Learning Prospective Pick and Place Behavior

David S. Wheeler, Andrew H. Fagg, Roderic A. Grupen
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts 01003
{dwheeler, fagg, grupen}@cs.umass.edu

Abstract— **When interacting with an object, the possible choices of grasp and manipulation operations are often limited by pick and place constraints. Traditional planning methods are analytical in nature and require geometric models of parts, fixtures, and motions to identify and avoid the constraints. These methods can easily become computationally expensive and are often brittle under model or sensory uncertainty. In contrast, infants do not construct complete models of the objects that they manipulate, but instead appear to incrementally construct models based on interaction with the objects themselves. We propose that robotic pick and place operations can be formulated as *prospective behavior* and that an intelligent agent can use interaction with the environment to learn strategies which accommodate the constraints based on expected future success. We present experiments demonstrating this technique, and compare the strategies utilized by the agent to the behaviors observed in young children when presented with a similar task.**

I. Introduction

The problem of grasping an object and moving it to another location has long been studied in robotics. One approach is to explicitly compute "pick-and-place" constraints and perform a search within the constrained space [1], [2]. In this work, it is acknowledged that constraints imposed late in a multi-step control process can influence decisions made early in that process. The classical example is the selection of an initial grasp of a peg that is compatible with a subsequent insertion of that peg into a hole. If the grasp involves surfaces of the peg that must fit into or mate with corresponding surfaces in the hole, a regrasp must be employed to free those surfaces. The approach cited above advocates a backward chaining algorithm that propagates the assembly process backward in time until it "finds" the initial state. The resulting grasps are all compatible with the desired outcome, so reversing the solution will avoid erroneous early control decisions. Not only is such an approach computationally expensive, but it requires that complete models of the task exist prior to acting, and does not elucidate the perceptual distinctions necessary to solve the next instance of the problem.

In contrast, humans are capable of robustly planning and executing grasps to objects about which their knowledge is incomplete. Furthermore, it appears that grasping strategies are acquired incrementally as a function of experience with different objects. For example, McCarty *et al.* studied the initial reach made by infants to a spoon laden with applesauce [3]. The youngest infants (9 months) demonstrated an almost "reflexive" strategy in which they grasped the spoon with their dominant hand and immediately brought their hand to their mouth. This strategy is successful when the spoon is presented in an orientation that results with the bowl of the spoon on the thumb side of the hand, but fails when the spoon is presented in the opposite orientation. In the latter case, the infants corrected their movement by either regrasping the spoon or rotating their hand into an awkward configuration. With age, the policy evolves to an anticipatory regrasping strategy, which is later subsumed by a process that predicts which arm to use so that regrasping is not necessary.

We hypothesize that human infants use exploration based learning to search for actions that will yield future reward, and that this process works in concert with the identification of features which discriminate between important interaction contexts. In this context, this paper proposes a control structure for acquiring increasingly sophisticated representations and control knowledge incrementally. Within this framework, we suggest that a robot can use Reinforcement Learning (RL) [4] to write its own programs for grasping and manipulation tasks that depend on models of manual interaction at many temporal scales. The robot learns to associate visual and haptic features with grasp goals through interactions with the task. Our approach is a computational account of a theory of the development of prospective behavior in human infants.

II. Grip Selection Learning in Infants

McCarty, *et al.* [3] have studied tool-use problem solving strategies used by 9, 14, and 19 month old children. In this experiment, the infant is presented with a spoonful of applesauce in one of two orientations (with respect to the infant, the bowl of the spoon is placed either to the left or to the right of the handle). Three[1] different grips were exhibited by the children:

- Radial - An overhand grip on the handle of the spoon

[1] An underhanded grip was also available to the children, but was only used by one child on one trial.

with the thumb side of the hand closest to the bowl of the spoon.

- Ulnar - An overhand grip on the handle of the spoon with the thumb away from bowl.
- Goal-end - An overhand grip on the bowl of the spoon.

In this task, the radial grip is the most effective grip to obtain the applesauce; nonradial grips lead to physically awkward postures or require a regrasp action.

In the child study, trials were categorized as *easy* or *difficult* based on the initial orientation of the spoon with respect to the child's dominant hand. On easy trials, the spoon was presented with the handle on the same side as the child's dominant hand; if the child used the dominant hand, a reach to the handle resulted in an efficient radial grip. On difficult trials, the spoon was presented with the bowl on the child's dominant hand side. A reach with the dominant hand on difficult trials resulted in a nonradial grip; if the child did not take corrective action, the handle of the spoon was placed in the mouth.

Irrespective of age, the children predominantly used their dominant hand on easy trials (and hence achieved a radial grip). However, on difficult trials, the children exhibited a range of strategies that varied with age. The three dominant strategies are illustrated in Figure 1 and are as follows:

- Late Correction Strategy - The child used an initial nonradial grip and placed the handle of the spoon in the mouth, then made a correction and obtained the applesauce.
- Early Correction Strategy - The child used an initial nonradial grip, then made a correction *without* first placing the handle in the mouth, and obtained the applesauce.
- Optimal Strategy - The child used an initial radial grip and obtained the applesauce without need for corrective action.

Figure 2 shows the percent of radial and nonradial grips used on difficult trials by children in the three age groups. The high incidence of nonradial grips used by the 9 and 14 month old children indicates that they were still following the preexisting tendency to use their dominant hand. At 19 months, radial grips were far more prevalent. The older children presumably realized that the orientation of the spoon on the table recommends the best strategy, and suppressed a dominant hand strategy in favor of reaching with the non-dominant hand on difficult trials.

McCarty *et al.* used the observed strategy on difficult trials to measure the the degree of advanced planning performed by the children. Late corrections indicate the child used preconceived actions with little or no consideration of the spoon orientation. Children who made early corrections began with preconceived actions, but recognized the error and made the appropriate correction. Children who used the optimal strategy exhibited the highest degree of planning. Results from the child study illustrated in Figure 3 show that the youngest children tended to make the same number of early and late corrections, while children in the middle age group exhibited a strong preference for early corrections. Children in the oldest age group adopted the optimal strategy, so corrective actions were not necessary.

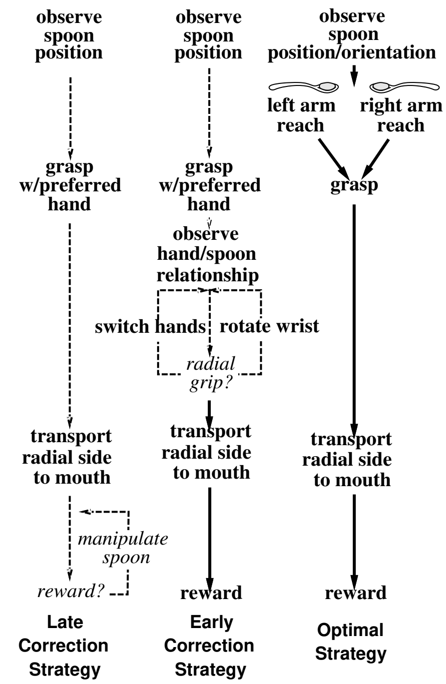


Fig. 1. Prospective Behavior revealed in the applesauce experiment. Dotted lines indicate exploration and solid lines indicate a developing policy. Infants initially (9 months) employ a dominant hand strategy to bring the spoon to the mouth (left). At 14 months, they learn to correct the strategy by performing regrasps before the spoon is inserted into the mouth (middle). By 19 months, toddlers grasp with the correct hand so that a re-grasp is not necessary (right).

III. A Robot Model of Learning Prospective Behavior

The experimental results described above indicate that children construct representations of tasks and objects based upon their experiences in actually solving the tasks. We hypothesize that the evolution of behaviors in the child study is the result of a search for action sequences that produce the highest expected future reward, and for features that recommend those actions. The youngest children in the study demonstrate a preconceived strategy to reach with the dominant hand and place the spoon in the mouth, presumably with the expectation of receiving a reward (applesauce) for their actions. When presented with a novel experience on difficult trials, the strategy fails, and they must reevaluate the value of their actions in

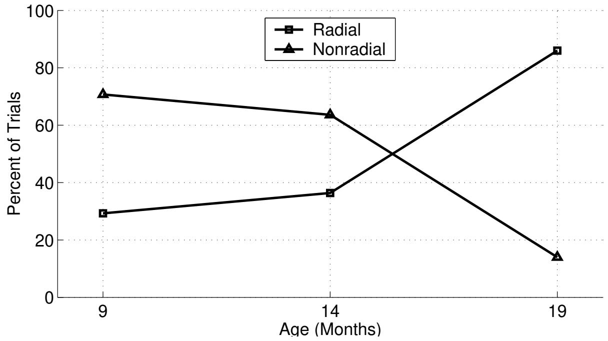


Fig. 2. Percentage of radial and nonradial grips used by children on difficult trials. (Data adapted from McCarty, *et al.* [3], Figure 3).

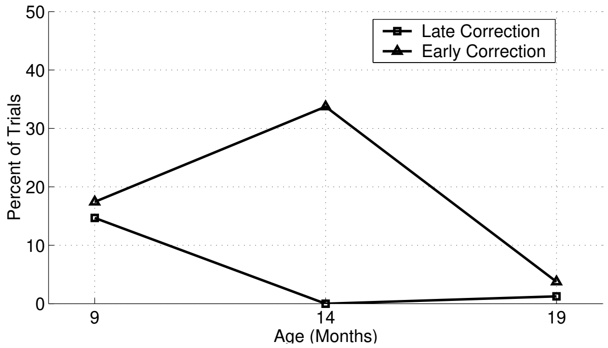


Fig. 3. Percentage over all trials of correction strategies used by children after an initial nonradial grip (adapted from McCarty, *et al.* [3], Figure 5 and Table 1).

the new situation. As they begin to differentiate between easy and difficult trials, one would expect the value of actions closest to the point of reward to be revised first. Since reward is received at the end of each trial, the first improvements should be evident in later actions. This is consistent with the child study, where an improvement at a later step (no longer placing the handle of the spoon in the mouth) was observed before an improvement in an earlier step (grasping the spoon with the appropriate hand).

We explore these hypotheses in the context of a robotic pick and place task. We present an intelligent agent with a task analogous to that used in the child study, and investigate the manipulation strategies exhibited by the agent as learning progresses. The agent controls a two-armed robot equipped with hands, tactile sensors, and a vision system. The objective is to grasp an object insert it into a receptacle. The object has two possible grasp targets, but a grasp by only one of the targets permits insertion. The object is presented to the robot in either of two orientations, such that each grasp target can be reached by one of the hands. The agent must learn the appropriate action sequences to grasp the object when presented in either orientation and insert it successfully.

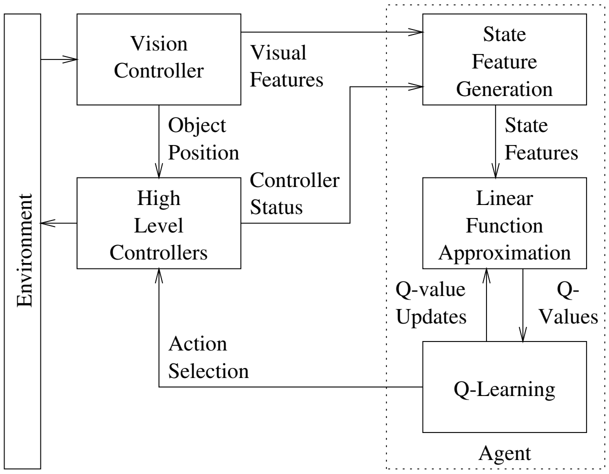


Fig. 4. System architecture.

A. Architecture

The system architecture used in our experiments is illustrated in Figure 4. A system of high-level closed loop controllers [5], [6] is used to implement six high-level actions that the agent can use to perform the task:

• Grasp Left - Use the left hand to pick the object up by the left grasp target.
• Grasp Right - Use the right hand to pick the object up by the right grasp target.
• Swap Left to Right - Swap the object from the left hand to the right, and grasp the object by the right grasp target.
• Swap Right to Left - Swap the object from the right hand to the left, and grasp the object by the left grasp target.
• Insert Left - Insert the object held in the left hand into the receptacle.
• Insert Right - Insert the object held in the right hand into the receptacle.

The agent incorporates a Discrete Event Dynamic System (DEDS) [5] layer to prohibit dangerous actions. DEDS can also accelerate learning by prohibiting known unproductive actions, and can be used to implement shaping (temporarily disabling selected actions to reduce the initial exploration space).

A vision controller provides visual input to the agent, and also provides position information directly to the high level reach and grasp controllers. Due to occlusion problems associated with visual sensing of a grasped object, we obtain a single image of the object prior to grasping.

High level action controllers are composed of lower level gross motion, fine motion, and grasp controllers. Gross motion controllers use harmonic path planning in configuration space [7] to implement large movements of the arms while providing collision avoidance. Fine motion controllers implement cartesian motions to perform initial positioning of the hands for grasping. Grasp controllers use contact position and normal feedback to minimize force and moment residuals [8], resulting in statically stable grasps.

The agent implements the Q−learning [9], [10], [4] algorithm with $\epsilon-$greedy exploration. Q−learning divides a task into steps, where a step consists of (1) sensing the system state, (2) performing one action, and (3) receiving a numeric reward. Based on previous experience, Q−learning estimates a *Q−value* for each action in each state, and selects actions based on their Q−value. The Q−value is an estimate of the expected future reward of performing a given action in a given state (i.e.: the immediate reward for taking the action, plus the discounted sum of the rewards that can be expected in subsequent state(s) that result from that action). The Q−value is estimated using the following incremental update equation:

$$Q(\phi_t, a_t) \leftarrow Q(\phi_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(\phi_{t+1}, a) - Q(\phi_t, a_t)],$$

where $Q(\phi_t, a_t)$ is the Q−value of action a_t (the action taken at time t) given the vector of state features ϕ_t. The new Q−value is estimated incrementally based on the previous estimate, the immediate reward, r_{t+1}, and $\max_a Q(\phi_{t+1}, a)$, the value of the "best" action that is available in the next state. γ is a discount factor that determines the importance of future rewards in the Q−value estimate. α is a learning rate chosen to provide fast learning while filtering out stochastic rewards, actions and state transitions.

With $\epsilon-$greedy exploration, Q−learning usually chooses the action with the highest Q−value; however, with some small probability, ϵ, an action is chosen at random to explore alternative actions which may lead to improved performance.

In our implementation, a linear function approximator (LFA) approximates the Q−values as a function of the state features, providing implicit feature-based state discrimination while allowing the agent to generalize experience across many different but related states. Function approximation also allows for direct use of continuous feature values, obviating the need to discretize the features and enumerate all possible feature combinations. An LFA was chosen over alternatives such as a multilayer neural network for improved learning speed. The function implemented by an LFA is represented in the form of a parameter vector (θ_{a_t}); the LFA estimates the Q−value of an action as:

$$Q(\phi_t, \theta_{a_t}) = \theta_{a_t}^T \phi_t,$$

where θ_{a_t} is the parameter vector associated with action a_t. When the Q−value of an action is updated, the LFA parameter vector is modified as follows:

$$\theta_{a_t} \leftarrow \theta_{a_t} + \alpha[r_{t+1} + \gamma \max_a(\theta_a^T \phi_{t+1}) - \theta_{a_t}^T \phi_t]\phi_t.$$

A state feature generator produces state features in the form of compositions of the visual and haptic input features. Visual features consist of a 100-element intensity histogram. Haptic features consist of two bits, one per hand, indicating the convergence status of the grasp controllers (1 if the hand is holding the object, otherwise 0). The output of the generator is a 303 element state feature vector containing the Cartesian product of the visual and haptic features, the haptic features themselves, and a constant 1 to provide a bias term for the LFA.

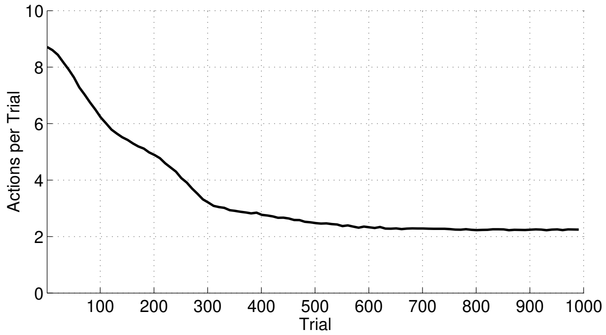
B. Experiments

Simulation runs were performed for scenarios with and without a "dominant hand". Runs with a dominant hand were used for comparison to the child study. Prior to the study, the children had already developed a dominant hand and an associated preconceived grasping strategy. We gave the agent a similar dominant hand strategy by pretraining it for 400 trials with the object always presented in the same orientation. After pretraining, the agent was analogous to the youngest children in the child study; it had developed a dominant hand strategy that made no distinction as to object orientation. As with the children, it must learn to suppress that strategy to perform optimally on difficult trials. After pretraining, the agent performed 600 trials with the object presented randomly in one of two orientations. Runs without a dominant hand provided baseline performance data. On those runs, the agent was not pretrained, and simply performed 1000 trials with random orientation.

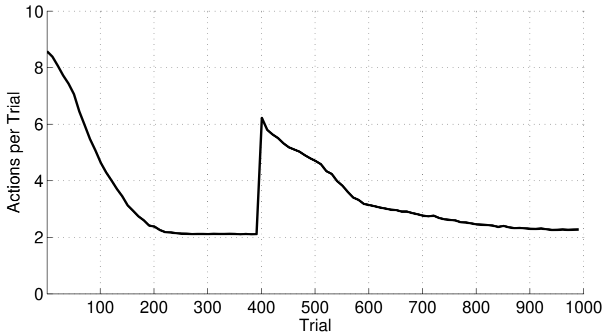
For both scenarios, a trial began with the object presented to the agent in one of two orientations, and ended when the object was inserted in the correct orientation, or after a maximum of ten actions. The agent was given a reward of -1 on each action; in order to maximize the expected future reward, the agent must accomplish the task with the fewest actions. All runs used parameters α (learning rate) $= 0.004$, ϵ (exploration rate) $= 0.05$, and γ (discount factor) $= 0.99$. Intensity histograms were generated from video images of a peanut butter jar in two different orientations. Simulation results were averaged longitudinally in sets of 10 trials over 1000 runs.

Figure 5a shows the average number of actions per trial without a dominant hand. Initially, the untrained agent performed poorly, with most trials timing out after ten actions. Over time, the agent learned to distinguish between the two object orientations based on the visual features, and learned the optimal action sequence for each orientation. After approximately 650 trials, the agent exhibited nearly optimal performance (with a fixed nonzero exploration rate, performance will never reach optimality).

Figure 5b shows the average number of actions per

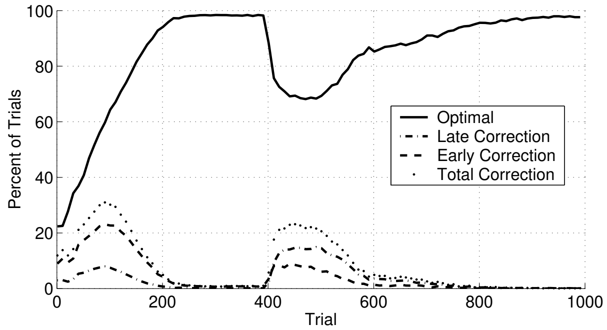


(a)

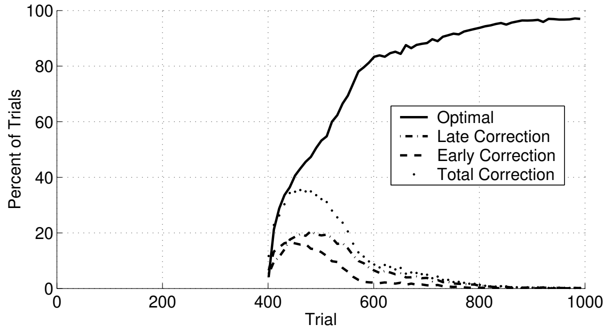


(b)

Fig. 5. Average number of actions per trial (a) without a dominant hand and (b) with a dominant hand (including 400 pretraining trials). For both cases, optimal performance is two actions: A grasp with the left or right hand (depending on the initial object orientation) followed by an insertion using the same hand.



(a)



(b)

Fig. 6. Percentage of optimal, late correction, early correction, and total correction (sum of late correction and early correction) strategies utilized by the agent during (a) easy trials and (b) difficult trials. The first 400 trials are pretraining trials consisting of all easy trials.

trial with a dominant hand. During pretraining starting at trial 0, the untrained agent initially exhibited poor performance, but approached optimality after approximately 250 trials. The agent learned faster during pretraining because the object was always presented in the same orientation, so the agent could rely on a dominant hand strategy. At trial 400, random orientations were introduced. Average actions per trial rose to over six primarily due to poor performance on difficult trials, but also due to interference with easy trials (refer to discussion of Figure 6a). Learning then progressed at the same rate as for trials without a dominant hand.

Figure 6 shows the percentage of optimal, early correction, and late correction strategies utilized by the agent in easy (a) and difficult (b) trials. Our results generally show similar characteristics whenever the agent encounters a new experience (when the untrained agent is presented with pretraining for easy trials, and when random orientations are introduced): There is an initial increase in nonradial grasps and subsequent corrective actions which tapers off as the agent learns the optimal strategy.

During pretraining for a dominant hand (early trials in Figure 6a), the agent explores various strategies as the Q−values transition from arbitrary initial values to appropriate action value estimates. However, with the object always presented in the same orientation, visual distinctions are unnecessary, and the agent quickly learns that late corrections are a poor strategy, early corrections are better, and the optimal strategy is best. A comparison to the child study at this stage is unrealistic; the purpose of the pretraining is to develop a dominant hand strategy in the agent, but the children had already developed a dominant hand prior to the study.

In the easy trials after random orientations are introduced (starting at trial 400 in Figure 6a), our results show a temporary drop in the optimal strategy as the agent explores nonradial grips. We attribute this to "overgeneralization." When random orientations are introduced, dominant hand grasps are the best strategy only half of the time. Eventually, the agent learns to differentiate between the orientations, but the initial effect is that the agent begins to learn that the average value of a dominant hand grasp has decreased. It is not apparent whether this effect was observed in the McCarty *et al.* study; planned longitudinal studies will provide additional insight.

In the difficult trials (starting at trial 400 in Figure 6b), the increase in corrective actions is consistent with results from the child study; the agent uses the preconceived strategy and chooses a nonradial grip, but then must make a correction to complete the task. However, two differences from the child study are apparent:

1. With young children, the nonradial corrective strategies are dominant, whereas in our results, corrective strategies never dominate the optimal strategy.
2. In the child study, late corrections disappear before early corrections, but in our data, early corrections disappear first.

We attribute these effects to a difference in the way features are processed. Our system makes discriminations based on all available features, while children may employ a discovery process to identify features with high discrimination utility. With respect to (1), our agent immediately begins to learn the optimal strategy, while children would tend to use the preconceived strategies until discriminating feature(s) had been discovered. With respect to (2), our agent must learn significantly different action values using visual features, most of which have little or no discrimination utility in our system. Once children had discovered discriminating features, they could readily reject the late correction strategy.

IV. Conclusions and Future Work

We have shown that accommodation of pick and place constraints can be expressed as prospective behavior, and can be learned by an intelligent agent using experience gained through interaction with the environment. This allows the agent to identify features that recommend various actions, and to adapt to nonstationary environments. We are currently porting the agent architecture to an actual robot, the UMass Torso. This effort will provide a foundation for additional enhancements such as the work of Coelho [11], in which observations of the dynamics of grasp controllers are used to identify haptic categories. These categories provide valuable shape information to the grasp controller itself, and can also provide useful haptic features to a higher level agent.

We have also proposed a computational model for behavioral development in children, and made a prediction testable by future child studies. We have shown that RL can account for some aspects of the behavioral development in children, but recognition of discriminating features also appears to play an important role. We used an LFA to implement a form of discrimination based on function approximation, which considers the combined effect of all features. However, differences in behavioral development between our system and human children suggest that children may employ a different technique which identifies the most salient features as needed. One such technique is suggested by the work of Piater, *et al.* [12], [13] in which Bayesian networks are employed to estimate the utility of features or compositions of features for discriminating between externally defined classes. Piater's work allows for a flexible definition of class membership; in addition to experimenter-defined classes, preliminary experiments by Coelho *et al.* [14] indicate that visual features can be used to recognize classes representing haptic categories defined using grasp controllers.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CISE/CDA-9703217, and IRI-9704530, DARPA MARS DABT63-99-1-0004, and NASA/RICIS. The authors gratefully acknowledge the valuable input provided by Rob Platt and Danny Radhakrishnan, our colleagues in the UMass Torso group.

References

[1] T. Lozano-Pérez, "Automatic planning of manipulator transfer movements," *IEEE Transactions Systems, Man, and Cybernetics*, vol. 11, no. 10, pp. 681–689, 1981.

[2] 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.* May 1990, pp. 683–688, IEEE.

[3] M. E. McCarty, R. K. Clifton, and R. R. Collard, "Problem solving in infancy: The emergence of an action plan," *Developmental Psychology*, vol. 35, no. 4, pp. 1091–1101, 1999.

[4] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[5] M. Huber and R.A. Grupen, "A hybrid discrete event dynamic systems approach to robot control," Tech. Rep. 96-43, University of Massachusetts Amherst Computer Science Department, October 1996.

[6] M. Huber, W.S. MacDonald, and R.A. Grupen, "A control basis for multilegged walking," in *Proceedings of the 1996 IEEE Conference on Robotics and Automation*, Minneapolis, MN., 1996, pp. 2988–2993, IEEE.

[7] C. Connolly and R. Grupen, "On the applications of harmonic functions to robotics," *Journal of Robotics Systems*, vol. 10, no. 7, pp. 931–946, 1993.

[8] Jefferson A. Coelho Jr. and Roderic A. Grupen, "A control basis for learning multifingered grasps," *Journal of Robotic Systems*, vol. 14, no. 7, pp. 545–557, 1997.

[9] C.J.C.H. Watkins, *Learning from Delayed Rewards*, Ph.D. thesis, Cambridge University, 1989.

[10] C.J.C.H. Watkins and P Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[11] J. Coelho, *Multifingered Grasping: Grasp Reflexes and Control Context*, Ph.D. thesis, University of Massachusetts, Amherst, MA, 2001.

[12] 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.

[13] J.H. Piater and R.A. Grupen, "Feature learning for recognition with bayesian networks," in *Proceedings of the Fifteenth International Conference on Pattern Recognition.* September 2000, IEEE.

[14] J. Coelho, J.H. Piater, and R.A. Grupen, "Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot," in *First IEEE-RAS International Conference on Humanoid Robots.* September 2000, IEEE.