2018 Joint IEEE International Conference on Development and Learning and
Epigenetic Robotics (ICDL-EpiRob)
Tokyo, Japan, September 16-20, 2018

Modeling the Development of Infant Imitation using Inverse Reinforcement Learning

Ahmet E. Tekden[1,2], Emre Ugur[1], Yukie Nagai[2] and Erhan Oztop[3]
[1]Bogazici University, [2]NICT and [3]Ozyegin University

***Abstract*— Little is known about the computational mechanisms of how imitation skills develop along with infant sensorimotor learning. In robotics, there are several well developed frameworks for imitation learning or so called learning by demonstration. Two paradigms dominate: Direct Learning (DL) and Inverse Reinforcement Learning (IRL). The former is a simple mechanism where the observed state and action pairs are associated to construct a copy of the action policy of the demonstrator. In the latter, an optimality principle or reward structure is sought that would explain the observed behavior as the optimal solution governed by the optimality principle or the reward function found. In this study, we explore the plausibility of whether some form of IRL mechanism in infants can facilitate imitation learning and understanding of others' behaviours. We propose that infants project the events taking place in the environment into their internal representations through a set of features that evolve during development. We implement this idea on a grid world environment, which can be considered as a simple model for reaching with obstacle avoidance. The observing infant has to imitate the demonstrator's reaching behavior through IRL by using various set of features that correspond to different stages of development. Our simulation results indicate that the U-shape performance change during imitation development observed in infants can be reproduced with the proposed model.**

I. INTRODUCTION

Imitation is a dynamic process that evolves along with sensorimotor development [1]. Action experience changes the perception of others' actions [2] and improves the goal-prediction capacity in action observation in infants [3]. Several innate sensorimotor programs for triggering social bonding and facilitating imitation development, such as face detection and basic facial mimicry may be present in early infancy. However, it is not likely that a full fledged imitation or goal extraction capacity exists soon after birth. Goal emulation, a form of imitation characterized by the replication of the observed end effect [4], starts after early infancy. However, infants become skilled at imitating unseen movements only after 12 months of age [5]. Infants' means of imitation changes over time. While younger infants are more inclined in achieving the goal of a demonstrated action, older infants tend to exactly imitate (and in later stages over-imitate) the observed target action sequence even if those actions are not physically related to the goal [6]. As a possible hypothesis, we consider the idea that at some point in development infants start to use a rudimentary inverse mechanism to explain observed actions, i.e. they try to assign reasons or goals to observed actions. This rudimentary system then evolves into a complex adult inverse mechanism [7] where several levels of goals can be instantiated by the observer. A computational mechanism that may capture such an inverse computation is generally called Inverse Reinforcement Learning (IRL) [8] which we adopt in this study.

The capacity of infants to reason about their environment is constrained by what features they can perceive/compute from their environment. As such, their understanding can be envisioned as a projection of outside world into their feature representations. As development shapes how the infants perceive the world and form representations [9], [10] for expanding their manipulation capability, it also enriches the feature space where actions are mapped to and reasoned upon. This is analogous to mirror neurons [11], [12] in that as one becomes skilled at performing a particular task, mirror neurons develop that can recognize the actions used to perform the task [12], [13]. Many IRL methods try to recover the reward function with the assumption that the state and action costs are a function of a predetermined set of features. In this study, we explore whether these features may be taken akin to some representation that are formed by infant sensorimotor system, and more importantly, whether developmental changes in these representations can explain infant imitation development.

To conduct this modeling study we adopt a tractable grid-world like environment that can be considered as a simple model of reaching in two dimensions, and focus our attention on a simple set of features that may correspond to different stages of infant development. For example, many basic reflexes such as rooting and sucking are elicited by tactile stimulation in human infants [14] in the first month of life. Therefore, it may be argued that contingency of tactile sensation and vision of touching an object may make hand-object contact a salient visual stimuli early in infancy. During the sensorimotor stage, infants repeat actions centered on their own body in primary circular reactions period (months 1-4), and then get more involved with the objects around in the secondary circular reactions period (months 4-8) [15]. Based on this data, perception of the agent is gradually improved by first introduction of tactile perception and the visual representation of it, then egocentric hand position, and finally object centered hand position to the set of features for acting and understanding others' actions. In the same vein, in

This research was partially funded by Bogazici Resarch Fund (BAP) Startup project no 12022, by Slovenia/ARRS - Turkey/TUBITAK bilateral collaboration grant (ARRS Project no: BI-TR/16-18-001; TUBITAK Project no:215E271), and by JST CREST Cognitive Mirroring (Grant Number: JPMJCR16E2), Japan.

our experiments, we have our agent use progressively more complex set of features in inferring the underlying reward function of the observed actions. While this leads to an increased imitation performance in a fixed environment, it causes a U-shaped performance curve in novel environments, as the feature complexity increases.

II. Method

A. Reinforcement Learning

Reinforcement Learning (RL) is a computational method for an agent to learn to act to maximize long term rewards in a possibly stochastic environment (see [16] for an extensive introduction). RL builds on the formal framework of Markov Decision Processes (MDP) and the agent is envisioned in an interaction cycle with the environment, where the agent acts, and in return, obtains scalar reward signals. The state of the environment and the agent is captured in a representation called *state* that represents the past interaction of the agent with the environment. In this study, we consider discrete and deterministic MDPs, accordingly an MDP is defined with (S, A, P, R) where S is the set of states, A is the set of actions, and P is the transition function that maps $(state, action)$ pairs to *states*, and finally R is the reward function that maps given $(state, action)$ pair to an immediate reward (or cost) value. In MDPs the goal is to find an optimal policy to maximize the total reward in the long run, which can be solved by Dynamic Programming or Monte Carlo sampling based methods (e.g. see [16] Chapter 4). RL also aims to solve the same problem, but in this case the MDP can be partially specified if the agent is allowed to explore. In particular, RL methods that can deal with MDPs with unknown P or R are abundant. In this study, we are not really concerned with how the agent obtains an optimal behavior, but rather with the underlying reward function leading to the observed optimal behavior. So when we need to generate optimal behavior samples for our agent, we adopt a simple MDP solution method, namely Value Iteration [16] which can easily find optimal policies for our simple grid-world like environment. The simulated optimal behavior samples are then used by an IRL algorithm to infer the reward, which we describe next in the context of imitation learning.

B. Inverse Reinforcement Learning

In Imitation Learning or Learning by Demonstration [17], the goal is to extract a representation of the demonstrated action that would allow an agent to generate actions to replicate the observed action. An easy way to do is direct imitation (DI) where the actions of the demonstrator are associated with the state of the agent and the environment. DI is applicable for agents with similar kinematics and dynamics on tasks that do not require generalization over context. In other cases, a more suitable approach is to see the behavior of the demonstrator as the result of an optimal control or planning problem, and try to recover the objective function of the agent. Finding the objective function is called the problem of Inverse Optimal Control or IRL. Most IRL algorithms try to infer the missing immediate reward function given the optimal behavior samples. So input to an IRL algorithm is an incomplete MDP, $MDP/R = (S, A, P, \_)$ and a set of optimal behavior demonstrations, and the output is R, the immediate reward function. IRL is an ill-conditioned inverse problem, so the solution space must be constrained. Among other alternatives (e.g. [18], [19]), we have adopted Maximum Entropy IRL [20] that postulates that trajectories with equal total reward appear with equal probability; and those with higher reward appear exponentially more often. This makes the method robust against noise and imperfect demonstration.

C. Imitation Setup

In most IRL settings the reward function is assumed to be a linear or non-linear function of a set of predefined features of the task and environment. To use IRL as a tool to explore our hypothesis about imitation development, we analyze the effects of different feature sets on imitation capacity obtained through IRL. In particular, we propose connections between specific feature sets and developmental stages of human infants.

To create an imitation scenario we use RL together with IRL: the former allows us to provide expert demonstration samples upon which IRL can be run. Thus, we first compute the optimal state value function, $V(s)$, and then use it to derive the optimal policy for our task. In turn, using the optimal policy, we generate optimal action trajectories for a set of task configurations. These trajectories serve as the demonstrations for the observer who uses IRL to imitate the observed act. In this study, we limit the task configurations so that the effects of different perceptual capacities can be observed. Thus, accordingly, the observer is not given the opportunity to observe all possible optimal trajectories. To make sense of the demonstrations, the observer uses a set of features, $F(s)$, dictated by its perception capacity, to infer the reward function of the demonstrator. We postulate that these features are subject to development and hence could potentially predict the observed properties of infant imitation development.

To asses the imitation performance of the observer, the extracted reward function, $r'(F)$ is fed to RL and converted into a state value function for the observer $V'(s)$. $V'(s)$ can then be used to generate optimal actions, which would be only a faithful imitation only if the observer has used a sufficiently rich or an appropriate feature set during IRL computation. To assess the imitation capacity quantitatively we do not directly compare r to r' or V to V' as different reward functions (and hence value functions) may lead to the same optimal policy. Instead, we compare the optimal policy inferred by the observer with the demonstrator's policy. For this, we define a test configuration as a set of possible initial states (i.e. target and hand locations) with which action rollouts (i.e. task executions) can be generated. In the current experiments, two different test configurations have been used: one replicates the observed configuration, and the other defines a novel environment based on the given

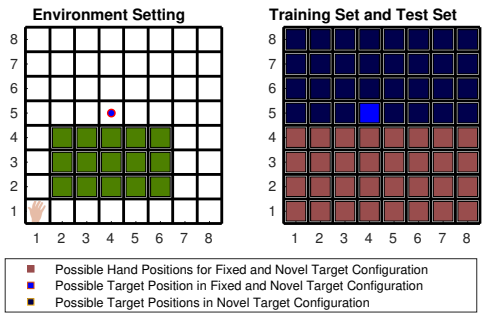

Fig. 1: Environment setup. Left: Hand and target position, and the rough (3x5) region at a time-step. Right: The hand and the target are initialized at a point at the bottom and top regions, respectively.

configuration (e.g. target appears at different locations from the observed one). Using these test configurations, rollouts are run according to the optimal policy based on the IRL extracted rewards (r'), and the total rewards (R') collected along these trajectories are computed. To assess the imitation performance, the total reward computation is not based on the estimated reward function, r'; but, on the original reward function r. The total reward R' is compared against the true optimal total reward R that would be collected by the true optimal policy of the demonstrator. A perfect imitation by the observer would necessarily imply that for each rollout, the true total reward R would be equal to R', the one collected based on the inferred reward function.

III. Experimental Setup

The hand of the agent (or simply the agent) moves in a 8x8 grid world that includes a target in one of its grids. The hand moves with left, right, up and down actions in a deterministic setting where two kinds of grids are defined: normal and rough. Movements in rough grids incur higher costs than moving in normal grids. In the imitation scenario, the observing agent is provided a number of demonstrations where the demonstrator starts from various grid locations and reaches the target through the optimal trajectories. The observer then tries to infer the reward structure of the demonstrator and imitate the observed behavior via IRL. The aim of our experiments is to investigate how the imitation performance is affected from the way the agent models the reward structure of the demonstrator.

We use IRL toolkit[1] of Sergey Levine for environment modeling and running RL and IRL algorithms. In particular, Value Iteration and Maximum Entropy IRL [20] implementations are used. The basic grid-world setup is modified so that the agent receives -10 reward in the normal grids and -30 reward in the rough grids. Selected reward values enables agent to act differently depending on its starting cell. A too little or a too large difference between the two reward values would make the agent ignore the rough grids or always avoid the rough grids, thus forming a poor data set for IRL. An example snapshot from this environment is shown in Fig. 1 where the hand is in grid position (1,1), the target is in grid position (4,5), and the rough grids are shown with green color in the figure on the left. In the current experiments, the hand and the target are allowed to be initialized to any position in the lower and the upper part of the environment, respectively. The shape of the rough grid area is fixed and always appears centered just below the target object as shown in Fig. 1. Max Entropy IRL models the reward function as weighed summation of (state) feature vectors and estimates the weights. In our experiments we define the following features that may correspond to certain developmental stages of infants related to target object position (tar_x, tar_y) and current agent hand position $(curr_x, curr_y)$.

- F_1: Indicates whether the agent is on the target or not.

$$F_1 = \begin{cases} 1, & \text{if } ||tar_x - curr_x, tar_y - curr_y|| = 0 \\ 0, & \text{otherwise} \end{cases}$$

- F_2: The vector encoding the location of the hand of the agent which is obtained by flattening the 8×8 egocentric response matrix $[M^E_{ij}]$ defined by

$$M^E_{ij} = e^{-||i-curr_x, j-curr_y||^2/0.1}$$

- $F_{1,2}$: Concatenation of F_1 and F_2
- F_3: The vector encoding the location of the agent with respect to the target, which is obtained by flattening the 8×8 allocentric response matrix $[M^A_{ij}]$ defined by

$$M^A_{ij} = e^{-||(i-4)-(tar_x-curr_x),(j-5)-(tar_y-curr_y)||^2/0.1}$$

- $F_{1,2,3}$: Concatenation of F_1, F_2 and F_3.

The composite features defined above captures the notion that the perception system of a developing agent becomes more refined through development, by starting off with the mere visual perception of hand-object contact (F_1), then by the addition of egocentric hand position ($F_{1,2}$), and finally by the incorporation of object centered hand position information ($F_{1,2,3}$). Note that an infant's ability to use vision to guide his/her hand does not necessarily mean that he/she has the capacity to represent an observed hand in visual coordinates. First hand-object contact may become salient for the infant (F_1); then position of a moving hand can be encoded in visual coordinates (F_2), and finally the infant may learn to compute relative distances with respect to a salient object (F_3). This progression is plausible from a computational perspective and can be supported by infant literature that infants make use of vision later in development even for their own actions [21]. If one needs a speculation as to possible timing of these feature developments one can suggest that $F_{1,2}$ may correspond to 2-3 months olds, whereas $F_{1,2,3}$ may correspond to 6 months or older.

To generate demonstrator actions, a set of consistent hand trajectories bringing the hand to the target object must be computed. We used value iteration algorithm to compute an optimal policy that can be used to generate actions to reach the target with minimal cost from any given initial hand position. For optimal policy computation, the grid position of the agent was taken as the *state* and a move into a normal grid

[1] https://graphics.stanford.edu/projects/gpirl/

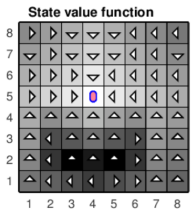

Fig. 2: Optimal policy found by RL. Demonstrations are extracted from this optimal policy and provided to the agent.

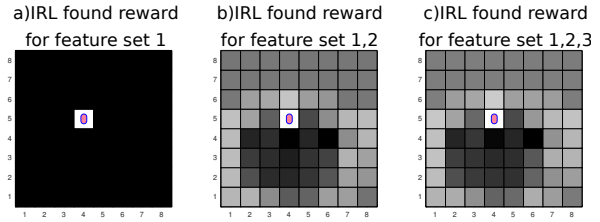

Fig. 3: Reward function maps obtained from different features sets F_1, $F_{1,2}$and $F_{1,2,3}$.

is assumed to incur a cost of 10 whereas a move into a rough grid is assumed to incur a cost of 30. Optimal policies for different target position were generated by sampling target positions (n=100) from the region shown in Fig. 1, and the the corresponding optimal policy for each target position was found. The value iteration algorithm works by using the state and reward function; so, unless the state definition is modified to account for target position, different optimal policies will be found for different target positions in general. As an example, Fig. 2 illustrates the optimal policy when the target is at grid position (4,5). Arrows in the grids indicate (one of the) optimal actions assuming the hand is on that grid; whereas the intensity of the grids indicates the corresponding state values. As can be seen, in general the demonstrator avoids the rough grids below the target. When put inside the rough region, the demonstrator directly moves within the rough region towards the target if the direct path is short; otherwise, it exits the region (for example from (2,2), (2,3), (6,2), and (6,3)), and detours the rough region while reaching the target.

IV. Results

A. Rewards and policies in fixed target configuration

The observing agent was given 100 demonstrations generated from the optimal policy found as described above. From these demonstrations, the observing agent (modeling the imitating infant) estimated the reward function by using the maximum entropy IRL [22] method. The reward function

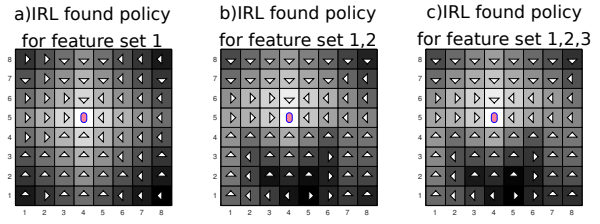

Fig. 4: Optimal policies computed from the reward functions, $f(F_1)$, $f(F_{1,2})$, and $f(F_{1,2,3})$, respectively.

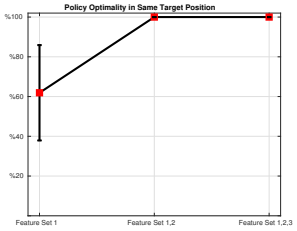

Fig. 5: Performance of agents in imitating reaching actions in the original environment where the demonstration took place. The performance quickly increased with reward function that used more complex feature sets.

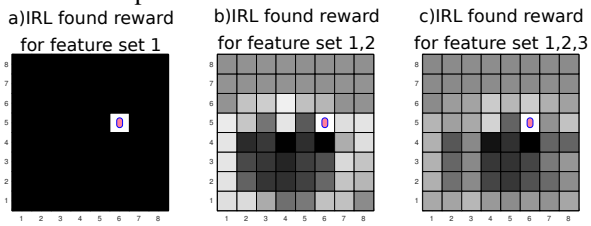

Fig. 6: Reward function maps for obtained from different features sets, F_1, $F_{1,2}$ and $F_{1,2,3}$.

was parameterized as a linear function of state features. We repeated IRL process by using progressively more complex feature sets: F_1, $F_{1,2}$, and $F_{1,2,3}$ to model observing agents with differing perceptual capacities. The resulting reward functions with these features are shown in Fig. 3, which we call reward function maps. In these maps, the intensity of each grid indicates the immediate reward value of the hand being at the corresponding grid. As observed in (a), when F_1 is used as the feature to encode the observed state, the hand being in the same grid with the target was found to be rewarding and other grids non-rewarding. F_1 could only represent the information whether the hand is on the target or not, and therefore it neglected the rough regions in computing rewards. On the contrary, because the hand position was encoded in $F_{1,2}$ and $F_{1,2,3}$in (b) and (c), low rewards were computed for situations when the hand is moving in the rough regions. As direct paths to the target through the rough regions were rare; such paths were only observed if the agent already started within the rough regions (as also shown in Fig. 4). Therefore, being on normal grids were found to be more rewarding compared to being on rough grids.

Next, we investigated imitation performance based on the feature sets considered above. To this end, we applied another round of value iteration to find the optimal policy and state-value functions using the IRL estimated rewards. The obtained policies and state-value functions are provided

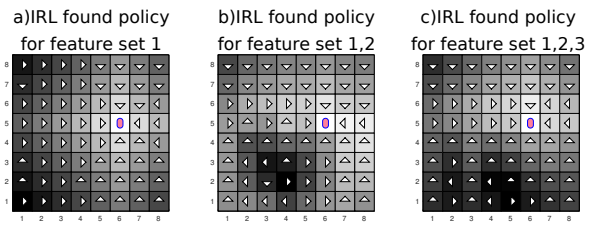

Fig. 7: Optimal policies when applied for new environment for the feature sets $f(F_1)$, $f(F_{1,2})$ and $f(F_{1,2,3})$.

in Fig. 4. As shown in (a), the optimal policy that used $f(F_1)$ finds the shortest, and thus more costly path to the target neglecting the rough region. Other policies generally avoid the rough region because being in that region was found to have low reward. This is so because the features except $f(F_1)$ contain the hand position information, either in absolute or relative coordinates with respect to the target.

After examining the policies generated from the extracted reward functions based on the different features, we quantitatively compared the performances of those policies. For this, the total rewards accumulated by the inferred optimal policies were calculated. The hand position was initialized randomly 32 times, and the actions were selected based on the inferred optimal policy. The total rewards were however computed using the original reward formulation (-10 for normal and -30 for rough grids). The distribution of the ratio of total rewards obtained from inferred optimal policy and optimal policy found using RL is provided in Fig. 5. As shown, as F_1 does neglect the rough area and move the hand directly to the target, it collected negative rewards, whereas the executions that followed the $F_{1,2}$ and $F_{1,2,3}$ based policies accumulated high rewards. The observations from these experiments can be summarized as follows:

- If the observer agent can only represent whether the demonstrator's hand has contact with the target or not, and use only this information in computing reward function for the demonstrated reach action, the policy derived for its imitation leads to a suboptimal performance.
- If the observer agent can also perceive additional information such as the location of the hand, and use this in the reward function estimation, the performance of the derived policies for imitation are close to the performance of the original optimal policy, albeit the fact that nonidentical reward function estimations can be observed depending on the features used.
- Using hand information in computing rewards enables the agent to develop optimal policies when the target position is same. In other words, the agent who computed the underlying reward function of the demonstrations using the hand position, can imitate the observed action when the target position is kept same.

B. Rewards and policies in novel target position

In the second experiment, we aimed to see what would be the imitation performance when the agent is put in a novel environment. Note that here, the agent does not apply a fresh new IRL, but uses the reward function obtained for the original environment. In particular, it does not see any demonstration for the new environment. The environment in our case is simply determined by the location of the target object. At first sight, it may look unfair to expect the learned behavior to generalize to a new object location. However this is possible with suitable feature definition. Here we wish to capture an infant's over-learning a behavior and deploying this behavior in a new environment. In such a case, certain feature representation would be beneficial for the infant, albeit requiring more complex computation. As part of the development, we suggest such feature selection formation takes place, and this simple experiment may serve a simplified model for the initial parts of feature selection.

In the original environment the agent forms a reward map based on the features, but these features (to be concrete, F_1 and F_3) are more like functionals parameterized by the target object location. Thus when the reward map is instantiated in the new environment, a new map forms as the object location is different. With this in mind, we first examine the resulting reward function maps in the new environment for each feature set. Fig. 6 provides an example of reward function map that was obtained when the target placed at position (6,5) instead of its original position (4,5). As shown, the reward for (a) is the same as in our previous case: F_1 can only encode a contact with the target, therefore $f(F_1)$ gives a binary reward depending on the contact. Similarly, $f(F_{1,2})$ and $f(F_{1,2,3})$ assign a high reward when the hand is on target; but, importantly these features can also discriminate other grids by for example, producing lower rewards when the hand is in the rough area. One interesting observation is that the rewards found based on $f(F_{1,2})$ and $f(F_{1,2,3})$ for the hand positions near the target (e.g. (5,5)) are sometimes lower than those further away from the target (e.g. (4,6)). For $f(F_{1,2})$, this was an expected result: the system failed to generalize the reward function to new target locations since F_2 did not consider object position. For $f(F_{1,2,3})$, one could expect that the object-centered coding in $F_{1,2,3}$ would override the effect of the fixed target position used in the demonstrations. However, this was not completely the case; the agent partially retained the effect of absolute hand location but also showed considerable generalization for the new object positions as explained next. The IRL algorithm we used forms a reward function as a linear combination of the components of the feature vector provided to it, this observation tells us that in $f(F_{1,2,3})$ map F_2 components are also assigned non-negligible weights.

To further investigate the imitation performance, we obtained the optimal policies (by using the inferred reward functions) for the new environments with novel target positions. The obtained policies and state-value functions are presented in Fig. 7. As shown in (a), the optimal policy that used the derived reward function $f(F_1)$ again follows the shortest and more costly path ignoring the rough region. On the other hand, for the other two cases,the features assumed by the imitator and their interaction result in different policies which are not optimal. To quantitatively investigate the performance of these policies, we computed the total rewards accumulated by these policies over a collection of action executions, i.e. roll-outs. The initial target and the hand positions allowed $32 \times 32 = 1024$ roll-outs which were executed by following the extracted policies. The total accumulated rewards were computed by using the original cost formulation (-10 for normal and -30 for rough grids). The distribution of the ratio of total rewards obtained from inferred optimal policy and optimal policy found using RL is provided in Fig. 8. As can be seen in Fig. 8, different

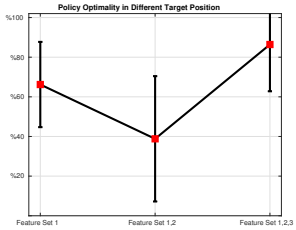

Fig. 8: Performance of agents in imitating reaching actions in novel environments. We observed a U-shaped performance change when progressively more complex feature sets are used in encoding the reward function of the observed original trajectories.

from the fixed-target case, the policies obtained by using the reward functions $f(F_{1,2})$ and $f(F_{1,2,3})$ also collected low rewards, i.e. they were not optimal. Notice that while the the policy derived from $f(F_{1,2})$ performed almost perfect in the original setting, it performed worst in the novel environment. The optimal policy derived from $f(F_{1,2,3})$ on the other hand, performed better compared to $f(F_1)$and $f(F_{1,2})$, as evident from the mean accumulated reward in the Fig. 8. Never the less, as indicated by the standard deviation around the mean, there are also suboptimal actions that lead to lower total rewards in the novel environment, which are probably due to the conflict created by the coexistence of F_2 and F_3 terms in the feature representation used in computing $f(F_{1,2,3})$.

V. CONCLUSION

In this study we explored the plausibility of an IRL based explanation of infant imitation development. In doing so, we assumed that during early development, basic RL and IRL capacity exists albeit in a rudimentary form and that the infant perceives an ongoing action in terms of a set of state features which becomes refined along with development. By inspiring from infant literature we proposed a set of features that starts with detection of hand-object contact and evolves towards a goal centered hand representation.

With this setting, our simulation experiments have shown that the use of such progressively more complex feature sets yields a U-shaped imitation performance when the infant is put into a novel environment. Interestingly, similar U-shaped learning phenomena have been reported in several learning problems that human infants face such as acquisition of the verb morphology for the English past-tense [23], and development of reaching and walking behaviors [24, p. 148]. It should be also checked whether this U-shape performance change can be obtained independent of the IRL method used. The result may give us a clue as to which IRL method is more likely to capture the IRL mechanism, if exist, of the brain.

In this study, as a first step for explaining imitation learning via IRL, we have used a fixed set of features corresponding to different developmental stages of an infant. Besides validation of this computational approach by actual human infant experiments, it would be very interesting to model the feature set formation and adaptation in a developmentally valid perspective, which we plan to address next.

REFERENCES

[1] S. S. Jones, "The development of imitation in infancy," *Philos Trans R Soc Lond B Biol Sci*, vol. 364, no. 1528, pp. 2325–35, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19620104

[2] J. A. Sommerville, A. L. Woodward, and A. Needham, "Action experience alters 3-month-old infants' perception of others' actions," *Cognition*, vol. 96, no. 1, pp. B1–B11, 2005.

[3] Y. Kanakogi and S. Itakura, "Developmental correspondence between action prediction and motor ability in early infancy," *Nature communications*, vol. 2, p. 341, 2011.

[4] S. C. Want and P. L. Harris, "How do children ape? applying concepts from the study of non-human primates to the developmental study of imitationin children," *Developmental Science*, vol. 5, no. 1, pp. 1–14, 2002.

[5] B. Elsner, "Infants imitation of goal-directed actions: The role of movements and action effects," *Acta psychologica*, vol. 124, no. 1, pp. 44–59, 2007.

[6] C.-T. Huang and T. Charman, "Gradations of emulation learning in infants imitation of actions on objects," *Journal of experimental child psychology*, vol. 92, no. 3, pp. 276–302, 2005.

[7] S. Collette, W. M. Pauli, P. Bossaerts, and J. O'Doherty, "Neural computations underlying inverse reinforcement learning in the human brain," *eLife*, vol. 6, 2017.

[8] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *in Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 2000, pp. 663–670.

[9] E. Ugur, Y. Nagai, E. Sahin, and E. Oztop, "Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 2, pp. 119–139, 2015.

[10] E. Ugur and J. Piater, "Emergent structuring of interdependent affordance learning tasks using intrinsic motivation and empirical feature selection," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 328–340, 2017.

[11] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature reviews neuroscience*, vol. 2, no. 9, p. 661, 2001.

[12] E. Oztop and M. A. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological cybernetics*, vol. 87, no. 2, pp. 116–140, 2002.

[13] C. Catmur, V. Walsh, and C. Heyes, "Sensorimotor learning configures the human mirror system," *Current biology*, vol. 17, no. 17, pp. 1527–1531, 2007.

[14] E. B. Goldstein, *The Blackwell handbook of sensation and perception*. John Wiley & Sons, 2008.

[15] J. Piaget and M. Cook, *The origins of intelligence in children*. International Universities Press New York, 1952, vol. 8, no. 5.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[17] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[18] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 1.

[19] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 729–736.

[20] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning." in *AAAI*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[21] E. Oztop, N. S. Bradley, and M. A. Arbib, "Infant grasp learning: a computational model," *Experimental Brain Research*, vol. 158, no. 4, pp. 480–503, 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15221160

[22] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning." in *AAAI*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[23] R. Brown, *A first language: The early stages.* Harvard U. Press, 1973.

[24] A. Cangelosi, M. Schlesinger, and L. B. Smith, *Developmental robotics: From babies to robots*. MIT Press, 2015.