View Estimation Learning based on Value System

Yasutake Takahashi, Kouki Shimada, and Minoru Asada

Abstract— **Estimation of a caregiver's view is one of the most important capabilities for a child to understand the behavior demonstrated by the caregiver, that is, to infer the intention of behavior and/or to learn the observed behavior efficiently. We hypothesize that the child develops this ability in the same way as behavior learning motivated by an intrinsic reward, that is, he/she updates the model of the estimated view of his/her own during the behavior imitated from the observation of the behavior demonstrated by the caregiver based on minimizing the estimation error of the reward during the behavior. From this view, this paper shows a method for acquiring such a capability based on a value system from which values can be obtained by reinforcement learning. The parameters of the view estimation are updated based on the temporal difference error (hereafter TD error: estimation error of the state value), analogous to the way such that the parameters of the state value of the behavior are updated based on the TD error. Experiments with simple humanoid robots show the validity of the method, and the developmental process parallel to young children's estimation of its own view during the imitation of the observed behavior demonstrated by the caregiver is discussed.**

I. INTRODUCTION

Estimation of a caregiver's view is one of the most important capabilities for a child in understanding behavior demonstrated by the caregiver, that is, to infer the intention of behavior, and/or to learn the observed behavior efficiently. Understanding the observed behavior means, in this paper, that the child recognizes the goal of the behavior, observes a reward received at the goal, and performs (or at least imagines) actions that will lead to the goal by itself. The child learns a lot of behavior through trial and error without instruction from the caregiver. The mapping from sensory information to motor skill, that is, the behavior representation, of the child would be based on an egocentric coordinate, not on an allocentric one yet, since young children and autistic children seem to have difficulty understanding the relationship between the caregiver and objects. Bekkering and Wohlschlager [1] showed that a young child has difficulty imitating behavior that was taught in a face-to-face situation. This implies that the ability of view estimation is not innate but acquired through behavior development of the child.

This behavior learning by a child seems to be motivated by some kinds of happiness or joy. The child feels an intrinsic reward when it reaches the goals of the behavior and learns the skills through trial and error to receive the reward. Experiments by Hollerman and Schultz [2] strongly suggest that the activity of dopamine neurons encode the error between the predicted reward and the actual one. The prediction error influences the behavior of the child. This can be modeled as reinforcement learning [3].

Reinforcement learning has been studied well for motor skill learning and robot behavior acquisition in both single and multi-agent environments [4]. The reinforcement learning generates not only an appropriate behavior (a map from states to actions) to achieve a given task but also a utility of the behavior, an estimated discounted sum of the reward value that will be received in the future while the agent is taking an appropriate policy. This estimated discounted sum of the reward is called "state value." Estimation error of the state value is called "temporal difference error" (hereafter TD error) and the agent updates the state value and the behavior based on the TD error. Eventually, the agent represents its behavior based on the state value.

On the other hand, Meltzoff suggests [5] "Like me" hypothesis that a child uses the experience of self to understand the actions, goals, and psychological states of a demonstrator[1] including its caregiver. From a viewpoint of reinforcement learning framework, this hypothesis indicates that the reward and state value of the demonstrator might be estimated through observing the behavior. Takahashi et al. proposed a method to understand observed behavior based on the state value estimation [6] and a method for mutual development of acquisition and recognition of observed behavior [7]. From the above viewpoints, the TD error might be utilized not only for behavior learning but also for the view estimation.

When the child observes behavior of a caregiver and tries to understand and imitate the behavior, it needs to estimate the view of the caregiver and map the trajectory of palms or objects to the representation of its own behavior during the observation. How can the child acquire the mapping from the observation to the self-sensory information in order to imitate the observed behavior?

Most of the conventional approaches to imitative learning (for example, [8], [9]) assume a global coordinate system in order to mimic the observed motion; the observer can transform the trajectory of the demonstrator's motion into a Cartesian coordinate system of the environment or the joint space of the demonstrator and the observer imitates manipulative tasks or gestures. On the other hand, Asada et al. [10], [11] proposed a view-based imitation learning system that estimates the view of the demonstrator based on an opt-geometric constraint called an "epipolar constraint"

Yasutake Takahashi, Kouki Shimada, and Minoru Asada are with the Dept. of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Yamadaoka 2-1, Suita, Osaka, 565-0871, Japan (phone: +81 6 6879 4123; email: {yasutake,kouki.shimada,asada}@ams.eng.osaka-u.ac.jp).

Minoru Asada is also with the JST ERATO Asada Synergistic Intelligence Project (email: asada@jeap.org).

[1]For reasons of consistency, the term "demonstrator" is used to describe any agent from which an observer can learn, even if the demonstrator does not have an intention to show its behavior to the observer.

between two cameras of the observer. Yoshikawa et al. [12] proposed a mechanism in which the observer learns view transformation incrementally based on the idea of view-based imitation. Their idea of view transformation learning is based on the opt-geometric constraint and estimation of the posture of the demonstrator therefore, it is basically independent from behavior learning by imitation. Yokoya et al. [13] proposed a view estimation method based on projecting a model of self-behavior. They strongly assume that the caregiver always mimics the motion of the child, then, the child estimates view estimation parameters by matching the observed trajectory and its own motion. Our motivation on this research is similar to the one of them in a way that the child develops the view estimation based on the self-behavior, but we stick to the idea of utilizing value system, TD error, for the development of the view estimation in order to integrate the view estimation and behavior learning through imitation based on value system seamlessly in future.

From the viewpoint of a reinforcement learning framework, Meltzoff's "Like me" hypothesis [5] indicates that a child estimates rewards received by his/her caregiver based on experiences of self. As Takahashi et al. [6], [7] have shown so far, estimation of rewards received during the observed behavior based on reward models of own behavior enables an agent to understand/recognize/learn the observed behavior deeply. View estimation is important to understand/imitate the observed behavior because the mapping from sensory information to motor skill would be based on an egocentric coordinate, not on an allocentric one. Based on the discussions above, we hypothesize that the child develops this ability in the same way as behavior learning motivated by an intrinsic reward, that is, he/she updates the model of the estimated view of his/her own during the behavior imitated from the observation of the behavior demonstrated by the caregiver based on minimizing the estimation error of the reward during the behavior.

Here, we propose a method by which the agent develops the ability of view estimation based on the TD error in the reinforcement learning framework. The parameters of the view estimation are updated based on the TD error, analogous to the way in which the parameters of the state value of the behavior are updated based on the TD error. Experiments with simple humanoid robots show the validity of the method, and the developmental process parallel to young children's estimation of its own view during the imitation of the observed behavior is discussed.

II. VIEW ESTIMATION BASED ON TD ERROR

A scenario of our experiment is shown first. Then we describe the reinforcement learning scheme, the state/action value function, recognition/understanding of observed behavior, updating strategy of estimation parameters based on TD error, and formulation of view estimation.

A. Scenario of Experiment

Fig. 1 shows the scenario of our experiment. There are two players in front of a table. A few objects are on the table. Both players have independently acquired behavior of reaching for each object and maintain a value system based on reinforcement learning. After the behavior learning, one of the players becomes a demonstrator and touches one of the objects. The other player, an observer, tries to estimate its own view during the imitation of the demonstrated behavior while the demonstrator is displaying the behavior. Parameters for the view estimation depend not on the demonstrated task but on relation between positions of the demonstrator and the observer. The observer can learn/understand the observed behavior with the view estimation ability developed based on its own behavior model including value system. Therefore, the development of the view estimation ability follows the processes below:

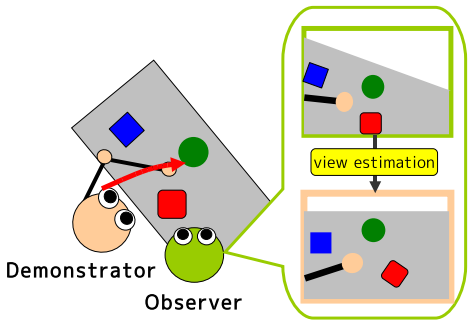


Fig. 1. Scenario of the experiment : Two players are in front of a table. A number of objects are put on the table. One of the players becomes a demonstrator and touches one of the objects. The other player, an observer, tries to estimate the view of self during imitation of the observed behavior.

1) Behavior learning through trial and error
2) Development of view estimation ability for reward estimation
3) Understanding and imitation of observed behavior based on mapping from observation to self-sensory information

The behavior learning through trial and error has been studied in reinforcement learning society for decades. This paper focuses on the process 2), that is, development of view estimation ability for reward estimation. The view estimation enables the observer to understand and imitate the observed behavior based on mapping from observation to self-sensory information.

B. State value

An agent can discriminate a set S of distinct world states. The world is modeled as a Markov process, making stochastic transitions based on its current state and the action taken by the agent based on a policy π. The agent receives reward r_t at each step t after it follows the policy π. State value at state s_t, $V(s_t)$, the discounted sum of the reward received over time under execution of policy π, will be calculated as follows:

$$V(s_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots] \ . \qquad (1)$$

$0 < \gamma < 1$ is a discount rate. The agent receives a positive reward if it reaches a specified goal and zero otherwise,

therefore, the state value increases if the agent follows a good policy π. The agent updates its policy through trial and error in order to receive higher positive rewards in the future. From 1, the state value V_t can be derived as:

$$V(s_t) = E[r_t] + \gamma V(s_{t+1}) \ . \tag{2}$$

Then the state value V_t can be updated iteratively by:

$$V(s_t) \leftarrow V(s_t) + \alpha \Delta V(s_t) \tag{3}$$

$$\Delta V(s_t) = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{4}$$

where $\alpha (0 < \alpha \leq 1)$ is the update ratio. The $\Delta V(s_t)$ is called Temporal Difference error (TD error) and it is used for updating the parameter of estimation of the state value function and policy. Fig. 2(a) shows a diagram of the state value updating procedure. For further details, please refer to the textbook by Sutton and Barto [3] or a survey of robot learning [4].

C. Understanding Observed Behavior based on State Value of Self

Reinforcement learning generates not only an appropriate behavior (a map from states to actions) to accomplish a given task but also the state value. This value roughly indicates closeness to the goal state of the given task if the agent receives a positive reward when it reaches the goal and zero otherwise, that is, if the agent is getting closer to the goal, the value becomes higher. This suggests that the observer may recognize which goal the observed agent would like to achieve if the value of the corresponding task is going higher. Takahashi et al. [6] proposed a method of not only learning and executing a variety of behaviors but also recognizing and understanding the behavior of others, supposing that the observer has already acquired the values of all kinds of behaviors the observed agent can do.

In order to map the observed behavior to the state value of its own behavior, the view estimation during the imitation from the observation is needed. Here we introduce the following assumptions:

- State transition during the observation follows a Markov process.
- State value acquired by itself is fixed as a reference. Observation of the behavior of the demonstrator does not affect anything on the state value.
- Observed behavior of the demonstrator is always one of the behavior acquired beforehand, and the observer knows which behavior the demonstrator is taking by observing a reward received by the demonstrator.[2]

In order to understand the observed behavior based on the state value of the observer, the agent has to follow the procedure below: The agent

1) observe the behavior of the demonstrator,
2) estimate the view of self during imitation of the observed behavior,

[2]This assumption is very natural as we assume that we share "value" with colleagues, friends, or our family in our daily life.

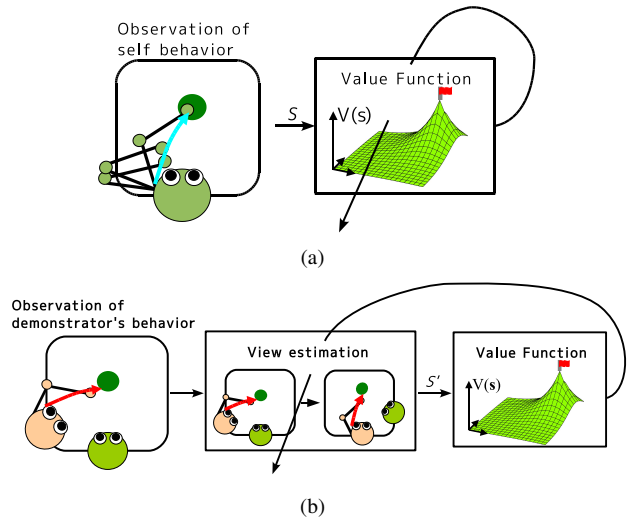


Fig. 2. (a)Update of state values based on TD error through trial and error, (b)Update of view estimation parameter based on the TD error

3) estimate the state value based on the estimated view, and
4) recognize the observed behavior based on the sequence of the estimated state value.

From the above assumptions, there is no room to change any parameters in items 1, 3, and 4 in the procedure. Then, the parameters for estimation of the view of the demonstrator in 2 should be accordingly updated in order to maintain consistency with the concept of behavior representation and recognition based on the state value.

D. View Estimation Parameter Update based on TD error

Fig. 2(b) shows a sketch of the parameter update of view estimation based on TD error. The observer receives sensory ${}^o\boldsymbol{x}$, first. The ${}^o\boldsymbol{x}$ contains, for example, position of a palm of the demonstrator. The sensory information from the view of self during the imitation of the observed behavior ${}^d\boldsymbol{x}$ is estimated from ${}^o\boldsymbol{x}$ by a transformation matrix ${}^d\boldsymbol{T}_o$:

$${}^d\boldsymbol{x} = {}^d\boldsymbol{T}_o \, {}^o\boldsymbol{x} \tag{5}$$

Parameters $\boldsymbol{\phi}$ in the view estimation matrix ${}^d\boldsymbol{T}_o$ are updated based on estimated TD error $\Delta\hat{V}_t$:

$$\Delta\hat{V}_t = \hat{r}_t + \gamma\hat{V}_{t+1} - \hat{V}_t \ , \tag{6}$$

where

$$\hat{r}_t = r(\hat{s}_t) \ , \ \hat{V}_t = V(\hat{s}_t) \ , \ \hat{s}_t \leftarrow F^{hash}({}^d\boldsymbol{x}_t) \ .$$

F^{hash} is a hash function that maps from sensory values ${}^d\boldsymbol{x}_t$ to a state $s \in \boldsymbol{S}$. The parameters of view estimation $\boldsymbol{\phi}$ are updated as TD error decreases as follows:

$$\phi_{ij} \leftarrow \phi_{ij} - \beta \frac{\partial |\Delta V_t|}{\partial \phi_{ij}} \tag{7}$$

where i, j is an indexes of the parameter of view estimation.

In the following experiments, the state space is quantized into a set of discrete states and the state value function

is represented in this space. When the differential of the state value is calculated, in order to avoid a problem of the discontinuity of the function, the state value is interpolated linearly as

$$\hat{V}_t \leftarrow V({}^d\boldsymbol{x}_t) \tag{8}$$

and the TD error of (6) is calculated with the interpolated state value. Then, $\frac{\partial|\Delta\hat{V}_t|}{\partial\phi_{ij}}$ is calculated in a numerical manner as below:

$$\frac{\partial|\Delta\hat{V}_t|}{\partial\phi_{ij}} \rightarrow \frac{|\Delta\hat{V}_t({}^d\boldsymbol{x}_t|^{\phi_{ij}+\delta\phi_{ij}})| - |\Delta\hat{V}_t({}^d\boldsymbol{x}_t|^{\phi_{ij}-\delta\phi_{ij}})|}{2\delta\phi_{ij}} \tag{9}$$

where ${}^d\boldsymbol{x}_t|^{\phi_{ij}+\delta\phi_{ij}}$ and ${}^d\boldsymbol{x}_t|^{\phi_{ij}-\delta\phi_{ij}}$ are estimated sensory information vectors of the demonstrator. The parameter of the view estimation matrix ϕ_{ij} is increased or decreased by $\delta\phi_{ij}$, respectively.

E. View Estimation

The affine transformation matrix can be the model for view transformation only for simplicity although it is not the only one suitable for the given task below. Each agent has a perspective camera on the head and acquires, for example, a palm position of the agent on the camera image. Let ${}^o x = ({}^o x, {}^o y)$ and ${}^d x = ({}^d x, {}^d y)$ be the palm positions on the views of the observer and the demonstrator, respectively. Then, the estimated palm position on the view of the demonstrator can be calculated with the view estimation matrix and the palm position of the view of the observer as below:

$$\begin{pmatrix} {}^d x \\ {}^d y \\ 1 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} {}^o x \\ {}^o y \\ 1 \end{pmatrix}.$$

The view estimation matrix depends on the relationship between positions of the observer and the demonstrator. As we assume in II-C, the relationship between the observer and the demonstrator is fixed during the learning of the view estimation in the following experiments.

III. TARGET TOUCH GAME WITH HUMANOID ROBOTS

A. Humanoid Simulator

The scenario of the experiments in this paper was briefly described in II-A. This section explains a humanoid simulator on which our experiments are made, and experimental setups such as the demonstrator's behavior during the estimation of view field by the observer and the position configuration of both players. Fig.3 shows the simulator. There are two humanoid robots and colored objects on a table. The robot has two-degree-of-freedom arms each of which has elbow and shoulder joints and can sweep a palm horizontally. Simple color image processing is applied to detect the palm and the objects on the image captured by a camera mounted on the head of the humanoid robot. The size of a table and the distance between observer and demonstrator are shown in Fig.5(a).

The observer and the demonstrator are supposed to be a child and a caregiver. The observer has 2/3 size of body of demonstrator, that is, the length of the arm, body and head of the observer is 2/3 of the ones of the demonstrator. The position of the observer's camera is lower than the one of the demonstrator, then, the observer watches the behavior of the demonstrator at closer position than the demonstrator. Therefore, the view estimation system must have robustness against not only parallel and rotation translation but also scale change.

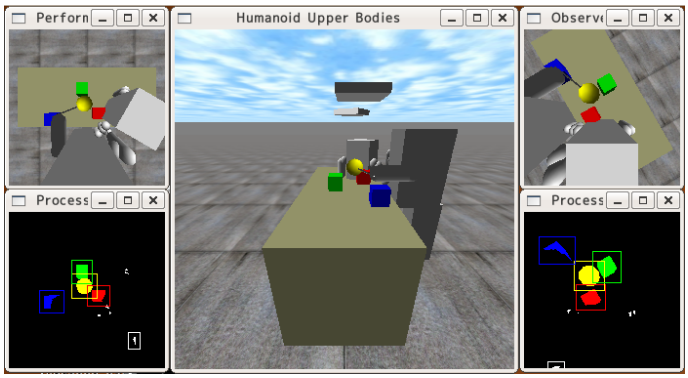


Fig. 3. Viewer of simulator. *Top-left*: view of a demonstrator *Bottom-left*: objects and a palm detection on the camera image of the demonstrator *Center*: overview of the experiment with two humanoids and three objects on a table *Top-right*: view of an observer *Bottom-right*: objects and a palm detection on the camera image of the observer

B. View Estimation

First of all, the observer learns the state value functions of reaching for the colored objects in the manner of reinforcement learning. A state space for the state value estimation is constructed with two state variables: x and y coordinates of palm from observer's view. Each state variable is divided into 30 slots in order to construct a discretized state space. Positive reward (+1) is given when the robot touches the colored object and 0 reward else. The robot updates the estimated state value by reaching for one of the objects from all possible positions of its palm. One state value function is assigned to each behavior of reaching for one of the objects. The robot maintains a total of three state value functions because there are three objects on the table.

Fig.4 shows the state value function of the behavior of reaching for three objects. The state value is not distributed like a cone; the shape of the state value function is like a mountain with gentle and steep slopes because of the manipulability of the arm and the constraint of the manipulator configuration: that is, it has a limitation of the reachable area because of the length of the arm and its structure.

Next, the robot observes behavior of the demonstrator. The demonstrator shows the reaching behavior that the observer learned before. The observer watches the behavior from beside the demonstrator as shown in Fig. 1. The observer stays in the position during the view estimation for its imitation. The demonstrator shows the reaching behavior under various initial palm conditions; it moves its palm to various reachable positions first, then, reaches for one

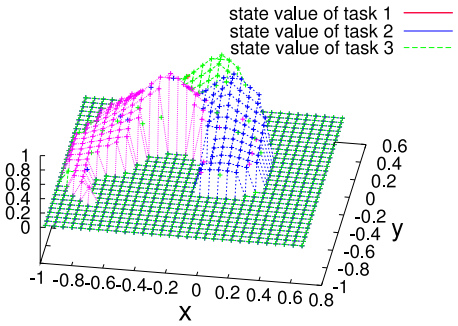
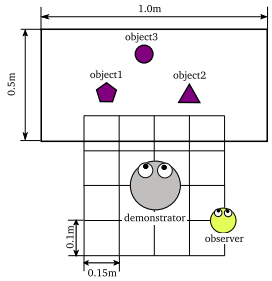
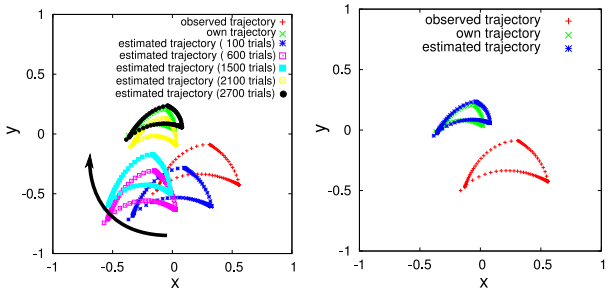


Fig. 4. Three state value functions for reaching behavior to each of three objects



(a) experimental set-up



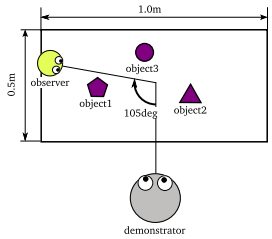
(b) changes of trajectory on estimated view (c) final estimation

Fig. 5. Estimation of view of self during imitation of observed behavior: Observer posture is parallel to the demonstrator.

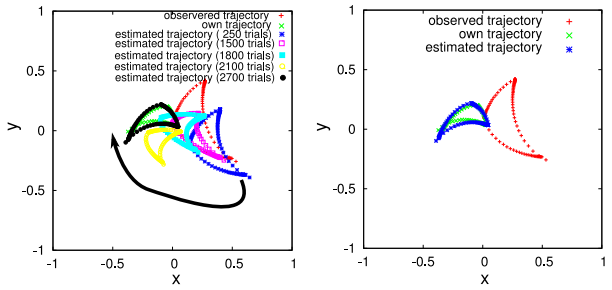
of the objects on the table. It repeats the behavior under the various initial conditions many times. The observer updates the view estimation matrix during the observation of the demonstrator's behavior. The view estimation matrix is initialized as one unit. The update ratio β in equation (7) is fixed to 0.01 in the experiments in this paper.

In order to evaluate the estimated view transformation matrix, the observer estimates the view of self for imitation of the observed behavior while the demonstrator moves its palm to three objects one-by-one. The trajectory is like a triangle shape. The estimated trajectory in the view of the demonstrator is compared with the true trajectory in the view of self during the imitation of the observed behavior.

Figs. 5, 6, 7, and 8 show the configurations of the experiments and the results of the view estimation. Figs. 5(a) and 6(a) show the position and posture conditions of the observer: the observer stands 0.1m to the back and 0.3m



(a) experimental set-up



(b) changes of trajectory on estimated view (c) final estimation

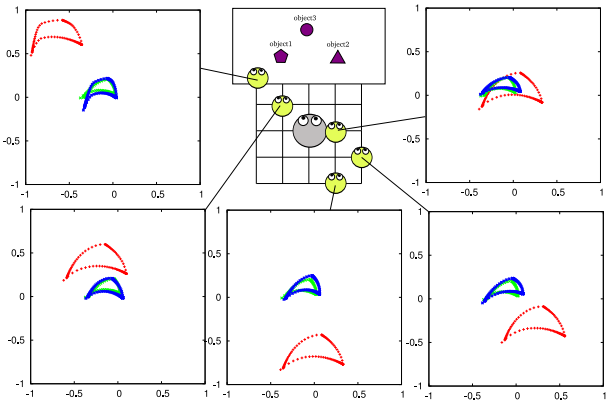
Fig. 6. Estimation of view of self during imitation of observed behavior: Observer orients to the demonstrator with 105 degrees rotation.

to the right of the demonstrator in Fig. 5(a), and the observer stands 105 degrees from the demonstrator around the table in Fig. 6(a). Figs. 5(c) and 6(c) are transit of the trajectories on the estimated views during the learning parameters of the view estimation matrix in the cases, respectively. Figs. 5(b) and 6(b) shows the final results of the estimated trajectories. In both cases, the parameters of the view estimation matrix become correct enough to estimate the trajectory in the view of self during the imitation. Fig.7 shows the results under many conditions of the observer's position and posture. When the observer moves in parallel, the estimated trajectories are close to ones from the view of the demonstrator even though 0.2m away from the demonstrator[3] ,(see Figs. 7(a) and 8(a)). When the observer moves around the table, the observer can estimate the trajectories from the demonstrator until 105 degrees from the demonstrator (see Figs. 7(b) and 8(b)).

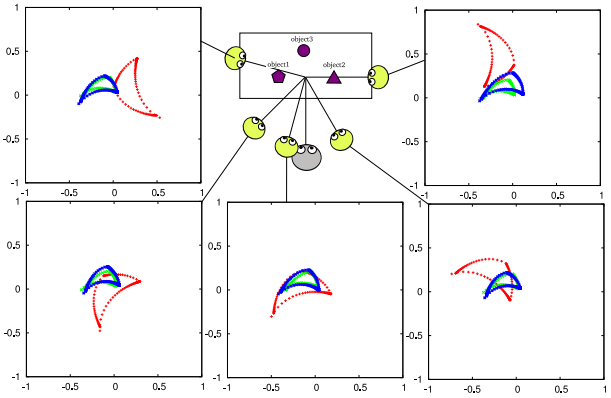
IV. CONCLUSION

We proposed a hypothesis that an agent develops the ability of the view estimation based on the TD error in the reinforcement learning framework. From the view point of reinforcement learning framework, Meltzoff's "Like me" hypothesis indicates that a child estimates rewards received by his/her caregiver based on experiences of self. Estimation of rewards received during the observed behavior based on reward models of own behavior enables an agent to understand/recognize/learn the observed behavior deeply[6], [7]. View estimation is important to understand/imitate the

[3]The furthest distance that the observer can watch the trajectory of the demonstrator is 0.2m in this experiment.

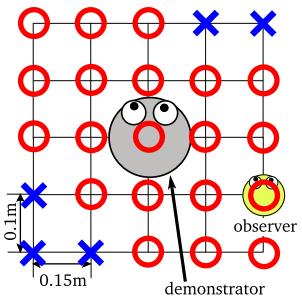


(a) parallel translation

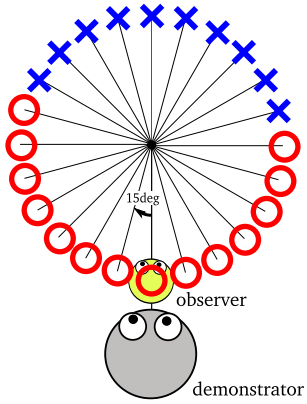


(b) rotational translation

Fig. 7. Final estimated trajectories in various postures. Red, green, and blue curves indicate trajectories in case of observation, estimation, and imitation.



(a) parallel translation



(b) rotational translation

Fig. 8. Success or failure of view estimation in various posture: Circle and cross indicate success and failure, respectively

observed behavior because the mapping from sensory information to motor skill would be based on an egocentric coordinate, not on an allocentric one. Based on the discussions above, we hypothesize that the child develops this ability in the same way as behavior learning motivated by an intrinsic reward, that is, he/she updates the model of the estimated view of his/her own during the behavior imitated from the observation of the behavior demonstrated by the caregiver based on minimizing the estimation error of the reward during the behavior. The parameters of the view estimation are updated based on the TD error, analogous to the way the parameters of the state value of the behavior are updated based on the TD error. Experiments with simple humanoid robots showed the validity of this idea and we expect our method can be helpful in explaining the process of infant development of the view estimation for understanding/imitation of observed behavior based on the value system of self.

REFERENCES

[1] H. Bekkering and A. Wohlschlager, "Imitation of gestures in children is goal-directed," *The Quarterly Journal of Experimental Psychology*, vol. 53, pp. 153–164, 2000.

[2] J. R. Hollerman and W. Schultz, "Dopamine neurons report an error in the temporal prediction of reward during learning," *Nature Neuroscience*, vol. 1, no. 4, pp. 304–309, 1998.

[3] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998. [Online]. Available: citeseer.ist.psu.edu/sutton98reinforcement.html

[4] J. H. Connell and S. Mahadevan, *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.

[5] A. N. Meltzoff, "'like me': a foundation for social cognition," *Developmental Science*, vol. 10, no. 1, pp. 126–134, 2007.

[6] Y. Takahashi, T. Kawamata, M. Asada, and M. Negrello, "Emulation and behavior understanding through shared values," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2007, pp. 3950–3955.

[7] Y. Takahashi, Y. Tamura, and M. Asada, "Mutual development of behavior acquisition and recognition based on value system," in *From Animals to Animats 10 (Proceedings of 10th International Conference on Siulation of Adaptive Behavior, SAB 2008)*, July 2008, pp. 291–300.

[8] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation," pp. 199–218, 2004.

[9] T. Inamura, Y. Nakamura, and I. Toshima, "Embodied symbol emergence based on mimesis theory," *International Journal of Robotics Research*, vol. 23, no. 4, pp. 363–377, 2004.

[10] M. Asada, Y. Yoshikawa, and K. Hosoda, "Learning by observation without three-dimensional reconstruction," in *Proceedings of Intelligent Autonomous Systems (IAS-6)*, 2000, pp. 555–560.

[11] Y. Yoshikawa, M. Asada, and K. Hosoda, "View-based imitation learning by conflict resolution with epipolar geometry," in *Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2001, pp. 1416–1427.

[12] ——, "Developmental approach to spatial perception for imitation learning: Incremental demonstrator's view recovery by modular neural network," in *Proceedings of the 2nd IEEE-RSA International Conference on Humanoid Robot*, 2001, pp. 107–114.

[13] R. Yokoya, T. Ogata, J. Tani, K. Komatani, and H. G. Okuno, "Discovery of other individuals by projecting a self-model through imitation," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, San Diego, Oct 2007, pp. 1009–1014.