Imitation Learning with Hierarchical Actions

Abram L. Friesen and **Rajesh P. N. Rao**
Department of Computer Science and Engineering
University of Washington
Seattle, WA, 98195, USA
{afriesen, rao}@cs.washington.edu

***Abstract*—Imitation is a powerful mechanism for rapidly learning new skills through observation of a mentor. Developmental studies indicate that children often perform goal-based imitation rather than mimicking a mentor's actual action trajectories. Further, imitation, and human behavior in general, appear to be based on a hierarchy of actions, with higher-level actions composed of sequences of lower-level actions. In this paper, we propose a new model for goal-based imitation that exploits action hierarchies for fast learning of new skills. As in human imitation, learning relies only on sample trajectories of mentor states. Unlike apprenticeship or inverse reinforcement learning, the model does not require that mentor actions be given. We present results from a large-scale grid world task that is modeled after a puzzle box task used in developmental studies for investigating hierarchical imitation in children. We show that the proposed model rapidly learns to combine a given set of hierarchical actions to achieve the subgoals necessary to reach a desired goal state. Our results demonstrate that hierarchical imitation can yield significant speed-up in learning, especially in large state spaces, compared to learning without a mentor or without an action hierarchy.**

***Index terms* – Human learning and development, action hierarchy, implicit imitation, reinforcement learning, temporal abstraction.**

I. Introduction

Humans appear to be born with an innate ability and desire to imitate others [1]. Within hours of birth, infants can reliably imitate simple gestures and facial movements [2]. By imitating their parents, peers, and other adults, infants circumvent the curse of dimensionality inherent in large state-action spaces as well as the potentially dangerous consequences of purely trial-and-error-based learning approaches such as reinforcement learning (RL) [3]. The result is rapid and relatively safe learning of a wide range of skills tailored to fit the infant's environment and culture.

How do infants translate observed actions to imitative behavior? Developmental studies conducted in a number of laboratories indicate a progression of imitative abilities, starting from imitation of facial and body movements to more complex forms of imitation, such as imitation of actions on objects, and, eventually, imitation based on inferring the underlying goal of an observed or attempted behavior (see [4] for a review).

Of particular interest is the fact that, unless required by the task, children (and adults) are particularly good at abstracting away the particulars of observed movements, focusing instead on subgoals achieved along the way to solving a given task [5], [6]. This ability dovetails nicely with the fact that human action is hierarchically organized; higher-level actions (or "plans") are composed of a sequence of lower-level actions (which can themselves be either plans or primitive actions). There is evidence for action hierarchies from multiple fronts, e.g., neuroanatomy, neurophysiology, and behavior [7].

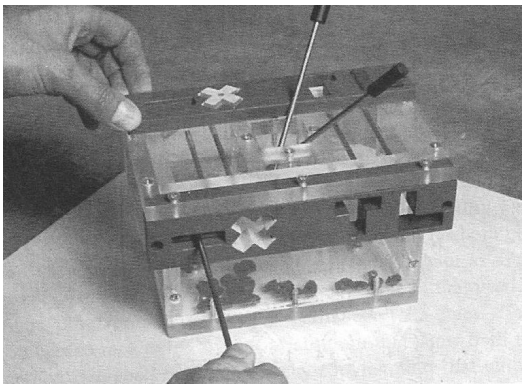

Fig. 1. The 'keyway-fruit' puzzle box task used by Whiten et. al. [8] to demonstrate imitation of hierarchical action structure in children.

Humans are also able to perceive and imitate hierarchical structure in the actions of others. For example, Whiten et al. [8] demonstrated hierarchical imitation in three-year-old children using the 'keyway-fruit' puzzle box shown in Figure 1. The box has four skewers holding the lid in place; each skewer can be removed with the same sequence of four steps, requiring a total of 16 steps to open the box. Children who witnessed solutions significantly outperformed children that did not. Additionally, children who witnessed solutions were able to acquire hierarchical rules that they later used to solve an extended version of the task, consisting of an additional skewer. Hierarchical structure was evident because the children did not copy the specific sequences of lower-level actions; rather, they inferred subgoals and used existing skills to achieve them.

In this paper, we present a new model for hierarchical imitation learning. Unlike previous models for imitation, such as apprenticeship or inverse reinforcement learning [9], the model does not assume that mentor actions (motor commands or muscle activations) are available. Rather, learning occurs solely from observation of mentor states (e.g., mentor location, limb positions, or joint angles extracted from visual or other sensory observations). We report results from experiments on a large-scale grid world task that is modeled after the "keyway-fruit" puzzle box task described above. Given only example sequences of mentor states, we show that the model rapidly

learns to combine a set of hierarchical actions ("options") to achieve the subgoals necessary to reach the desired goal state. As in the case of children and the puzzle box task, the model does not imitate the actual trajectory demonstrated by the mentor but rather utilizes the mentor's demonstrations to perform goal-based imitation. The results show significant speed-up in learning, particularly in large state spaces, compared to learning without a mentor or without an action hierarchy.

II. A Model for Hierarchical Imitation Learning

We begin with the standard reinforcement learning (RL) formalism [3]: the goal of the agent is to learn a "policy" which maps world states to appropriate actions for maximizing expected future reward. At each time step, the agent executes an action according to its policy which causes the agent to move to a new state and potentially receive a reward. The world is represented as a Markov decision process (MDP), which is a tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is a finite set of states; $\mathcal{A}$ is a set of actions; $P(s'|s,a)$ is a transition model (or "dynamics" of the environment) that specifies the probability of reaching a successor state $s' \in \mathcal{S}$ given a current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$; the reward function $R(s,a,s')$ is the reward for executing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ and transitioning to state $s' \in \mathcal{S}$; and γ is a discounting factor which specifies the weight assigned to future rewards (see equation below).

First consider the case where there is no mentor and no hierarchical actions. As the agent explores the world, it receives samples of the transition and reward functions. These allow the agent to estimate the expected reward of each state and encode it in a *value-function*, which is computed recursively using the following *Bellman equation*:

$$V(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right\},$$

where $\mathcal{A}_s$ is the set of actions available in state s, $r(s,a)$ is the *expected* reward for taking action a in state s, and $p(s'|s,a)$ is the *estimated* probability of transitioning to s' if a is executed in state s.

Next, consider introducing hierarchical actions to the formalism above. We adopt the same approach as the *options* framework for temporal abstraction in RL [10]. Options are temporally-extended actions which span multiple time steps and can execute other options, creating a hierarchical control architecture. Essentially, an option is a predefined policy over lower-level actions on a subset of the state space. Each option $o \in \mathcal{O}$ consists of three components: an initiation set, $\mathcal{I} \in \mathcal{S}$, termination probabilities $\beta(s)$ for each state s, and a policy π mapping states to actions. The option $\langle \mathcal{I}, \beta, \pi \rangle$ is available in state s if and only if $s \in \mathcal{I}$. If the option is executed, then actions are selected according to π until the option terminates stochastically in state s' with probability $\beta(s')$ [11]. Using semi-Markov decision process (SMDP) theory, the value function can be reformulated to incorporate options as follows:

$$V(s) = \max_{o \in \mathcal{O}} \left\{ r(s,o) + \sum_{s' \in \mathcal{S}} p^o_{ss'} V(s') \right\},$$

where $r(s,o)$ is the expected (discounted) reward gained from executing option o in state s, and $p^o_{ss'} = \sum_{k=1}^{\infty} p(s',k)\gamma^k$ where $p(s',k)$ is the probability that the option terminates in s' after k steps. Thus, $p^o_{ss'}$ represents a *multi-time model*, which combines the likelihood that o terminates in state s' with a measure of how delayed that termination is relative to γ [10].

Now, consider the case where a mentor is available to provide samples of state trajectories for imitation, but the mentor's actions or rewards are not available. This corresponds most closely to the case of imitation in humans and animals. Price and Boutilier proposed the framework of "implicit imitation" [12] to tackle this problem for primitive actions only. In their approach, the value function is computed using an augmented version of the Bellman equation:

$$V(s) = R_{obs}(s) + \gamma \max \left\{ \max_{a \in \mathcal{A}_{obs}} \left\{ \sum_{s' \in \mathcal{S}} p_{obs}(s'|s,a) V(s') \right\}, \sum_{s' \in \mathcal{S}} p_m(s'|s) V(s') \right\}, \quad (1)$$

where the observer's *known* reward function for state s is $R_{obs}(s)$, the observer's actions are $\mathcal{A}_{obs}$, the estimated transition probability of the observer is p_{obs}, and the observed transition probability model for the mentor is p_m. The second summation represents the expected value of duplicating the mentor's (unknown) action $\pi_m(s)$.[1] In order to guarantee accurate convergence of estimates of the mentor's dynamics, π_m must be stationary and ergodic over the subset of states that the mentor visits [12].

Finally, consider incorporating hierarchical actions (options) into imitation learning. As in implicit imitation, we assume that the state space and primitive (lowest-level) action space of the mentor and observer are the same, i.e. $\mathcal{S}_m = \mathcal{S}_{obs} = \mathcal{S}$, $\mathcal{A}_m \subseteq \mathcal{A}_{obs}$, and $p_m(s'|s,a) = p_{obs}(s'|s,a)$ for common actions a.[2] We also assume that the observer can always execute its primitive actions, allowing it to repeat any sequence of actions that the mentor executes. The observer can then estimate the value function from observations of a mentor's trajectory, even when both the observer and mentor execute possibly different options, using the *augmented Bellman equation with options*:

$$V(s) = \max \left\{ r(s) + \max_{o \in \mathcal{O}_{obs}} \left\{ \sum_{s' \in \mathcal{S}} p^o_{ss'} V(s') \right\}, R_m(s) + \gamma \sum_{s' \in \mathcal{S}} p_m(s'|s) V(s') \right\}, \quad (2)$$

where $r(s)$ is the expected reward of acting in state s, $\mathcal{O}_{obs}$ is the set of options available to the observer, $R_m(s)$ is the reward function that the observer assigns to the mentor, and $p_m(s'|s)$ is the observed transition probability for the mentor.

[1] The outer $\max$ term becomes redundant when the dynamics estimates converge because the mentor's action is identical to one of the observer's actions.

[2] These assumptions can be relaxed using appropriate mappings between the observer's states/actions and the mentor's states/actions.

Equation (2) extends the augmented Bellman equation (1) to the options case, where the second term in the outer $\max$ still denotes the expected value of duplicating the mentor's action. To guarantee convergence of the value function, $R_m(s)$ must not overestimate the observer's actual reward function and the observer must be able to duplicate the mentor's actions.[3] As above, π_m must be stationary and ergodic.

In our experiments we assume a known reward function $R_m(s) = R_{obs}(s)$ (the same assumption that [12] makes); however, this reward function could also be estimated. One particularly interesting estimation method would be to use *intrinsic reward*. As the mentor traverses the state space, it triggers salient stimuli (changes in light or sound intensity, for example), which generate intrinsic reward for the observer [13], [14]. This would allow the observer to estimate an intrinsic value function that it could use for exploration, similar to the intrinsic reward mechanism that has been postulated to exist in humans and animals [15], [16].

When learning to use options in a new task, an option's value is determined by the reward it achieves and the value of the states in which it terminates, analogous to primitive actions in standard RL. With only primitive actions, there always exists some sequence of actions that terminates in states the mentor has visited, allowing the agent to utilize information from the mentor. In imitation with options, in addition to being able to execute primitive actions, we have two scenarios: if an option terminates in states that the mentor has visited, the option's value can be updated using standard option learning methods. On the other hand, options that do not terminate on the mentor's trajectory may not benefit from observations of the mentor, however, neither will imitation be detrimental.

Additionally, it has been shown that options that take the agent to key states, such as bottlenecks or hard-to-reach areas [17], can be very useful in certain domains (e.g., a doorway in the four-rooms domain [10] or a light-switch in the playroom domain [13])). Thus, if the mentor is using a near-optimal policy for the task or solving a similar task, it is likely that the mentor will also visit these key states, ensuring an overlap between states visited by useful options and a mentor's trajectory. In the case of no overlap, exploration using primitive actions allows the agent to propagate value outward from the mentor's trajectory and spread it to a state in which an option terminates.

III. Experiments and Results

We first describe the experimental domain used to illustrate the performance of the model. We then present results from two experiments, one corresponding to the case where an agent has a large library of options from which to choose (more options than necessary to solve the task) and another where the agent has some, but not all, of the options required to complete the task. Both cases serve to illustrate goal-based imitation because the agent and the mentor are given different options but with overlapping subgoals. Finally, we investigate the issue of scalability and study the performance of the model as a function of the size of the state space.

[3]This guarantees that the outer $\max$ becomes redundant as before.

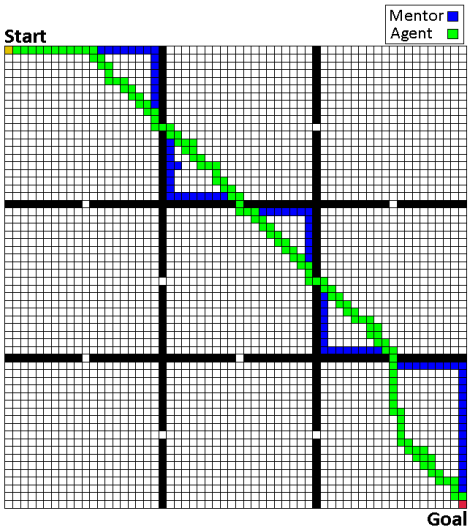

Fig. 2. Multirooms gridworld domain with 60x60 states and 3x3 rooms showing the mentor's trajectory and the imitation agent's learned trajectory through the state space.

Experimental Paradigm

Our experiments utilize a multirooms gridworld domain intended to model aspects of the keyway-fruit puzzle box task used by Whiten et al. [8] to study hierarchical imitation in children. The multirooms domain can also be viewed as an extension of the four-rooms domain previously used in hierarchical RL [10]. As shown in Figure 2, the domain consists of a large gridworld separated into rooms which are connected by doorways. The agent starts in the top-left corner and attempts to find the goal in the bottom-right corner. Once the agent reaches the goal, the episode terminates and the agent is reset to the start state.

The multirooms domain shares several similarities with the keyway-fruit puzzle box task in Figure 1. The puzzle box task can be regarded as a four dimensional task where each skewer is a separate dimension and each action on a skewer is a step along that dimension. Since there are four steps to remove each skewer, the task requires four steps in each dimension, where "steps" are complex sequences of primitive actions that achieve a specific subgoal. Similarly, the multirooms gridworld can be regarded as a two dimensional task where doorways are subgoals and the agent must choose the appropriate options to achieve the subgoals in order to reach the final goal state.

The options in the multirooms domain are composed of sequences of four primitive actions, each of which move the agent one step in one of the four compass directions. However, the environment is stochastic and, 10% of the time, an action fails, causing the agent to move in a random direction. If an action would cause the agent to move into a wall, it instead remains in its current state.

The agent's repertoire of options includes: *primitive options*, which execute a single primitive action and terminate imme-

diately, and *doorway options*, each of which takes the agent to a particular doorway when in a particular room, i.e. for each doorway in each room, a doorway option exists that takes the agent to that doorway. All options used in our experiments have pre-defined policies; the challenge is for the agent to learn which options to use to reach the goal state.

The multirooms gridworld is quite versatile and can be used to model any general hierarchical task where the effects of actions are local and where specific states (subgoals) must be reached to proceed to the next stage. For example, the multirooms gridworld is comparable to the playroom domain [13]: different rooms correspond to different states of the objects in a playroom and the doorways represent the states that trigger those objects. If more connections were added between rooms, the domains would become identical.

While the basic gridworld domain is not complicated, separating it into multiple rooms with single-state connectors (doorways) greatly increases its complexity. The task can be made more difficult by increasing the number of states, increasing the number of rooms, closing doorways to make the environment more maze-like, adding options which do not benefit the agent, and moving the position of the goal to a location that can only be reached by executing primitive options.

To illustrate the general case, we give the agent and mentor different options for reaching the doorways from within a room: the mentor travels along the walls while the agent travels directly towards the doorway. This is analogous to a child using different actions to achieve the same goal or subgoal as an adult mentor [6].

Implementation Details

The hierarchical imitation model was implemented with gamma equal to 0.99. The agent receives a reward of 1.0 for reaching the goal state and a reward of -0.01 for each step it takes. The mentor executes an optimal policy (with respect to its given options) which takes it along the diagonal from the top-left state to the bottom-right state along the walls of the rooms, as shown in Figure 2. Option values are learned according to the augmented Bellman equation (2). Value is propagated through the state space using improved prioritized sweeping [18], an approximation which enables the agent to only update the values of "important" states, with a fixed number of backups generally equal to the number of options needed to reach the goal.

We use ϵ-greedy exploration, meaning that the agent acts greedily (chooses the highest value action) with probability $(1-\epsilon)$ and acts randomly otherwise, where ϵ begins at 0.75 and starts to decay to a minimum of 0.01 after the agent first finds the goal. The agent thus initially acts mostly randomly in order to explore the environment and gracefully transitions to exploiting its learned knowledge.

We note that convergence to the optimal value function is guaranteed both when using the mentor and when not, as ϵ never decays below 0.01. The benefit of imitation is in considerably speeding up learning of the optimal policy, as demonstrated by the experimental results below.

Experiment 1: Hierarchical Imitation with a Large Library of Options

The first experiment explores the case where the agent has a large library of options from which to choose, with some options not relevant to the task at hand. This represents a common scenario in imitation learning where the agent has sufficient skills and capabilities but also many extraneous skills which are not helpful for the task at hand and whose use might actually worsen performance. The agent must discover which options are beneficial using subgoals inferred from the mentor. This is similar to the puzzle box task in that the children ignore the low-level skewer manipulation actions performed by the mentors and instead use their own learned skills to achieve the same subgoals and solve the task.

We use a large multirooms gridworld with 6400 (80×80) states and 100 (10×10) rooms. Using more rooms increases the complexity of the task because options become less effective in moving the agent closer to the goal. To further increase the complexity of the task and make it more realistic, we add eight extraneous options to each room, which take the agent to or near a corner. We will refer to these eight options as the *corner options*. These are in addition to the already mentioned doorway and primitive options. Thus, the optimal solution is to take doorway options to the bottom-right room and then take the corner option to the goal state.

Figure 3 plots the number of goals achieved in the last 5000 steps as a function of the total number of steps taken by an agent, smoothed by averaging over 5 runs of the agent through the world. The optimal path in the presence of noise is approximately 190 steps in this domain. The results show that the imitating agent significantly outperforms agents using only options or only primitive actions. The imitating agent both improves its policy at a faster rate and converges to a much better policy, even though all agents use identical exploration policies. Given enough time and exploration, a non-imitating agent will eventually converge to the optimal policy, but the bias introduced by imitation focuses the imitating agent's exploration on more useful options, allowing it to rapidly learn a better policy. Note that even though it was executing different options than the mentor, the imitating agent was still able to significantly benefit from observations of the mentor.

Experiment 2: Insufficient Options Task

Our second experiment illustrates the case where an agent has most of the skills needed for a task but is missing the final skill to take it to the goal state. Again, we use a large multirooms gridworld, this time with 1600 (40×40) states and 100 (10×10) rooms. Only the primitive and doorway options are used (no corner options). We use a smaller domain because the removal of the corner options makes this a very difficult task for the non-imitating agent. This is because the agent explores semi-randomly with ϵ-greedy exploration. With only the primitive and doorway options available, approximately half of the randomly chosen options will take the agent to a doorway (and away from the goal if it is in the goal room). Thus, to solve this task without corner options, the non-imitating agent

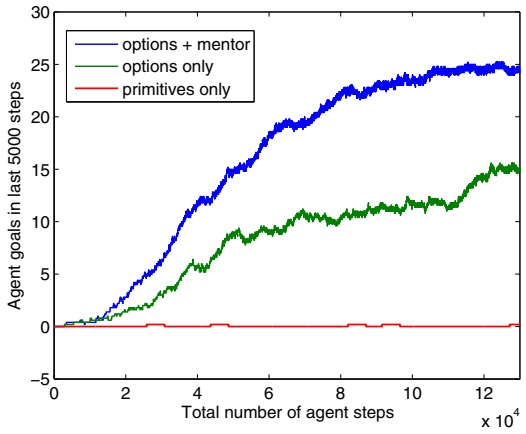

Fig. 3. **Hierarchical Imitation with a Large Library of Options.** The plots show the number of goals achieved by the agent in the last 5000 steps as a function of the total number of steps taken, averaged over 5 runs through the world. Hierarchical imitation with options yields significant gains in performance compared to using options or primitive actions only. Note that the optimal average reward is approximately 26.3 in this domain.

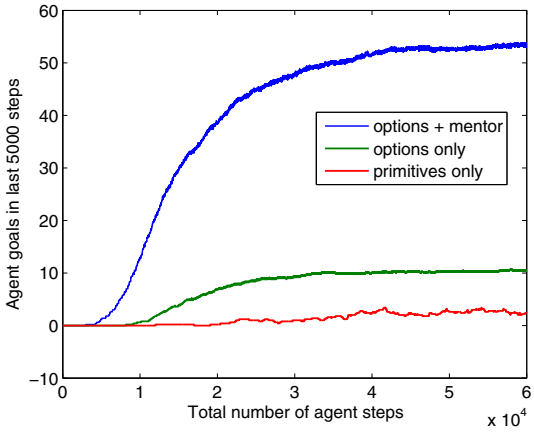

Fig. 4. **Learning with Insufficient Options.** The plots show the number of goals achieved by the agent in the last 5000 steps (averaged over 5 runs through the world) as a function of the total number of agent steps. The imitating agent demonstrates an even greater gain in performance than in experiment 1 in terms of both learning accuracy and speed, compared to the non-imitating agents. In this domain, the optimal average reward is approximately 53.2.

must successively choose primitive options in the final room in order to reach the goal, a very unlikely occurrence.

To successfully solve the task, the agent must be able to learn a new skill in the final room to take it to the goal state and realize that the doorway options are not beneficial. Note that the agent does not actually learn a new option when it reaches the goal. Instead, it learns a sequence of primitive actions which take it to the goal. The model could potentially be extended to allow the agent to create and learn new options when faced with such situations. We leave such an extension to future work.

The results, shown in Figure 4, demonstrate that, with the help of a mentor, the agent is able to quickly and reliably reach the goal even though it does not have the necessary

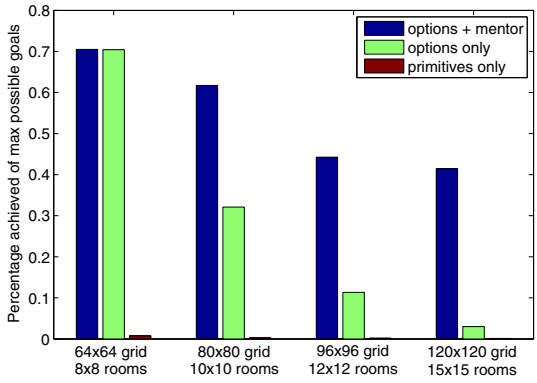

Fig. 5. **Scaling Up to Larger Worlds.** The bar graph shows the total number of goals received in 130,000 steps as a percentage of the maximum achievable number of goals (if the agent acted optimally). The benefits of imitation become increasingly apparent as the size of the state space is increased, with greater payoffs for larger state spaces.

set of options for this task. The non-imitating agent with the same options requires a much longer time to reach the goal. In four out of five sample runs, the non-imitating agent was unable to find the goal. Increasing the difficulty of the task (by increasing the size of the gridworld, for example) causes the non-imitating agent to consistently fail to find the goal within a reasonable number of steps.

Experiment 3: Scaling Up

Our final experiment focuses on how the performance gains due to hierarchical imitation scale up with the size of the environment. We computed the total number of goals that the agent was able to achieve as a percentage of the maximum number of goals that the agent could achieve if it were to execute an optimal policy for a fixed number of steps (130,000). Note that an agent cannot perform at the level of the optimal policy because it must first explore the environment and find the goal before it can begin learning a policy. The results, shown in Figure 5, indicate that in domains with smaller numbers of rooms, the task is relatively easy to solve and imitation provides little benefit over standard reinforcement learning with options. However, as the domain size increases, the benefit of imitation becomes increasingly apparent, with dramatic performance gains as one goes from a 64×64 world to 80×80, 96×96, and 120×120 worlds.

IV. Conclusions

This paper presents a new model for hierarchical goal-based imitation. The model exploits action hierarchies in the form of options to quickly learn new skills given only the mentor's states (without the mentor's actions or rewards). The model leverages advances in two parallel lines of research: reinforcement learning based on options [10] and implicit imitation [12]. We propose that the reward function could be learned via intrinsic reward generated from salient attributes of the mentor's states. We illustrate the performance of the model on a multirooms domain that captures the essential aspects

of a puzzle box task used to study hierarchical imitation in children. Our results show that the model can learn from observing a mentor even when the mentor uses very different actions (options) from its own by implicitly inferring subgoals from the mentor's trajectory and appropriately combining its options. The results demonstrate that the model yields significant performance gains compared to traditional reinforcement learning based on options; furthermore, these gains become more substantial as the problem size increases. The model's ability to scale with task complexity confirms the long-held intuition that imitation learning is an effective way to counter the curse of dimensionality present in large state-action spaces.

Learning by imitation has received considerable attention in both cognitive modeling and robotics. In robotics, the traditional approach to imitation has been trajectory following [19], [20], where the agent attempts to directly imitate a mentor's state trajectory rather than imitating based on inferred subgoals. More recent variations have combined dimensionality reduction with probabilistic techniques to achieve trajectory imitation [21]–[23]. Approaches to goal-based imitation have focused on using graphical models [24], [25] and apprenticeship learning via inverse reinforcement learning [9]. In the former, graphical models are used for learning and representing the transition dynamics of an agent, and then employed to probabilistically infer both goals and actions for imitation. However, rewards are not considered during learning. Inverse reinforcement learning, on the other hand, attempts to learn an unknown reward function by observing an expert. The approach however is not biologically realistic as it assumes full knowledge of the expert's actions, in addition to the expert's states. None of the above models incorporate hierarchical actions for imitation.

Several important questions remain to be addressed. The proposed model allows a set of previously learned hierarchical actions or options to be combined to achieve a desired goal state. How can a collection of such hierarchical actions be learned from experience? One approach is to use bottleneck states (such as the doorways in the multirooms domain) to segment recent action sequences and cache the segments as options [17]. The imitation paradigm offers a potentially faster way to learn options through observation of various mentors. We intend to investigate such methods for option learning in future work. Another area worth investigating is combining hierarchical state abstraction with the action hierarchy currently used in the model. This may facilitate transfer of knowledge between different tasks. Finally, we hope to extend our model to continuous domains such as humanoid robotics where the ability of the model to scale to large state-action spaces may pay rich dividends.

ACKNOWLEDGMENT

This work is supported by a grant from the Office of Naval Research Cognitive Science program and a Packard Fellowship to RPNR.

REFERENCES

[1] A. N. Meltzoff, "Imitation and other minds: The "like me" hypothesis," in *Perspectives on Imitation: From Neuroscience to Social Science*, S. Hurley and N. Chater, Eds. MIT Press, 2005, pp. 55–77.

[2] A. N. Meltzoff and M. K. Moore, "Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms," *Developmental Psychology*, vol. 25, pp. 954–962, 1989.

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

[4] 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, 2004, pp. 217–247.

[5] M. Gattis, H. Bekkering, and A. Wohlschläger, "Goal-directed imitation," in *The Imitative Mind: Development, Evolution, and Brain Bases*, A. N. Meltzoff and W. Prinz, Eds. Cambridge University Press, 2002.

[6] A. N. Meltzoff, "The "like me" framework for recognizing and becoming an intentional agent," *Acta Psychologica*, p. 2643, 2007.

[7] M. M. Botvinick, "Hierarchical models of behavior and prefrontal function," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 201–208, 2008.

[8] A. Whiten, E. Flynn, K. Brown, and T. Lee, "Imitation of hierarchical action structure by young children," *Developmental Science*, vol. 9, no. 6, pp. 574–582, 2006.

[9] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM Press, 2004.

[10] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, pp. 181–211, 1999.

[11] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dynamic Systems*, vol. 13, pp. 341–379, 2003.

[12] B. Price and C. Boutilier, "Implicit imitation in multiagent reinforcement learning," in *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 325–334.

[13] A. G. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of International Conference on Development and Learning*, 2004, pp. 112–119.

[14] S. Singh, A. G. Barto, and N. Chentanez, "Intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems*, 2005.

[15] R. W. White, "Motivation reconsidered: The concept of competence," *Psychological Review*, vol. 66, pp. 297–333, 1959.

[16] P. Dayan and B. W. Balleine, "Reward, motivation, and reinforcement learning," *Neuron*, vol. 36, pp. 285–298, 2002.

[17] M. Stolle and D. Precup, "Learning options in reinforcement learning," in *SARA*, ser. Lecture Notes in Computer Science, S. Koenig and R. C. Holte, Eds., vol. 2371. Springer, 2002, pp. 212–223.

[18] H. B. McMahan and G. J. Gordon, "Fast exact planning in Markov decision processes," in *Proc. of the 15th International Conference on Automated Planning and Scheduling (ICAPS-05)*, 2005.

[19] M. Y. Kuniyoshi and H. Inoue, "Learning by watching: Extracting reusable task knowledge from visual observation of human performance," in *IEEE Transaction on Robotics and Automation*, vol. 10, no. 6, 1994, pp. 799–822.

[20] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[21] D. B. Grimes, R. Chalodhorn, and R. P. N. Rao, "Dynamic imitation in a humanoid robot through nonparametric probabilistic inference," in *Proceedings of Robotics: Science and Systems*, 2006.

[22] R. Chalodhorn, D. B. Grimes, K. Grochow, and R. P. N. Rao, "Learning to walk through imitation," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007.

[23] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2007, pp. 255–262.

[24] D. Verma and R. P. N. Rao, "Goal-based imitation as probabilistic inference over graphical models," in *Advances in Neural Information Processing Systems*, 2006.

[25] D. Verma and R. P. N. Rao, "Imitation learning using graphical models," in *Proceedings of the 2007 European Conference on Machine Learning*, 2007, pp. 757–764.