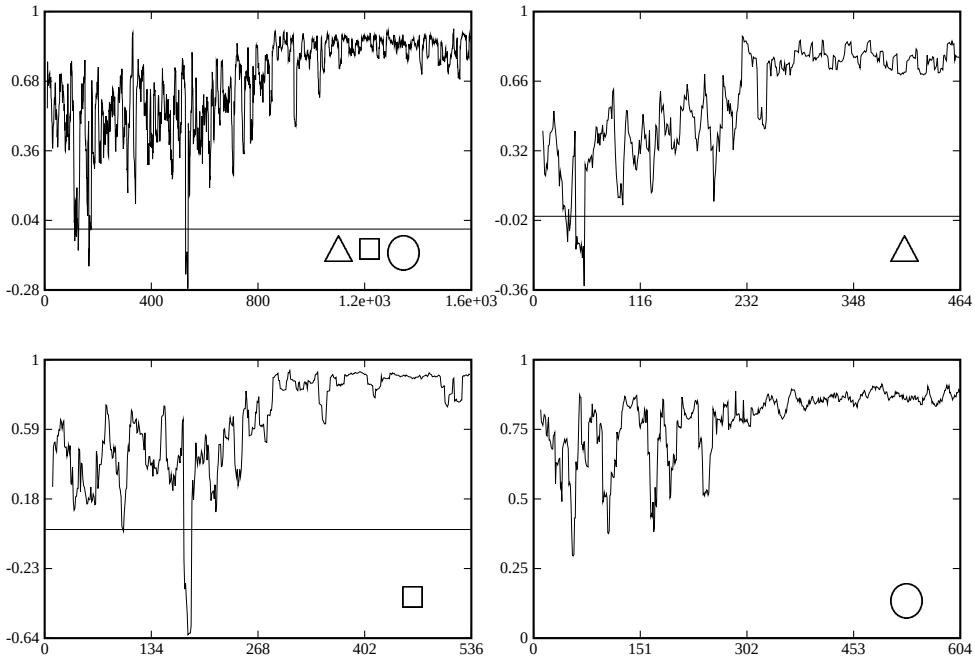

Figure 4: △□◯ Typical learning curve for all objects. Other curves (labeled △, □, ◯) are the corresponding learning curves for the individual objects. Vertical axes represent grasp scores $(1 - \mu_0)$, and horizontal axes are the trial number.

Figure 4 depicts a typical learning curve (curve labeled △□◯, top left curve). The curve was smoothed using a sliding window 10 data points wide. Each point corresponds to the grasp score associated with the terminal grasp configuration. The data corresponding to each object can be separated into three classes, corresponding to the three families of objects. The resulting learning curves are labeled △, □, ◯. Because the objects are not perfectly symmetrical and the control execution is halted before $\dot{\epsilon} = 0$, it is unrealistic to expect an average

score of 1. The curves for the individual objects are close to the optimal, within the limitations of the Q-learning algorithm.

Figure 5 presents the performance histograms for the controllers in the control basis ("native" controllers), and for the composite controller after learning. Each histogram depicts data collected over 1600 trials, involving all objects. For the top histogram, a controller was selected from the control basis at random, and executed until convergence. The end grasp configuration was scored and included in the histogram.

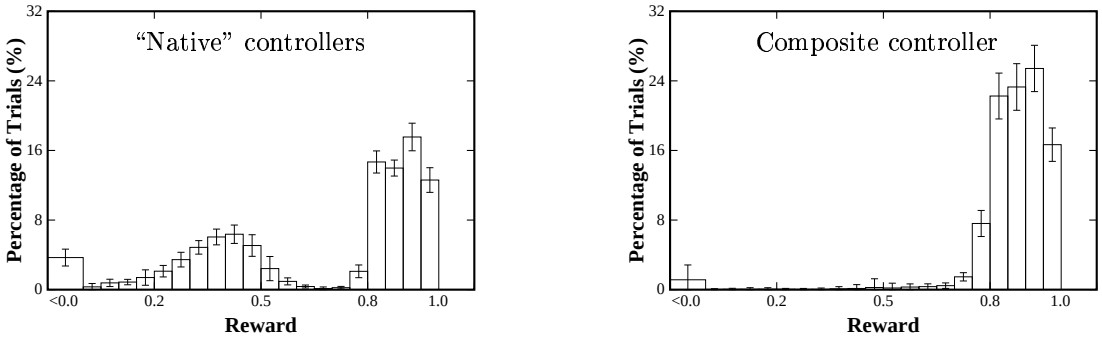

Figure 5: Grasp score distributions for the control basis controllers ("native controllers") and the composite controller after learning; higher scores are better. Histograms are based on data collected over 100 trials.

As one can see, the acquired grasp controller avoids the majority of low quality solutions: 88% of the solutions have scores higher than 0.8, compared to 59% for the native controllers. The variance associated with solution quality is also substantially smaller.

7 Conclusion and Future Work

All successful organisms exploit some form of native structure (neurological, muscular, skeletal) to settle into a stable dynamic relationship with their environment by exploiting the intrinsic dynamics of bodies and tasks. Humans learn, in addition, to exploit favorable dynamic relationships with the world by using acquired control knowledge.

In this work, we have described how context-dependent grasp strategies can be constructed through the activation of the native control primitive most highly recommended given the perceived operational context. The definition of context or state is itself based on empirically derived dynamic models. The performance gains achieved are the result of a better match between the task requirements and the capabilities of the system.

The method proposed was successfully applied to the multifingered grasp platform in Figure 3, in which a robotic

hand must deploy its grasp resources (fingertip contacts) according to the haptic context given unknown object identity and orientation. The number of trials required for real versus simulated robots is an issue, due to the cost of the probing operation. Biological systems, likewise, take a relatively long time to acquire these skills. In future work, we intend to learn similar policies without resorting to simulation to accelerate policy formation. We expect that the number of trials should be roughly equivalent to the simulation trials cited (1600), however, we have generated results similar to Figure 5 with as few as 100 trials in simulation. We anticipate that learning trials run exclusively on our robot platform should incorporate non-ideal aspects of the real hand as well.

Finally, in solving this task, the system develops its own notion of "object identity," and uses it effectively. While we have conducted experiments in which this approach identifies the three classes of objects we used, it typically requires many grasps on a single object. This is due to our design; object identification is not an explicit goal of the system. In the future, the resulting haptic categories will be used in conjunction with visual categorization to form multi-modal object representations that should be capable of efficient recognition behavior.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CISE/CDA-9703217, IRI-9704530 and IRI-9503687.

References

[1] Elizeth G. Araujo and Roderic A. Grupen. Learning control composition in a complex environment. In Pattie Maes, Maja Mataric, Jean-Arcady Meyer, Jordan Pollack, and Stewart W. Wilson, editors, *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 333–342, Cambridge, Massachusetts, September 1996. MIT Press.

[2] A.E. Aronson et. al. *Clinical Examinations in Neurology.* W.B. Saunders Co., Philadelphia, PA, 1981.

[3] Andrew G. Barto, Richard S. Sutton, and C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cyber.*, 13(5):834–846, 1983.

[4] R.A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, March 1986.

[5] Randy C. Brost. Automatic grasp planning in the presence of uncertainty. In *Proceedings 1986 IEEE Conference on Robotics and Automation*, volume 3, pages 1575–1581, 1986.

[6] J.S. Bruner. Organization of early skilled action. *Child Development*, 44:1–11, 1973.

[7] D. Chang and M. Cutkosky. Rolling with deformable fingertips. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, pages 194–199, 1995.

[8] J. Coelho and R. Grupen. Constructing effective multifingered grasp controllers. In *Proceedings of the 1994 Conference on Robotics and Automation*, San Diego, CA, May 1994. IEEE.

[9] J. Coelho and R. Grupen. Optimal multifingered grasp synthesis. In *Proceedings of the 1994 Conference on Robotics and Automation*, San Diego, CA, May 1994. IEEE.

[10] J. Coelho and R. Grupen. Online grasp synthesis. In *Proceedings of the Conference on Robotics and Automation*, Minneapolis, MN, April 1996. IEEE.

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

[12] A. Cole, P. Hsu, and S. Sastry. Dynamic regrasping by coordinated control of sliding for a multifingered hand. In *Proceedings of the 1989 Conference on Robotics and Automation*, pages 781–786, Scottsdale, AZ, May 1989. IEEE.

[13] R. H. Crites and Andrew G. Barto. Improving elevator performance using reinforcement learning. In *NIPS8*. Morgan Kaufmann, 1995.

[14] M.R. Cutkosky and P.K. Wright. Friction, stability and the design of robotic fingers. *Journal of Robotics Research*, 5(4), Winter 1986.

[15] B. Faverjon and J. Ponce. On computing two-finger force-closure graps of curved 2d objects. In *Proceedings of the IEEE Conference on Robotics and Automation*, volume 1, pages 424–429, April 1991.

[16] R.S. Fearing. Simplified grasping and manipulation with dextrous robot hands. *IEEE Journal of Robotics Research*, 2(4):188–195, January 1983.

[17] C. Ferrari and J. Canny. Planning optimal grasps. In *Proc. 1992 IEEE Int. Conf. Robotics Automat.*, volume 3, pages 2290–2295, Nice, FRANCE, May 1992.

[18] A. M. Fraser and A. Dimitriadis. Forecasting probability densities by using hidden markov models with mixed states. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 265–281. Santa Fe Institute/Addison-Wesley, 1993.

[19] Luiz Marcos Garcia Gonçalves, Gilson A. Giraldi, A. F. Oliveira Antonio, and Roderic A. Grupen. Learning policies for attentional control. In *Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA '99)*. IEEE, 1999.

[20] R.A. Grupen, K. Biggers, T.C. Henderson, and S. Meek. Task defined internal grasp wrenches. Technical Report UUCS-88-001, Department of Computer Science, University of Utah, 1988.

[21] R.A. Grupen, J. Coelho, and K. Souccar. On-line grasp estimator: A partitioned state space approach. Technical Report CS Technical Report 92-75, CS Department, University of Massachusetts, October 1992.

[22] Roderic A. Grupen, Manfred Huber, Jefferson A. Coelho, Jr., and Kamal Souccar. A basis for distributed control of manipulation tasks. *IEEE Expert*, 10(2):9–14, 1995.

[23] V. Gullapalli, R. Grupen, and A. Barto. Learning reactive admittance control. In *Proceedings of the 1992 Conference on Robotics and Automation*, pages 1475–1480, Nice, FRANCE, May 1992. IEEE.

[24] L. Han and J.C. Trinkle. Dextrous manipulation by rolling and finger gaiting. In *Proceedings of the 1998 Conference on Robotics and Automation*, pages 730–735, 1998.

[25] Kjeldy A. Haugsjaa, Kamal Souccar, Christopher I. Connolly, and Roderic A. Grupen. A computational model for repetitive motion. In C. E. Collyer and D. A. Rosenbaum, editors, *Timing of Behavior: Neural, Computational, and Psychological Perspectives*. MIT Press, 1996.

[26] J. Hoff and G. Bekey. An architecture for behavior coordination learning. In *Proceedings of the 1995 IEEE International Conference on Neural Networks*, pages 2375–2380, Perth, Australia, November 1996. IEEE.

[27] W. Howard and V. Kumar. On the stability of grasped objects. *IEEE Transactions of Robotics and Automation*, 12(6):904–917, 1996.

[28] M. Huber and R. Grupen. Learning to coordinate controllers - reinforcement learning on a control basis. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI)*, Nagoya, JP, August 1997. IJCAII.

[29] Manfred Huber and Roderic A. Grupen. A feedback control structure for on-line learning tasks. *Robots and Autonomous Systems*, 22(3/4):303–315, 1997.

[30] J.L. Jones and T. Lozano-Pérez. Planning two-fingered grasps for pick-and-place operations on polyhedra. In *Proceedings of 1990 Conference on Robotics and Automation*, pages 683–688. IEEE, May 1990.

[31] I. Kao and M. R. Cutkosky. Quasistatic manipulation with compliance and sliding. *International Journal of Robotics Research*, 11(1):20–40, 1992.

[32] J. Kerr and B. Roth. Analysis of multifingered hands. *Journal of Robotics Research*, 4(4):3–17, Winter 1986.

[33] Daniel E. Koditschek. The application of total energy as a lyapunov function for mechanical control systems. In *Dynamics and Control of Multibody Systems*, volume 97 of *Contemporary Mathematics*, pages 131–157. American Mathematical Society, 1989.

[34] Daniel E. Koditschek. The control of natural motion in mechanical systems. *Journal of Dynamic Systems, Measurement, and Control*, 113:547–551, 1991.

[35] P. Maes and R. Brooks. Learning to coordinate behaviors. In *Proceedings of the 1990 AAAI Conference on Artificial Intelligence*. AAAI, 1990.

[36] S. Mahadevan and J. Connell. Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55:311–365, 1992.

[37] X. Markenscoff, L. Ni, and C Papadimitriou. The geometry of grasping. *International Journal of Robotics Research*, 9(1):61–74, 1990.

[38] M.T. Mason and J.K. Salisbury. *Robot Hands and the Mechanics of Manipulation.* The MIT Press, Cambridge, MA, 1985.

[39] M. Mataric. Reinforcement learning in the multi-robot domain. *Autonomous Robots*, 4:73–83, 1997.

[40] D. Michie and R. Chambers. BOXES: An experiment in adaptive control. In E. Dale and D. Michie, editors, *Machine Intelligence 2*. Edinburgh, 1968.

[41] D. J. Montana. The kinematics of contact and grasp. *International Journal of Robotics Research*, 7(3):17–32, 1988.

[42] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation.* CRC Press, Boca Raton, CA, 1994.

[43] Roderick Murray-Smith and Tor Arne Johansen (Eds.). *Multiple Model Approaches to Nonlinear Modelling and Control.* Taylor and Francis, London, 1997.

[44] V.D. Nguyen. The synthesis of stable grasps in the plane. In *Proceedings of the 1986 Conference on Robotics and Automation*, volume 2, pages 884–889, San Francisco, CA, April 1986. IEEE.

[45] J. Piaget. *The Origins of Intelligence in Childhood.* International Universities Press, 1952.

[46] J.H. Piater and R.A. Grupen. Toward learning visual discrimination strategies. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, 1999. IEEE.

[47] J. Pratt and G. Pratt. Exploiting natural dynamics in the control of a planar bipedal walking robot. In *Proceedings of the* 36^{th} *Annual Allerton Conference on Communication, Control, and Computing*, 1998.

[48] M.H. Raibert. *Legged Robots that Balance.* MIT Press, Cambridge, MA, 1986.

[49] E. Rimon and D. Koditschek. The construction of analytic diffeomorphisms for exact robot navigation on star worlds. In *Proceedings of the 1989 Conference on Robotics and Automation*, volume 1, pages 21–26, Scottsdale, AZ, May 1989. IEEE.

[50] A.A. Rizzi, L.L. Whitcomb, and D.E. Koditschek. Distributed real-time control of a spatial robot juggler. *IEEE Computer Magazine*, 25(5), May 1992.

[51] J.K. Salisbury. *Kinematic and Force Analysis of Articulated Hands.* PhD thesis, Stanford University, May 1982.

[52] Stefan Schaal and D. Sternad. Programmable pattern generators. In *International Conference on Computational Intelligence in Neuroscience (ICCIN'98)*, pages 48–51, 1998.

[53] J.T. Schwartz, M. Sharir, and J. Hopcroft, editors. *Planning, Geometry, and Complexity of Robot Motion.* Ablex Publishing Corporation, Norwood, NJ, 1986.

[54] G. Taga. A model of the neuro-musculo-skeletal system for anticipatory adjustment of human locomotion during obstacle avoidance. *Biological Cybernetics*, 78(2):9–17, 1998.

[55] E. Thelen and L. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action.* MIT Press, Cambridge, MA, 1994.

[56] J. C. Trinkle. On the stability and instantaneous velocity of grasped frictionless objects. *IEEE Transactions of Robotics and Automation*, 8(5), 1992.

[57] J. C. Trinkle and R. P. Paul. Planning for dexterous manipulation with sliding contacts. *International Journal of Robotics Research*, 9(3):24–48, 1990.

[58] M.M. Williamson. Neural control of rhythmic arm movements. *Neural Networks*, 11(7-8):1379–1394, 1999.