International Journal of Humanoid Robotics
Vol. 4, No. 2 (2007) 387–406

World Scientific
www.worldscientific.com

A COGNITIVE MODEL OF IMITATIVE DEVELOPMENT IN HUMANS AND MACHINES

AARON P. SHON[*,‡], JOSHUA J. STORZ[*,§], ANDREW N. MELTZOFF[†,¶] and RAJESH P. N. RAO[*,||]

[*]*Department of Computer Science and Engineering, University of Washington, Paul G. Allen Center, Seattle, WA 98195, USA*

[†]*Institute for Learning and Brain Sciences, University of Washington, Seattle, WA 98195, USA*

[‡]*aaron@cs.washington.edu*
[§]*jstorz@cs.washington.edu*
[¶]*meltzoff@u.washington.edu*
[||]*rao@cs.washington.edu*

Received 23 September 2006
Revised 21 December 2006

Several algorithms and models have recently been proposed for imitation learning in humans and robots. However, few proposals offer a framework for imitation learning in noisy stochastic environments where the imitator must learn and act under real-time performance constraints. We present a novel probabilistic framework for imitation learning in stochastic environments with unreliable sensors. Bayesian algorithms, based on Meltzoff and Moore's AIM hypothesis for action imitation, implement the core of an imitation learning framework. Our algorithms are computationally efficient, allowing real-time learning and imitation in an active stereo vision robotic head and on a humanoid robot. We present simulated and real-world robotics results demonstrating the viability of our approach. We conclude by advocating a research agenda that promotes interaction between cognitive and robotic studies of imitation.

Keywords: Imitation; Bayesian inference; predictive models; plan recognition.

1. Introduction: Imitation Learning in Animals and Machines

The capacity of human infants to learn and adapt is remarkable. A few years after birth, a child is able to speak, read, write, interact with others, and perform myriad other complex tasks. In contrast, digital computers possess limited capabilities to learn from their environments. The learning they exhibit arises from explicitly programmed algorithms, often tuned for very specific applications. Human children accomplish seemingly effortlessly what the artificial intelligence community has labored more than 50 years to accomplish, with varying degrees of success.

(a) (b)

Fig. 1. Neonates and infants are capable of imitating body movements: (a) Tongue protrusion, opening the mouth, and lip protrusion are imitated by neonates.[23] (b) A 14-month-old child replays an observed, novel sequence required to cause a pad to light up.[21] The precise action needed to obtain the goal state of lighting up the pad is unlikely to be discovered by the child through random exploration of actions. Demonstration by an adult determines the child's later interactions with the pad. Prior models (discussed below) encode social preference or other contextual biases to guide action selection.

One key to solving this puzzle is that children are highly adept at observing and imitating the actions of others (see Fig. 1). Imitation is a versatile mechanism for transferring knowledge from a skilled agent (the *instructor*) to an unskilled agent (or *observer*) using direct demonstration rather than manipulating symbols. Various forms of imitation have been studied in monkeys and apes,[7,33,34] in children (including infants only 42 min old),[22–24] and in an increasingly diverse selection of machines.[11,18,28] The reason behind the growing interest in imitation in the machine learning and robotics communities is obvious: a machine with the ability to imitate has a drastically lower cost of reprogramming than one which requires programming by an expert. Imitative robots also offer testbeds for cognitive researchers to test computational theories, and provide modifiable agents for contingent interaction with humans in psychological experiments.

In this article, we discuss several findings and a basic model from the psychological literature on imitation. We show how a Bayesian model captures some developmental features of imitative learning in young children. Simulated results of an agent in a maze-like environment demonstrate the viability of the approach. We also briefly discuss two separate robotic platforms that illustrate how our Bayesian framework might lead to more flexible types of learning in robots. Our framework illustrates the potentially rich connections between cognitive modeling and probabilistic methods for learning in agents.

2. Related Work

Our approach shares some similarities with other recent model-based architectures for robotic imitation and control. Billard and Matarić developed a system that

learned to imitate movements using biologically-inspired algorithms.[5] The system learned a recurrent neural network model for motor control based on visual inputs from an instructor. A software avatar performed imitation of human arm movements. Demiris and Hayes proposed coupled forward and inverse models for robotic imitation of human motor acts.[10] Their architecture also draws inspiration from the AIM model of infant imitation. Moving beyond this deterministic, graph-based approach, we see advantages in using probabilistic models to handle real-world, noisy data. Wolpert and Kawato likewise proposed learning probabilistic forward and inverse models to control a robotic arm.[12,36] Another promising approach by Demiris is an architecture for robotic imitation that learns probabilistic forward models,[8,9] where learned Bayesian networks represent a forward model for a robotic gripper. These proposals still require the (potentially difficult) explicit learning of inverse models. Further, it is unclear how well these architectures mirror developmental stages observed in human infants.

Our framework differs from most previous efforts by employing Bayesian inference to decide which actions are most efficacious. A unique feature of our approach is the combination of forward models to predict environmental state and prior models expressing the instructor's preference over actions, rather than learning combinations of forward and inverse models. Related ideas are discussed in Refs. 3, 4, 27, 29 and 31.

2.1. *The AIM model*

Figure 2(a) provides a conceptual schematic of Meltzoff and Moore's active intermodal mapping (AIM) hypothesis for imitation in infants.[23,24] The key claim is that imitation is a matching-to-target process. The active nature of the matching process is captured by the proprioceptive feedback loop. The loop allows infants' motor performance to be evaluated against the seen target and serves as a basis for correction. Imitation begins by mapping perceptions of the teacher and the infant's own somatosensory or proprioceptive feedback into a supramodal representation, allowing matching to the target to occur.

2.2. *Developmental changes in imitative behavior*

A number of psychological theories of imitation have been proposed. Experimental evidence obtained by Meltzoff and colleagues indicates a developmental evolution of infants' imitative capabilities, beginning with exploratory movements of muscle groups and progressing to imitating bodily movements, imitating actions performed on objects, and finally inferring intent during imitation. Any system that attempts to capture the imitative capabilities of the developing child must address these characteristics.

The ability of human neonates as young as 42 min to imitate shows that imitation of a class of simple acts is present at birth; it is a biologically-endowed capability universal among typically developing children. It is also clear that newborns do not

begin life with the ability to perform imitative acts of arbitrary complexity. Neonates are capable of imitating facial gestures and gross bodily movements; infants gradually acquire more advanced imitative capabilities.[21,24] Based on the developmental work enumerated below, we suggest that the fundamental algorithms used by infants to perform imitation may not change significantly over time. Rather, more complex imitation occurs because infants gradually learn more complex models of their environment (and other agents in that environment). Gradual acquisition of more complex models for interacting with the environment, in turn, argues that these environmental models are distinct from one another (and possibly hierarchical).

2.3. *Imitative learning via inferring intent*

A later developmental step of imitation, and one where humans far exceed the capabilities of machines, is inferring intent — knowing what the instructor "means to do" even before the instructor has achieved a goal state (or when the instructor fails to reach a goal state). The ability to infer the intent of others represents a key step in forming a "theory of mind" for other agents, that is, being able to simulate the internal mental states of others. In one study,[20] 18-month-old infants were shown an adult performing an unsuccessful motor act. For example, the adult "accidentally" over- or under-shot his target object with his hands, or his hands "slipped" several times in manipulating a toy, preventing him from transforming the toy in some way. The results showed that infants did not re-enact what the adult actually did, but rather what he was trying to achieve (whereas control infants did not). This suggests that by, 18 months old, infants can infer which actions were intended, even when they have not seen the goal successfully achieved.

3. A Bayesian Framework for Goal-Directed Imitation Learning

Imitation learning systems that only learn deterministic mappings from state to actions are susceptible to noise and uncertainty in stochastic real-world environments. Development of a probabilistic framework for robotic imitation learning, capable of scaling to complex hierarchies of goals and subgoals, remains a largely untouched area of research. The following section sketches a proposal for such a framework. Systems that use deterministic models rather than probabilistic ones ignore the stochastic nature of realistic environments. We propose a goal-directed Bayesian formalism that overcomes both of these problems.

We use the notation s_t to denote the state of an agent at time t, and a_t to denote the action taken by an agent at time t. s_G denotes a special "goal state" that is the desired end result of the imitative behavior. Imitation learning can be viewed as a model-based goal-directed Bayesian task by identifying:

- **Forward model:** Predicts a probability distribution over future states given current state(s), action(s), and goal(s): $P(s_{t+1}|a_t, s_t, s_G)$. "Simulator"

that models how different actions affect the state of the agent and environmental state.

- **Inverse model:** Infers a distribution over actions given current state(s), future state(s), and goal(s): $P(a_t|s_t, s_{t+1}, s_G)$. Models which action(s) are likely to cause a transition from a given state to a desired next state.
- **Prior model:** Infers a distribution over actions given current state(s) and goal(s): $P(a_t|s_t, s_G)$. Models the policy (or preferences) followed by a particular instructor in transitioning through the environment to achieve a particular goal.

Learning inverse models is a notoriously difficult task,[15] not least because multiple actions may cause transitions from s_t to s_{t+1}. However, using Bayes' rule, we can infer an entire distribution over possible actions using the forward and prior models:

$$P(a_t|s_t, s_{t+1}, s_G) = \frac{P(a_t, s_t, s_{t+1}, s_G)}{P(s_t, s_{t+1}, s_G)} \tag{1}$$

$$= \frac{P(s_{t+1}|a_t, s_t, s_G)P(a_t, s_t, s_G)}{P(s_t, s_{t+1}, s_G)} \tag{2}$$

$$= \frac{P(s_{t+1}|a_t, s_t, s_G)P(a_t|s_t, s_G)}{P(s_t, s_{t+1}, s_G)P(s_t, s_G)} \tag{3}$$

$$\propto P(s_{t+1}|a_t, s_t, s_G)P(a_t|s_t, s_G) \tag{4}$$

$$\propto P(s_{t+1}|a_t, s_t)P(a_t|s_t, s_G), \tag{5}$$

where Eq. (5) follows because environmental dynamics are assumed first-order Markov (higher-order models taking into account a window of states $s_t \ldots s_{t+N}$ could, of course, be employed).

Equation (1) can be used to either select the maximum *a posteriori* (MAP) action, or to sample over the distribution of actions. The latter method occasionally picks an action different from the MAP action and potentially allows better exploration of the action space (cf. the exploration-exploitation tradeoff in reinforcement learning). Sampling from the distribution over actions is also called *probability matching*. Evidence exists that the brain employs probability matching in at least some cases.[13,16]

Figure 2(b) depicts a block diagram of our architecture. Like AIM, our system begins by running several feature detectors (skin detectors, face trackers, etc.) on sensor inputs from the environment. Detected features are monitored over time to produce state sequences. In turn, these sequences define actions. The next step is to transform state and action observations into instructor-centric values, then map from instructor-centric to observer-centric coordinates. Observer-centric values are employed to update probabilistic forward and prior models in our Bayesian inference framework. Finally, combining distributions from the forward and prior models as in Eq. (1) yields a distribution over actions. The resulting distribution over actions is converted into a single action the observer should take next.

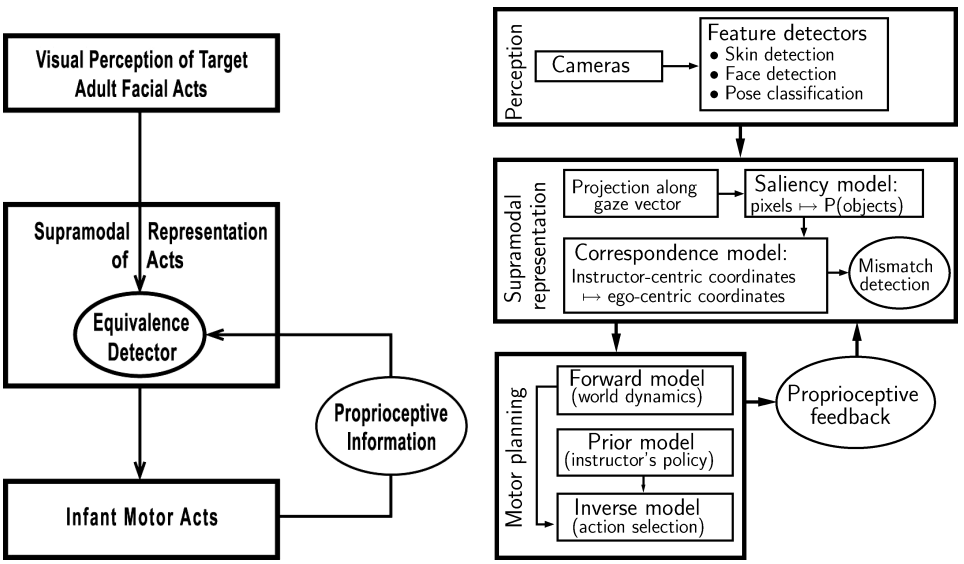


Fig. 2. Overview of model-based Bayesian imitation learning architecture: (a) Meltzoff and Moore's AIM model[24] argues that infants match observations of adults with their own proprioceptive information using a modality-independent ("supramodal") representation of state. Our computational framework suggests a probabilistic implementation of this hypothesis. (b) As in the AIM proposal, the initial stages of our model correspond to the formation of a modality-independent representation of world state. Mappings from instructor-centric to observer-centric coordinates and from the instructor's motor degrees of freedom (DOF) to the observer's motor DOF play the role of equivalence detector in our framework, matching the instructor's motor output to the motor commands of the observer. Proprioceptive feedback from the execution of actions closes the motor control loop.

3.1. *A Bayesian algorithm for inferring intent*

Being able to determine the intention of others is a crucial requirement for any social agent, particularly an agent that learns by watching the actions of others. One appealing aspect of our framework is that it suggests a probabilistic algorithm for determining the intent of the instructor. That is, an observer can determine a distribution over goal states based on watching what actions the instructor executes over some period of time. This could have applications in machine learning systems that predict what goal state the user is attempting to achieve, then offer suggestions or assist in performing actions that help the user reach that state.

Our algorithm for inferring intent uses applications of Bayes' rule to compute the probability over goal states given a current state, action, and next state obtained by the instructor, $P\left(s_G|s_{t+1}, a_t, s_t\right)$. This probability distribution over goal states represents the instructor's intent. In Markov decision process (MDP) problems, performing inference over which policy another agent is following is known as *plan*

recognition.[4,31] For robotic agents that need to react in real time, full-blown inference may prove impossible. Our approach represents a fast, greedy approximation to solving the full plan recognition problem for MDPs. One point of note is that $P(s_{t+1}|a_t, s_t, s_G) \equiv P(s_{t+1}|a_t, s_t)$, i.e. the forward model does not depend on the goal state s_G, since the environment is indifferent to the desired goal. Our derivation proceeds as follows:

$$P\left(s_G|s_{t+1}, a_t, s_t\right) = kP\left(s_{t+1}|s_G, a_t, s_t\right) P\left(s_G|a_t, s_t\right) \tag{6}$$

$$\propto P\left(s_{t+1}|a_t, s_t\right) P\left(s_G|a_t, s_t\right) \tag{7}$$

$$\propto P\left(s_{t+1}|a_t, s_t\right) P\left(a_t|s_G, s_t\right) P\left(s_t|s_G\right) P\left(s_G\right). \tag{8}$$

The first term in the equation above is the forward model. The second term represents the prior model (the "policy"; see above). The third term represents a distribution over states at time t, given a goal state s_G. This could be learned by, e.g. observing the instructor manipulate an object, with a known intent, and recording how often the object is in each state. Alternatively, the observer could itself "play with" or "experiment with" the object, bearing in mind a particular goal state, and record how often each object state is observed. The fourth term is a prior distribution over goal states characterizing how often a particular goal is chosen. If the observer can either assume that the instructor has a similar reward model to itself (the "Like-Me" hypothesis[21,22]), or model the instructor's desired states in some other way, it can infer $P(s_G)$.

Interestingly, the four terms above roughly match the four developmental stages laid out in Refs. 21 and 24. The first term is the forward model, whose learning is assumed to begin very early in development during the "body babbling" stage. The second term corresponds to a distribution over actions as learned during imitation and goal-directed actions. The third term refers to distributions over states of objects given a goal state. Because the space of actions an agent's body can execute is presumably much less than the number of state configurations objects in the environment can assume, this distribution requires collecting much more data than the first. Once this distribution is learned, however, it becomes easier to manipulate objects to a particular end — an observer that has learned $P(s_t|s_G)$ has learned which states of an object or situation "look right" given a particular goal. The complexity of this third term in the intent inference equation could provide one reason why it takes children much longer to learn to imitate goal-directed actions on objects than it does to perform simple imitation of body movements. Finally, the last term, $P(s_G)$, is the most complex term to learn. This is both because the number of possible goal states s_G is huge, and the fact that the observer must model the instructor's distribution over goals indirectly (the observer obviously cannot directly access the instructor's reward model). The observer must rely on features of its own reward model, as well as telltale signs of desired states to infer this prior distribution. For example, states that the instructor tends to act to remain in, or that cause the instructor to change the context of its actions, could be potential goal

states. The difficulty of learning this distribution could help explain why it takes so long for infants to acquire the final piece of their preverbal imitative toolkit — determining the intent of others.

4. Results

4.1. *Maze-world simulation*

We tested the proposed Bayesian imitation framework using a simulated maze environment. The environment [shown in Fig. 3(a)] consists of a 20×20 discrete array of states (thin lines). Thick lines in the figure denote walls, through which agents cannot pass. Three goal states exist in the environment; these are indicated by shaded ovals. Lightness of ovals is proportional to the *a priori* probability of the instructor selecting each goal state (reflecting, e.g. relative reward value experienced at each state). In this example, prior probabilities for each state (from highest to lowest) were set at $P(s_G) = \{0.67, 0.27, 0.06\}$. All instructor and observer trajectories begin at the lower left corner, maze location (1,1) (black asterisk).

Figures 3 and 4 demonstrate imitation results in the simulated environment. The task was to reproduce observed trajectories through a maze containing three different goal states [maze locations marked with ovals in Fig. 3(a)]. This simulated environment simplifies a number of the issues mentioned above: the location and value of each goal state is known by the observer *a priori*; the movements of the instructor are observed free from noise; the forward model is restricted so that only moves to adjacent maze locations are possible; and the observer has no explicit knowledge of walls (hence any wall-avoiding behavior results from watching the instructor).

The set of possible actions for an agent [Fig. 3(b)] includes moving one square to the north (N), south (S), east (E), or west (W), or simply staying put at the current location (X). The stochastic nature of the maze environment means an agent's selected actions will not always have the intended consequences.

Figure 3(c) compares the true probabilistic kernel underlying state transitions through the maze (left matrix) with the observer's forward model (right matrix). Lighter squares denote greater probability mass. Here the E and W moves are significantly less reliable than the N, S, or X moves. The observer acquires the estimated state transition kernel $\hat{P}(s_{t+1}|a_t, s_t)$ by applying randomly chosen moves to the environment for 500 simulation steps before imitation begins. In a more biologically relevant context, this developmental period would correspond to "body babbling," where infants learn to map from motor commands to proprioceptive states.[24]

After the observer learns a forward model of the environmental dynamics, the instructor demonstrates ten different trajectories to the observer (three to the white goal, four to the light gray goal, three to the dark gray goal), allowing the observer to learn a prior model. Figure 4(a) shows a sample training trajectory (dashed line) where the instructor moves from location (1,1) to goal 1, the state at (19,19)

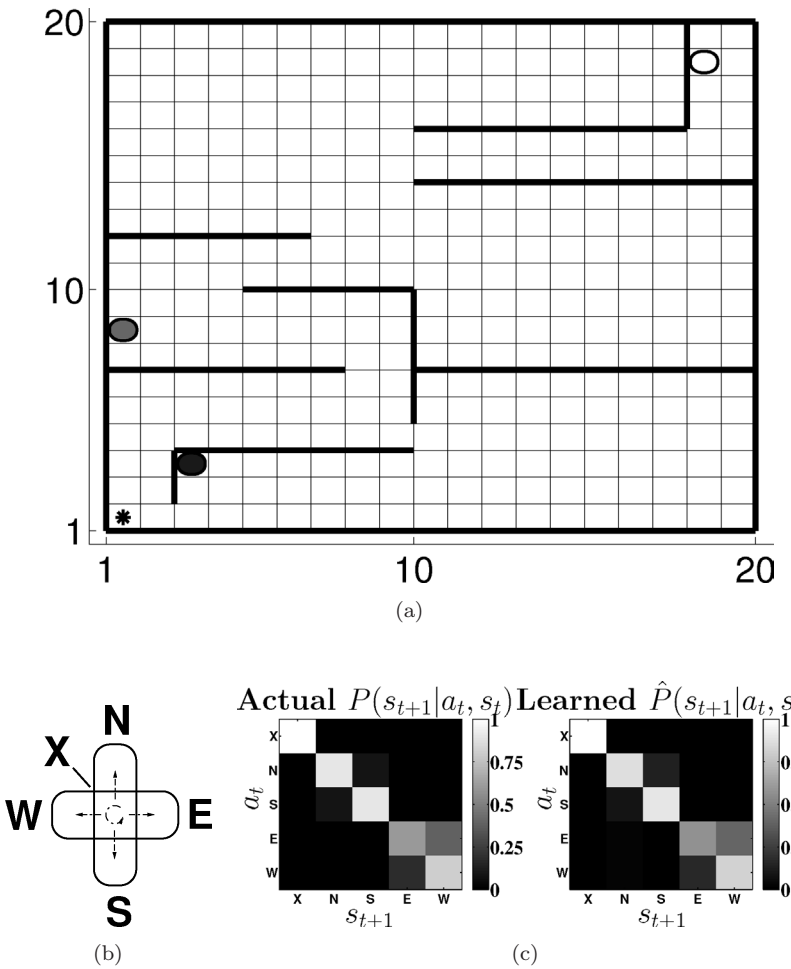


Fig. 3. Simulated environment for imitation learning: (a) Maze environment used to train observer. Thick black lines denote walls; ovals represent goal states. (b) Possible actions for the observer are relative to the current grid location: on each simulation step, the observer can choose to move North (N), South (S), East (E), West (W), or make no move (X). Because the environment is stochastic, it is not guaranteed that the chosen action will cause a desired state transition. (c) Actual (left) and estimated (right) state transition matrices. Each row of each matrix encodes a probability distribution $P(s_{t+1}|a_t, s_t)$ of reaching the desired next state given a current state and action.

indicated by the white oval. The solid line demonstrates the observer moving to the same goal after learning has occurred. The observer's trajectory varies somewhat from the instructor's due to the stochastic nature of the environment but the final state is the same as the instructor's.

We tested the intent inference algorithm using the trajectory shown in Fig. 4(b) where the instructor moves toward goal 1. The observer's task for this trajectory is

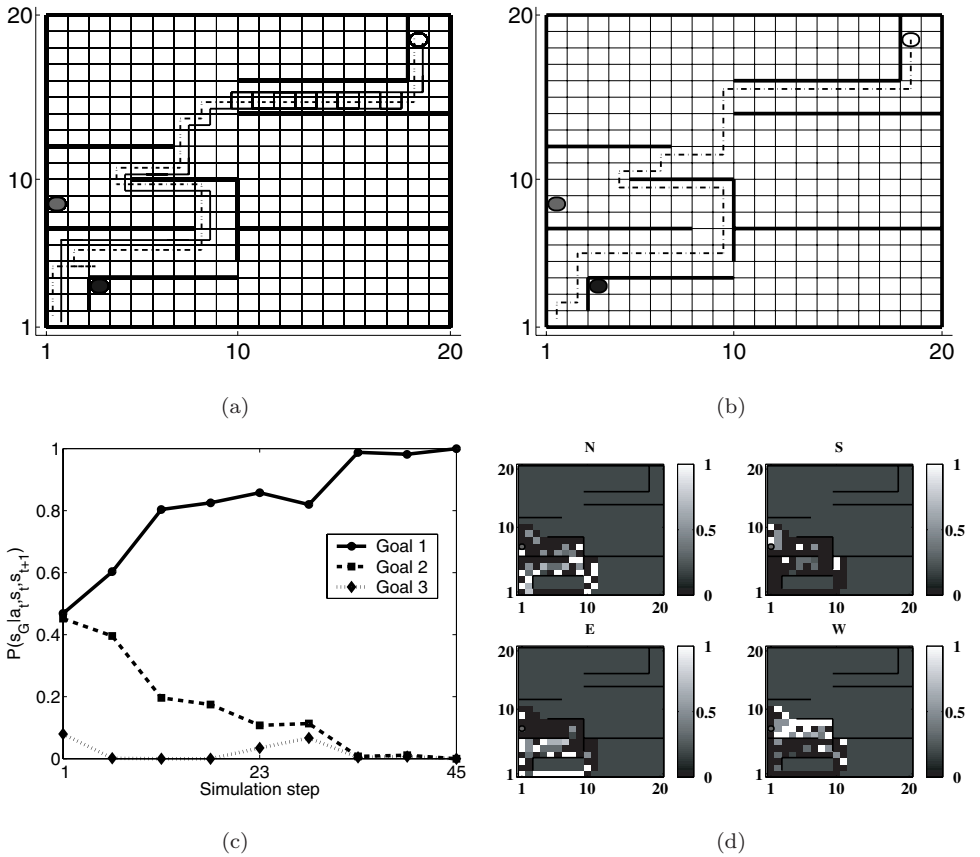


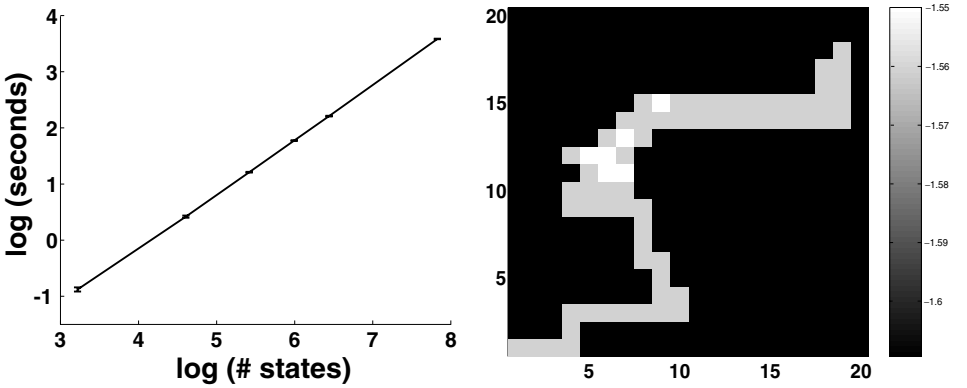
Fig. 4. Imitation and goal inference in simulation: (a) The observer (solid line) successfully imitates the instructor's trajectory (dashed line), starting from map location (1,1). Grayscale color of the lines indicates time: light gray denotes steps early on in the trajectory, with the line gradually becoming darker as the timecourse of the trajectory continues. Note that the observer requires more steps to complete the trajectory than does the instructor. This is because the environment is stochastic and because of the limited training examples presented to the observer. (b) Sample trajectory of the instructor moving from starting location (1,1) to the upper right goal (goal 1). The task is for the observer to use its learned models to estimate a distribution over possible goal states while the instructor is executing this trajectory. Total length of the trajectory is 45 simulation steps. (c) Graph showing a distribution over instructor's goal states inferred by the observer at different time points as the instructor is executing the trajectory in (b) . Note how the actual goal state, goal 1, maintains a high probability relative to the other goal states throughout the simulation. Goal 2 initially takes on a relatively high probability due to ambiguity in the training trajectories and limited training data. After moving past the ambiguous part of the trajectory, goal 1 (the correct answer) clearly becomes dominant. Each data point in the graph shows an average of the intent distribution taken over 5 time steps. (d) Example prior distributions $P(a_t|s_t, s_G)$ learned by the observer. In this case, we show the learned distribution over actions for each state given that the goal state s_G is goal 2, the gray oval.

to infer, at each time step of the trajectory, the intent of the instructor by estimating a distribution over possible goal states the instructor is headed toward. The graph in Fig. 4(c) shows this distribution over goals, where data points represent inferred intent averaged over epochs of five simulation steps each (i.e. the first data point on the graph represents inferred intent averaged over simulation steps 1–5, the second data point spans simulation steps 6–10, etc.). Because of the prior probabilities over goal states (given above), the initial distribution at time $t = 0$ would be $\{0.67, 0.27, 0.06\}$ (and hence in favor of the correct goal). Note that the estimate of the goal, i.e. the goal with the highest probability, is correct over all epochs. As the graph shows, the first epoch of five time steps introduces ambiguity: the probabilities for the first and second goals become almost equal. This is because the limited number of training examples were biased toward trajectories that lead to the second goal. However, the algorithm is particularly confident once the ambiguous section of the trajectory, where the instructor could be moving toward the dark gray or the light gray goal, is passed. Performance of the algorithm would be enhanced by more training; only ten sample trajectories were presented to the algorithm, meaning that its estimates of the distributions on the right-hand side of Eq. (8) were extremely biased.

Figure 4(d) shows examples of learned prior distributions $\hat{P}(a_t|s_t, s_G)$ for the case where the goal state s_G is goal 2, the light gray oval at map location (8,1). The plots show, for each possible map location s_t that the observer could be in at time t, the prior probability of each of the possible actions N, E, S, and W, given that s_G) is goal 2 (the X action is not shown since it has negligible probability mass for the prior distributions shown here). In each of the four plots, the brightness of a location is proportional to the probability mass of choosing a particular action at that location. For example, given goal 2 and the current state $s_t = (1, 1)$, the largest probability mass is associated with action E (moving east). These prior distributions encode the preferences of the instructor as learned by the observer.

Figure 5(a) provides an indication of the inference algorithm's scalability. Here we show a log–log plot of the number of states in the maze environment versus the runtime (in seconds)[a] required to learn the distributions $P(a_t|s_t, s_G), P(s_t|s_G), P(s_G)$. Each data point represents an average over ten randomly generated mazes, each trained with 50 training trajectories. Error bars represent standard deviations. Our model's runtime scales linearly with the number of states. This suggests the value of imitation: algorithms such as policy iteration for MDPs can require $O(N^3)$ time in the number of states N, and exponential time for variants such as simple policy iteration.[17,19] Although we have not yet proved efficiency bounds for our approach, the value of imitation in general is clear; while naive approaches to reinforcement learning can require time exponential in the number of states and actions, a knowledgeable instructor can focus a learning agent's attention on a small subset of states and actions that give rise to valuable

[a]Run on a Pentium Xeon using non-optimized Matlab code.



(a) (b)

Fig. 5. Performance data and state criticality:(a) Log–log plot of number of states ($5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25, 50 \times 50$) versus learning time required on a 2-processor workstation (in seconds). Each data point represents the average run time across ten randomly generated mazes. Error bars indicate standard deviations. 50 instructor training trajectories were used for each maze. The slope of 0.97 on the log–log plot shows that our learning algorithm's time complexity is roughly linear in the number of states. (b) Entropy of states in the maze shown in Fig. 4, after a sample training trajectory moving toward the goal at top right. Lower entropy (lighter squares) denote states where the mentor is surer of its action distribution, implying a critical state on the path to the goal.

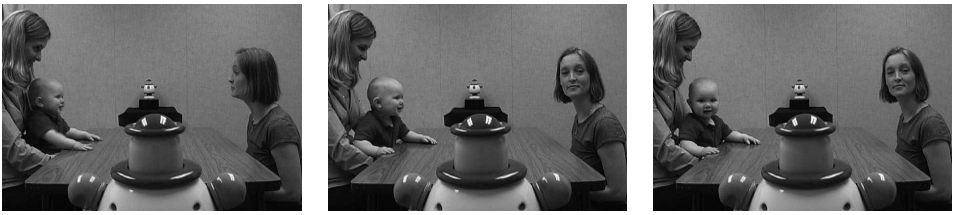
trajectories. In Fig. 5(b), we show entropy of the action distribution $P(a_t|s_t, s_G)$ for each state, derived from a single training trajectory. While state transition counts can determine $P(s_G)$, the probability that a particular state is a goal, entropy could be used to determine the "criticality" of a particular state. In this example, lighter squares denote locations where the mentor's action distribution is sharply peaked, implying that choosing a particular action from that state is especially critical. We anticipate further exploring a combination of computing $P(s_G)$ and action entropy to determine the relevance of environmental states during learning.

4.2. *Robotic gaze tracking*

In infants, following the gaze of another to distal objects is a critical ability for directing future learning (Fig. 6).

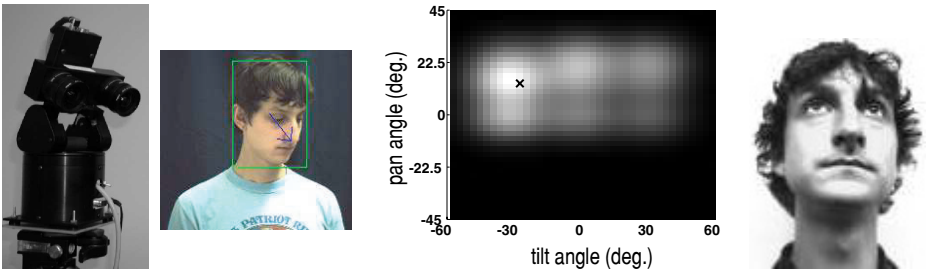
We previously demonstrated how our framework can be used to implement a robotic system for identifying salient objects based on instructor gaze.[14] Figure 7 shows an example of the system using vision to track the gaze of a human instructor.

Vision-based algorithms[32,37] find the instructor's face and estimate pan and tilt angles for the instructor's head direction. Figure 7(b) shows an example estimate for the face shown in (d). The system learns a prior model over which object sizes and colors are salient to a particular instructor. The prior model is combined with the likelihood term from the look direction of the instructor to determine the object



(a) (b) (c)

Fig. 6. Infants track gaze to establish shared attention: Infants as young as 9 months can follow an adult's head turn to an object; older infants ($\geq$ 10 months) use opened eyes as a cue to detect whether they should perform gaze tracking. In this example, an infant and instructor begin interacting (a). When the instructor looks toward an object (b), the infant focuses attention on the same object using the instructor's gaze as a cue (c). See Refs. 6 and 21 for relevant experiments.



(a) (b) (c)

Fig. 7. Gaze tracking in a robotic head: (a) Left: A Biclops active stereo vision head from Metrica, Inc. Right: The view from the Biclops' cameras as it executes our gaze tracking algorithm. The bounding box outlines the estimated location of the instructor's face, while the arrow shows the instructor's estimated gaze vector. (b) Likelihood surface for the face shown in (c), depicting the likelihood over pan and tilt angles of the subject's head. The region of highest likelihood (the brightest region) matches the actual pan and tilt angles (black $\times$) of the subject's face shown in (c).

from a cluttered scene to which the instructor is most likely attending. The set of motor encoder values of the system corresponds to the modality-independent space at the heart of the AIM hypothesis: the robotic head maps both its own motor acts and the observations of the human into pan and tilt joint angles, allowing it to determine which action will best match the human's looking direction.

4.3. *Humanoid implementation*

We have conducted preliminary experiments that show how the general framework proposed here might be implemented on a humanoid robot. We use a 12-camera Vicon motion capture system to train a HOAP-2 humanoid robot using motion models from a human instructor. Inverse kinematics software fits motion capture

marker data to a model of the robot's skeleton. The goal is to use imitation to teach the robot simple manipulation tasks, in this case lifting a box to chest height. The framework runs in real-time; all learning takes place online, and the system can be switched at any time from watching the human to attempting to achieve a goal.

The motion capture system reports a list of human joint positions at time t, $v_t = \{\theta_1 \ldots \theta_N\}$, and a 3D coordinate frame for each object i in the workspace, $s_t = \{x, y, z, \theta_x, \theta_y, \theta_z\}$. All objects are presently assumed to be non-articulated, rigid bodies. For the simple task of lifting a box, we collapse s_{t+1} and s_G into a single goal state; Eq. (5) then becomes

$$P(a_t|s_t, s_{t+1}) \propto P(s_{t+1}|a_t, s_t)P(a_t|s_t). \tag{9}$$

We represent all distributions of interest with sparse histograms. Histograms are easy to implement and lead to an efficient implementation; however, they do not generalize well over many different initial states and end states. Our ongoing work therefore includes representation of distributions using continuous models, in particular Gaussian processes.[35]

Actions are encoded as hidden Markov models (HMMs). Our system maintains a database of these models captured over time. Once a sequence of inferred joint angles has been assigned to an HMM, the model is trained using the well-known Baum–Welch algorithm.[26] When a sequence of human activity is observed, the system uses the Viterbi algorithm[26] to determine the log-likelihood that each model in the database generated the observed sequence. If the log-likelihood for at least one model in the database is above a threshold $\epsilon = 0$, we assign the sequence to the winning model and retrain it with the new sequence included; otherwise we create a new HMM and assign the sequence to it.

Given a prior object state S_t, an estimated action $\hat{a}_t$ executed by the human at time t, and a final object state S_{t+1} reached after the human's action is complete, the system updates its prior model $P(a_t|s_t)$ and its estimate of environmental dynamics for the human's actions $P_h(s_{t+1}|a_t, s_t)$. This updating process continues until the human is finished showing the robot a set of actions needed to accomplish the task. Three different coordinate origins, O_v, O_h, O_r, are used to align state spaces between the human and robot. Each origin is defined by a set of offsets and scaling factors: $O = \{x, y, z, \theta_x, \theta_y, \theta_z, \sigma_x, \sigma_y, \sigma_z\}$. $\sigma_x, \sigma_y, \sigma_z$ respectively indicate scaling factors along the X, Y, Z axes, relative to the motion capture world origin O_v. When the robot is "watching" the human, O_h is used; when the robot is imitating, O_r is used. The origin-corrected object position s_t defines an intermodal space in the same spirit as AIM. In this experiment, the robot selects an action to match the state of the box when it lifts with the state of the box when the human lifts, rotated and scaled appropriately.

When the robot is instructed to imitate, it determines an optimal action a_t^* by convolving $P(a_t|s_t)$ with $P_h(s_{t+1}|a_t, s_t)$. After it finishes executing the action, the robot collects an updated snapshot of the object state, s_{t+1}. This is used to update a different forward model, $P_r(s_{t+1}|a_t, s_t)$, representing the robot's estimate of how

its actions influence world state. Given a starting state s_t and a candidate action a_t, the robot uses P_h if the number of samples in $P_h(\cdot|a_t, s_t)$ is equal to or greater than the number of samples in $P_r(\cdot|a_t, s_t)$. The human's prior model therefore helps guide the robot toward exploring potentially profitable actions, while the robot builds up a probabilistic model of how reliable each of those actions is likely to be.

Figure 8 shows how Bayesian action selection as described in Eq. (5) operates as the robot acquires more information about environmental dynamics. The human demonstrated a total of 25 different attempts to lift the box; some attempts by the human instructor failed to attain the correct discretized state. Based on thresholding the log-likelihood of the Viterbi score as described above, the system discovers a total of eight actions used by the human instructor to lift the box; these actions represent two different styles of one-handed and two-handed lifting actions. For clarity, the figure only contains log likelihoods for the two most probable actions. One particular one-handed lift has much higher prior likelihood, since the human demonstrates it more often than other types of lift. The second most likely lift (according to the prior model) is a two-handed lift. The robot tries the one-handed action with high prior likelihood for 13 iterations; by the 14th iteration, the number of samples in P_r finally exceeds the number in P_h. This leads the robot to reestimate the log likelihood of reaching the goal using the one-handed action (which, since

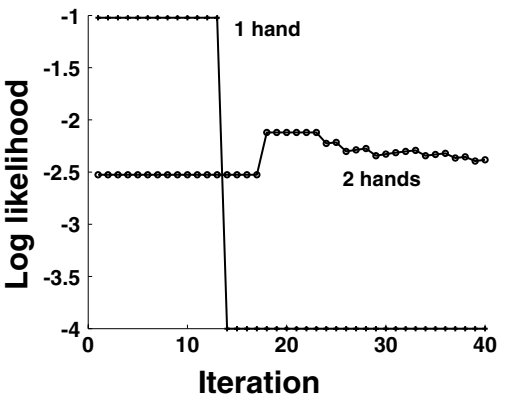


Fig. 8. Learning environmental dynamics using a combination of watching and exploring: We used motion capture to train a HOAP-2 humanoid robot to pick up a box. A human demonstrator (right) shows two types of actions to accomplish the goal of lifting the box to chest height (one-handed versus two-handed lifting). Actions are encoded as hidden Markov models, and fit to a model of the robot's body using inverse kinematics software. After watching, the robot interacts with the box to determine which actions are effective at reaching the goal. The limited capabilities of the robot ensure that it can only succeed if it tries a two-handed lift. Initially the robot tries a one-handed lift. After it has built up sufficient samples to reestimate its forward model of how the box acts (at iteration 14), the robot begins to use a two-handed lift instead. Because the one-handed lift never succeeds, the estimated log likelihood of its success becomes clamped at the minimum value of -4. Log likelihood of the two-handed action fluctuates; often it works, but occasionally it fails.

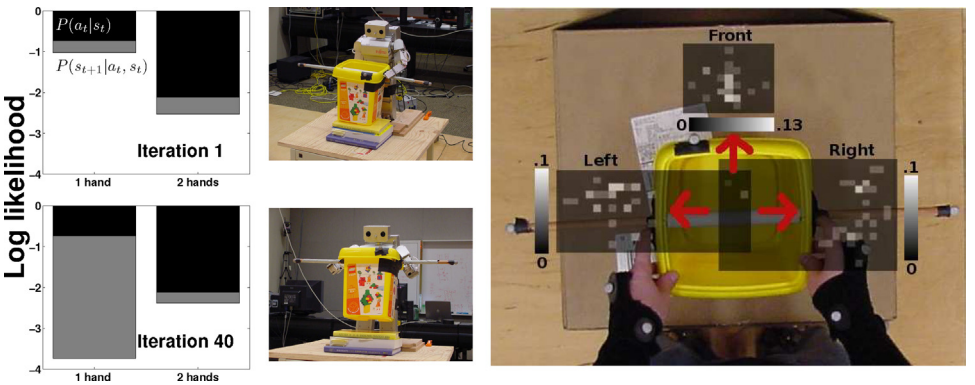


Fig. 9. Contribution of prior and forward models: At left, log likelihood estimates for two of the eight actions encoded using HMMs. Black bars represent prior model probabilities; grey bars represent forward model likelihood estimates. Note that while the prior probability stays unchanged over time (since it reflects preferences of the human instructor), the forward model likelihood changes once the robot has acquired sufficient samples from environmental interaction. Thus, early in the learning process (e.g. iteration 1), the robot favors the one handed lift action due to its higher likelihood value; with additional experience (e.g. iteration 40), it switches to the two-handed lift action, which is more reliable for the robot and has higher likelihood. Center column shows the robot attempting a one-handed action at iteration 1, and a two-handed action at iteration 40. At right, we show an example of sparse histograms representing the forward dynamics of the box when affected by three different actions. Grayscale squares reflect the probability of attaining a particular 1 cm^2 state location in the X–Y plane given the starting state of the box and one of the three actions (push forward, push left, push right). The human instructor provided 20 examples of each action, and the system correctly identified all 60 total action samples using the Viterbi algorithm.

it never succeeds, is clipped at a minimum log likelihood of −4). The robot then tries one of the two-handed actions, and succeeds. Small dips in the log likelihood of the two-handed action reflect the fact that it is not perfect; the robot sometimes drops the box even with two hands. As the number of trials increases, the robot's estimate of action reliability tends to reach a steady value reflecting the interaction between the robot and objects of interest. This has potential applications as a partial solution of the imitation correspondence problem;[1,2,24,25] the robot can determine quality of correspondence by examining how closely its estimated forward model matches that shown by the human. This example involved only two time steps, but 225,000 K total states (50 discretized X levels, 50 for Y, ten for Z, and three each for $\theta_x, \theta_y, \theta_z$). Only a small fraction of these states are actually encountered in the course of the human demonstrating the action, motivating our use of sparse histograms.

5. Conclusion

We have described a Bayesian framework for performing imitation-related tasks. The framework is modeled on cognitive findings and on the ideas of forward

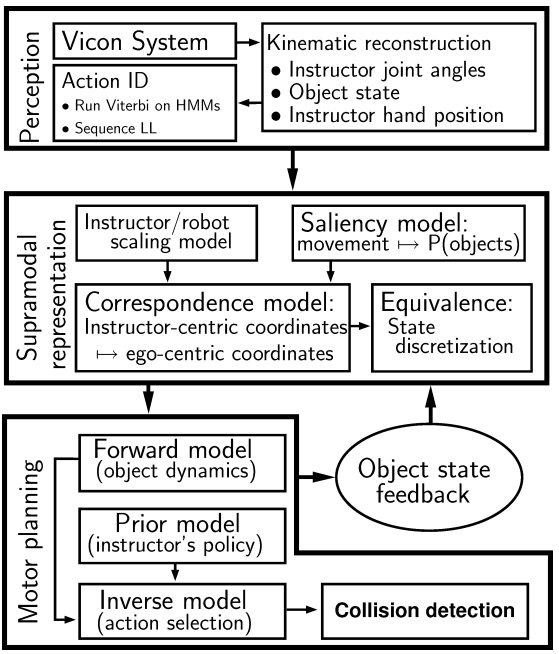


Fig. 10. Block diagram of HOAP-2 imitation on objects: Our system uses motion capture to identify actions and the effects of those actions on object states. A linear model (scaling, rotation, and translation) creates correspondences between states of the human actor and states of the humanoid. Object states are encoded as discretized 6D motion capture values $\{x, y, z, \theta_x, \theta_y, \theta_z\}$. Sparse histograms are used to represent forward and prior models. Contrast the architecture here with Fig. 2(b).

and inverse models from control theory. We described how model-based learning allows Bayesian-optimal action selection and plan recognition. We have implemented aspects of the framework in simulation, in an active vision head, and in a humanoid robot.

Ongoing work in our laboratories concentrates on both the cognitive and robotics sides of the imitation puzzle. Recently we have begun concentrating on a real-time implementation of our imitation system, allowing users to quickly program new tasks for our humanoid and to correct instructor mistakes.[30] We are also conducting studies of infant imitation and gaze following, using the humanoid as a flexible platform for testing which behavioral cues are important when children learn about social agents in their world.

Acknowledgments

We thank Keith Grochow and Danny Rashid for their respective assistance in implementing the motion capture and collision detection parts of our humanoid system. Thanks to Chris Baker, David Grimes, and Rechele Brooks for providing several

stimulating ideas. We gratefully acknowledge support from the NSF AICS program and the NSF SLC program (SBE-0354453). We thank the reviewers for their comments.

References

1. A. Alissandrakis, C. L. Nehaniv and K. Dautenhahn, Do as I do: Correspondences across different robotic embodiments, in *Proc. 5th German Workshop Artificial Life*, eds. D. Polani, J. Kim and T. Martinetz (2002), pp. 143–152.
2. A. Alissandrakis, C. L. Nehaniv and K. Dautenhahn, Imitating with ALICE: Learning to imitate corresponding actions across dissimilar embodiments, *IEEE Trans. Syst. Man Cybern.* **32** (2002) 482–496.
3. C. L. Baker, J. B. Tenenbaum and R. R. Saxe, Bayesian models of human action understanding, in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 18, eds. Y. Weiss, B. Schölkopf and J. Platt (2006), pp. 99–106.
4. C. L. Baker, J. B. Tenenbaum and R. R. Saxe, Goal inference as inverse planning, in *Proc. 29th Annual Conf. Cognitive Science Society* (2007).
5. A. Billard and M. J. Mataric, Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture, *Robot. Autonom. Syst.* **37**(2–3) (2001) 145–160.
6. R. Brooks and A. N. Meltzoff, The importance of eyes: How infants interpet adult looking behavior, *Dev. Psychol.* **38** (2002) 958–966 .
7. R. W. Byrne and A. E. Russon, Learning by imitation: A hierarchical approach, *Behav. Brain Sci.* **67** (2003) 667–721.
8. A. Dearden and Y. Demiris, Learning forward models for robotics, in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, pp. 1440–1445.
9. Y. Demiris and A. Dearden, From motor babbling to hierarchical learning by imitation: A robot developmental pathway, in *Proc. 5th Int. Workshop Epigenetic Robotics (EPIROB '05)*, eds. L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Komczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov and C. Balkenius (2005), pp. 31–37.
10. Y. Demiris and G. Hayes, Imitation as a dual-route process featuring predictive and learning components: A biologically-plausible computational model, in *Imitation in Animals and Artifacts*, eds. K. Dautenhahn and C. Nehaniv, Chap. 13 (MIT Press, 2002), pp. 327–361.
11. T. Fong, I. Nourbakhsh and K. Dautenhahn, A survey of socially interactive robots, *Robot. Autonom. Syst.* **42**(3–4) (2002) 142–166.
12. M. Haruno, D. Wolpert and M. Kawato, MOSAIC model for sensorimotor learning and control, *Neural Comput.* **13** (2000) 2201–2222.
13. R. J. Herrnstein, Relative and absolute strength of responses as a function of frequency of reinforcement, *J. Exp. Anal. Behav.* **4** (1961) 267–272.
14. M. W. Hoffman, D. B. Grimes, A. P. Shon and R. P. N. Rao, A probabilistic model of gaze imitation and shared attention, *Neural Networks* (2006).
15. M. I. Jordan and D. E. Rumelhart, Forward models: Supervised learning with a distal teacher, *Cogn. Sci.* **16** (1992) 307–354.
16. J. R. Krebs and A. Kacelnik, Decision making, in *Behavioural Ecology*, eds. J. R. Krebs and N. B. Davies, 3rd edn. (Blackwell Scientific Publishers, 1991), pp. 105–137.
17. M. L. Littmann, T. L. Dean and L. P. Kaelbling, On the complexity of solving Markov decision problems, in *Proc. Ann. Conf. Uncertainty in Artificial Intelligence* (*UAI 95*) (1995), pp. 394–402.

18. M. Lungarella and G. Metta, Beyond gazing, pointing, and reaching: A survey of developmental robotics, in *Proc. 3rd Int. Workshop Epigenetic Robotics* (*EPIROB '03*) eds. C. G. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov and C. Balkenius (2003), pp. 81–89.
19. Y. Mansour and S. Singh, On the complexity of policy iteration, in *Proc. 5th Ann. Conf. Uncertainty in Artificial Intelligence* (*UAI '99*) (1999) pp. 401–408.
20. A. N. Meltzoff, Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children, *Devel. Psychol.* **31** (1995) 838–850.
21. A. N. Meltzoff, Imitation and other minds: The "Like Me" hypothesis, in *Perspectives on Imitation: From Cognitive Neuroscience to Social Science*, eds. S. Hurley and N. Chater, Vol. 2, (MIT Press, Cambridge, MA, 2005), pp. 55–77.
22. A. N. Meltzoff, The "Like Me" framework for recognizing and becoming an intentional agent, *Acta Psychol.* **124** (2007) 26–43.
23. A. N. Meltzoff and M. K. Moore, Imitation of facial and manual gestures by human neonates, *Science* **198** (1977) 75–78.
24. A. N. Meltzoff and M. K. Moore, Explaining facial imitation: A theoretical model, *Early Devel. Parent.* **6** (1997) 179–192.
25. C. Nehaniv and K. Dautenhahn, The correspondence problem, in *Imitation in Animals and Artifacts* (MIT Press, 2002), pp. 41–61.
26. L. R. Rabiner, A tutorial on Hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77**(2) (1989) 257–286.
27. R. P. N. Rao and A. N. Meltzoff, Imitation learning in infants and robots: Towards probabilistic computational models, in *Proc. AISB* (2003).
28. R. P. N. Rao, A. P. Shon and A. N. Meltzoff, A Bayesian model of imitation in infants and robots, in *Imitation and Social Learning in Robots, Humans, and Animals* (Cambridge University Press, 2007).
29. A. P. Shon, D. B. Grimes, C. L. Baker and R. P. N. Rao, A probabilistic framework for model-base imitation learning, in *Proc. 26th Annual Meeting of the Cognitive Science Society* (*CogSci 2004*) (2004).
30. A. P. Shon, J. J. Storz and R. P. N. Rao, Towards a real-time Bayesian imitation for a humanoid robot, in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)* (2007).
31. D. Verma and R. P. N. Rao, Goal-based imitation as probabilistic inference over graphical models, in *Advances in Neural Information Processing Systems (NIPS),* Vol. 18, eds. Y. Weiss, B. Schölkopf and J. Platt (2006), pp. 1393–1400.
32. P. Viola and M. Jones, Robust real-time object detection, *Int. J. Comput. Vision* **57**(2) (2004) 137–154.
33. E. Visalberghy and D. Fragaszy, Do monkeys ape?, in *"Language" and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives* (Cambridge University Press, 1990), pp. 247–273.
34. A. Whiten, The imitator's representation of the imitated: Ape and child, in *The Imitative Mind: Development, Evolution, and Brain Bases*, eds. A. N. Meltzoff and W. Prinz (Cambridge University Press, 2002), pp. 98–121.
35. C. K. I. Williams and C. Rasmussen, Gaussian processes for regression, in *Advances in NIPS*, eds. D. Touretzky and M. Hasselmo, Vol. 8 (1996).
36. D. Wolpert and M. Kawato, Multiple paired forward and inverse models for motor control, *Neural Networks*, **11** (1998) 1317–1329.
37. Y. Wu, K. Toyama and T. Huang, Wide-range, person- and illumination-insensitive head orientation estimation, in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (AFGR00)* (2000), pp. 183–188.

Aaron P. Shon is a doctoral student at the University of Washington's Department of Computer Science and Engineering in the Neural Systems Group. His research interests include cognitive modeling, computational neuroscience, and biologically-inspired robotics. His awards include a National Defense Science and Engineering Graduate fellowship and a Microsoft Endowed fellowship.

Joshua J. Storz is a staff member at the University of Washington's Department of Computer Science and Engineering in the Neural Systems Group. His research interests include humanoid robotics and cognitive modeling.

Andrew N. Meltzoff is a Professor in the Department of Psychology at the University of Washington. He holds the Job and Gertrud Tamaki Endowed Chair in developmental science and is the Co-Director of the Institute for Learning & Brain Sciences. Meltzoff is the recipient of a NIH MERIT Award for outstanding research. He is a fellow in American Association for the Advancement of Science, the American Psychological Association, Association for Psychological Science, and the Norwegian Academy of Science and Letters. His research focuses on cognitive development in human children. Meltzoff is the co-author of two books, *Words Thoughts, and Theories* (MIT Press, 1997) and *The Scientist in the Crib: Minds, Brains, and How Children Learn* (Morrow Press, 1999), and co-editor of *The Imitative Mind: Development, Evolution, and Brain Bases* (Cambridge, 2002).

Rajesh P. N. Rao is an Associate Professor in the Computer Science and Engineering Department at the University of Washington, Seattle, USA, where he heads the Laboratory for Neural Systems. His research spans the areas of computational neuroscience, humanoid robotics, and brain-machine interfaces. He is the recipient of a David and Lucile Packard Fellowship, an Alfred P. Sloan Fellowship, an ONR Young Investigator Award, and an NSF Career award. Rao is the co-editor of two books: *Probabilistic Models of the Brain* (2002) and *Bayesian Brain* (2007).