Human Instruction Recognition and Self Behavior Acquisition based on State Value

Yasutake Takahashi, Yoshihiro Tamura, and Minoru Asada

Abstract— **A robot working with humans or other robots is supposed to be adaptive to changes in the environment. Reinforcement learning has been studied well for motor skill learning, robot behavior acquisition and adaptation of the behavior to the enviromental changes. However, it is not practical that the robot learns and adapts its behavior only through trial and error by itself from scratch because huge exploration is needed. Furtunately, it is nothing unusual to have predecessors in the environment and it is reasonable to learn something from the observation of predecessors' behavior. In order to learn various behavior from the observation, the robot must segment the behavior based on reasonable criterion for itself and feedback the data to behavior leanring byitself. This paper presents a case study for a robot to understand unfamiliar behavior shown by a human instructor through the collaboration between behavior acquisition and recognition of observed behavior, where the state value has an important role not simply for behavior acquisition (reinforcement learning) but also for behavior recognition (observation). The validity of the proposed method is shown by applying it to a dynamic environment where one robot and one human play soccer.**

I. INTRODUCTION

A robot working with humans or other robots is supposed to be adaptive to changes in the environment. Reinforcement learning has been studied well for motor skill learning, robot behavior acquisition and adaptation of the behavior to the enviromental changes. However, it is not practical that the robot learns and adapts its behavior only through trial and error by itself from scratch because huge exploration is needed. Furtunately, it is nothing unusual to have predecessors in the environment and it is reasonable to learn something from the observation of predecessors' behavior. Especially, in the multi-agent environment, observation of others make the behavior learning rapid and therefore much more efficient [1], [2], [3]. Actually, it is desirable to acquire various unfamiliar behavior with some instructions from others, for example, surrounding robots and/or humans in real environment because of huge exploration space and enormous learning time to learn. Therefore, behavior learning through observation has been more important. In order to learn various behavior from the observation, the robot must segment the behavior based on reasonable criterion for itself and feedback the data to behavior leanring byitself. From a viewpoint of the reinforcement learning framework, this means reading rewards of the observed behavior and estimating sequence of the value through the observation and feedback the estimated rewards to its own behavior learning system.

Yasutake Takahashi, Yoshihiro Tamura, 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,yoshihiro.tamura,asada}@ams.eng.osaka-u.ac.jp).

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

Takahashi et al.[4] proposed a method of not only to learn and execute a variety of behavior but also to recognize behavior of others supposing that the observer has already acquired the values of all kinds of behavior the observed agent can do. The recognition means, in this paper, that the robot categorizes the observed behavior to a set of its own behavior acquired beforehand. The method seamlessly combines behavior acquisition and recognition based on "state value" in reinforcement learning scheme. Reinforcement learning generates not only an appropriate behavior (a map from states to actions) to accomplish a given task but also a utility of the behavior, an estimated discounted sum of rewards that will be received in future while the robot is taking an appropriate policy. This estimated discounted sum of reward is called "state value." This value roughly indicates closeness to the goal state of the given task if the robot receives a positive reward when it reaches the goal and zero else, 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 likes to achieve if the value of the corresponding task is going higher.

Takahashi et al.[5] proposed an extended method that enhances behavior acquisition and recognition based on interaction between learning and observation of behavior. A robot learns its behavior through not only trial and error but also reading rewards of the observed behavior of others (including surrounding robots and humans). In this study, however, the instruction data is segmented by a coach even though he/she does not categolize nor inform them to the learner. This means that all the learner has to do is just classification of the instructions to some embedded behavior of itself and it does not have to segment the data by itself.

This paper presents a case study for a robot to understand unfamiliar behavior shown by a human demonstrator through the collaboration between behavior acquisition and recognition of observed behavior. The human demonstrator shows a set of behavior without any interuption so that the robot is not informed the segments of the observed behavior. The proposed method shows practical performance of learning and recognizing the observed behavior under a dynamic environment where one robot and a human demonstrator play soccer.

II. OUTLINE OF THE MECHANISMS

We review the behavior recognition/learning system proposed by Takahashi et al.[5], briefly.

A. Behavior Learning Based on Reinforcement Learning

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. State value V^π, the discounted sum of the reward received over time under execution of policy π, will be calculated as follows:

$$V^\pi = \sum_{t=0}^{\infty} \gamma^t r_t \quad . \tag{1}$$

In case that the agent receives a positive reward if it stays a specified goal and zero else, then, the state value increases if the agent follows an appropriate policy π. The agent updates its policy through a process of trial and error in order to receive higher positive rewards in future. Analogously, as animals get closer to former action sequences that led to goals, they are more likely to retry it. For further details, please refer to the textbook of Sutton and Barto[6] or a survey of robot learning[7].

Here we introduce a model-based reinforcement learning method. A learning module has a forward model which represents the state transition model and a behavior learner which estimates the state-action value function based on the forward model in a reinforcement learning manner.

Each learning module has its own state transition model. This model estimates the state transition probability $\hat{\mathcal{P}}^a_{ss'}$ for the triplet of state s, action a, and next state s':

$$\hat{\mathcal{P}}^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{2}$$

Each module has a reward model $\hat{\mathcal{R}}_s$, too:

$$\hat{\mathcal{R}}(s) = E\{r_t | s_t = s\} \tag{3}$$

All experiences (sequences of state-action-next state and reward) are simply stored to estimate these models. Now we have the estimated state transition probability $\hat{\mathcal{P}}^a_{ss'}$ and the expected reward $\hat{\mathcal{R}}_s$, then, an approximated state-action value function $Q(s,a)$ for a state action pair s and a is given by

$$Q(s,a) = \sum_{s'} \hat{\mathcal{P}}^a_{ss'} \left[\hat{\mathcal{R}}(s') + \gamma V(s') \right] \tag{4}$$

$$V(s) = \max_a Q(s,a) \ , \tag{5}$$

where γ is a discount factor.

B. Modular Learning System

In order to observe/learn/execute a number of behavior, a modular learning system is adopted. Many modular architectures have been proposed so far (for example [8], [9], [10]). Each module is responsible for learning to achieve a single goal. One arbiter or a gate module is responsible for merging information from the individual modules in order to derive a single action performed by the robot.

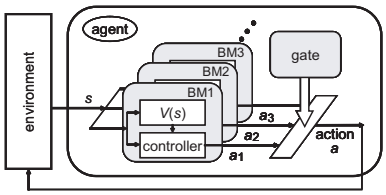


Fig. 1. Modular learning system

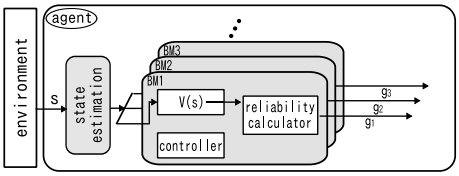


Fig. 2. Behavior inference diagram

Fig.1 shows a sketch of such a modular learning system. We prepare a number of behavior modules (BMs in the figure) each of which adopts the behavior learning method described in II-A. The module is assigned to one goal-oriented behavior and estimates one action value function $Q(s,a)$. A module receives a positive reward when it accomplishes the assigned behavior or zero reward else. The behavior module has a controller that selects the action with the maximum value based on the predictions of next state values. The gating module will then select one output from the inputs of the different behavior modules according to the player's intention.

Fig.2 shows a diagram to recognize an observed behavior. The same behavior modules are used for the behavior recognition. Each behavior module estimates the state value based on the estimated state of the observed demonstrator[1] and calculates reliability of observed behavior, that is, how likely the demonstrator is taking the behavior of the module. The details are described in following sections.

C. Behavior Recognition based on Estimated Values

While an observer watches a demonstrator's behavior, it uses the same behavior modules for recognition of observed behavior as shown in Fig.2. Each behavior module estimates the state value based on the estimated state of the observed demonstrator and sends it to the selector. The selector watches the sequence of the state values and selects a set of possible behavior modules of which state values are going up as a set of behavior the demonstrator is currently taking. As mentioned before, if the state value goes up during a behavior, it means that the module is valid for explaining the behavior. The observed behavior is recognized by a set of behavior whose modules' values are increasing.

[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.

Here we define behavior recognition reliability g that indicates how much the observed behavior would be reasonable to be recognized as a behavior

$$g = \begin{cases} g + \beta & \text{if } V_t - V_{t-1} > 0 \text{ and } g < 1 \\ g & \text{if } V_t - V_{t-1} = 0 \\ g - \beta & \text{if } V_t - V_{t-1} < 0 \text{ and } g > 0 \end{cases},$$

where β is an update parameter, and 0.1 in this paper. This equation indicates that the reliability g will become large if the estimated utility rises up and it will become low when the estimated utility goes down. Another condition is to keep g value from 0 to 1.

D. Learning from Observation

In the previous section, behavior recognition system based on state value of its own behavior is described. This system shows robust recognition of observed behavior [11] only when the behavior to be recognized has been well-learned beforehand. If the behavior is under learning, then, the recognition system is not able to show good recognition performance at beginning. Here, we shows how the estimated state value of observed behavior, $\hat{V}(s)$, gives feedback to learning and understanding unfamiliar observed behavior and this feedback loop enhances the performance of observed behavior recognition. Fig.3 shows a rough idea of our proposed method. $V(s)$ and $\hat{V}(s)$ are the state value updated by oneself and the state value estimated through observation, respectively.

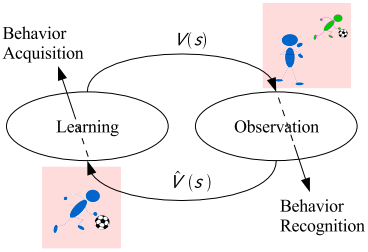


Fig. 3. Interaction between Learning and Observation of Behavior

The trajectory of the observed behavior can be a bias for learning behavior and might enhance the behavior learning based on the trajectory. The observer cannot watch actions of observed behavior directly and can only estimate the sequence of the state of the observed robot. Let s^o_t be the estimated state of the observed robot at time t. Then, the estimated state value $\hat{V}^o$ of the observed behavior can be calculated as below:

$$\hat{V}^o(s) = \sum_{s'} \hat{\mathcal{P}}^o_{ss'} \left[\hat{\mathcal{R}}(s') + \gamma V^o(s') \right] \quad (6)$$

where $\hat{\mathcal{P}}^o_{ss'}$ is state transition probability estimated from the behavior observation. This state value function $\hat{V}^o$ can be used a bias of the state value function of the learner V. The learner updates its state-action value function $Q(s,a)$ during the process of trial and error based on the estimated state value of observed behavior $\hat{V}^o$ as below:

$$Q(s,a) = \sum_{s'} \hat{\mathcal{P}}^a_{ss'} \left[\hat{\mathcal{R}}(s') + \gamma V'(s') \right] \quad (7)$$

while

$$V'(s) = \begin{cases} V(s) & \text{if } V(s) > \eta^n \hat{V}^o(s) \\ \eta^n \hat{V}^o(s) & \text{else} \end{cases}$$

This is a normal update equation as shown in (4) except using $V'(s)$. The update system switches the state value of the next state s' between the state value of own learning behavior $V(s')$ and the one of the observed behavior $\hat{V}^o(s')$ discounted by η^n. η^n is the discount factor ($0 < \eta 1$) based on the time of experiences n. This means that it receives bigger feedback with $\hat{V}^o(s')$ if the state-action transition "$(s,a) \rightarrow s'$" is not experienced so far while smaller else. This means the state value update system takes $\hat{V}^o(s')$ if the learner does not estimate the state value $V(s')$ because of lack of experience at the state s' from which it reaches to the goal of the behavior. $\hat{V}^o(s')$ becomes a bias for reinforcing the action a from the state s even though the state value of its own behavior $V(s')$ is small so that it leads the learner to explore the space near to the goal state of the behavior effectively.

A demonstrator is supposed to show a number of behavior which are not informed directly to the observer. The observer, who learns/recognizes the behavior, assumes that all agents share reward models of the behavior, that is, all agents, including human players, receive a positive reward when the goal of the behavior is achieved . This assumption is very natural as we assume that we share "value" with colleagues, friends, or our family in our daily life. In order to update the estimate values of the behavior the demonstrator is taking, the observer has to estimate which behavior the demonstrator is taking correctly. If the observer waits to learn some specific behavior by observation until it becomes able to recognize the observed behavior well, bootstrap of leaning unfamiliar behavior by observation cannot be expected. Therefore, the observer(learner) maintains a history of the observed trajectories and updates value function of the observed behavior with high reliability or high received reward. The observer estimates the state of the demonstrator every step and the reward received by the demonstrator is estimated as well. If it is estimated that the demonstrator receives a positive reward by reaching to the goal state of the behavior, then, the observer updates the state value of the corresponding behavior even if it has low reliability for the observed behavior. The update strategy enhances to estimate appropriate values of the observed behavior.

III. BEHAVIOR LEARNING BY OBSERVATION

Fig.4 shows a situation in which a human player shows passsing behavior to a behavior learning robot. All objects, e.g., an orange ball, and a goal., player on this field, are color-coded. A simple color image processing is applied in order to detect the color-coded objects and players in real-time. The

Fig. 4. Demonstration of a passing behavior by a human

mobile platform is based on an omni-directional vehicle. The robot and the human play soccer such as dribbling a ball, kicking it to a goal, passing a ball to the other, and so on. While playing with objects, the learning robot watches the behavior of the human player, try to understand observed behavior, and emulate them.

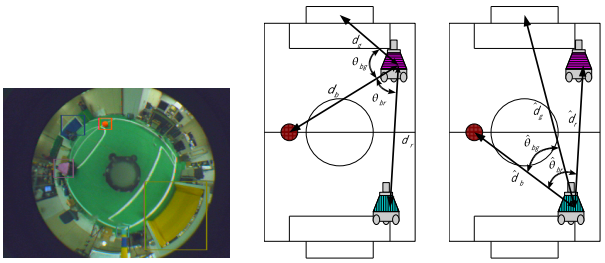


Fig. 5. Estimation of view of the demonstrator. Left : a captured image the of observer, Center : object detection and state variables for self, Right : estimation of view of the demonstrator

Each behavior module estimates a state value of observed behavior at an arbitrary time t to accomplish the specified task. An observer watches a demonstrator's behavior and maps the sensory information from an observer viewpoint to a demonstrator's one with a simple mapping of state variables. Fig.5 shows a simple example of this transformation. It detects color-coded objects on the omni-directional image, calculates distances and directions of the objects in the world coordinate of the observer, and shifts the axes so that the position of the demonstrator comes to center of the demonstrator's coordinate. Then it roughly estimates the state information in the egocentric coordinate and the state of the demonstrator. Every behavior module estimates a sequence of its state value from the estimated state of the observed demonstrator and the system selects modules which values are increasing. The learner tries to acquire a number of behavior shown in Table I. The table also describes necessary state variables for each behavior. Each state variable is divided into 11 in order to construct a quantized state space. 4 actions are prepared to be selected by the learning modules: Approaching the goal, approaching the teammate, going in front of the ball while watching the goal, and going in front of the ball while watching the teammate.

TABLE I

LIST OF BEHAVIOR LEARNED BY SELF AND STATE VARIABLES FOR EACH BEHAVIOR

Behavior	State variables
Approaching a ball	d_b
Approaching a goal	d_g
Approaching the teammate	d_r
Shooting a ball	d_b, d_g, θ_{bg}
Passing a ball	d_b, d_r, θ_{br}

A. Comparison Experiment Setup

In order to validate the effect of interaction between acquisition and recognition of behavior through observation, two experiments are set up. One is that the learner does not observe the behavior of the human player but tries to acquire shooting/passing behavior by itself. The other is that the learner observes the behavior of other and enhances the learning of the behavior based on the estimated state value of the observed behavior. In former experiment, the learner follows the learning procedure:

1) 2 minutes for behavior learning by itself
2) evaluation of self-behavior performance
3) evaluation of behavior recognition performance
4) goto 1.

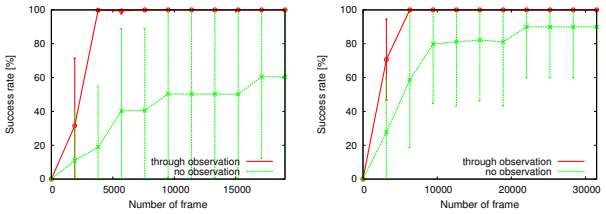
On the other hand, the later experiment, it follows :

1) 1 munite for observation of the behavior of the other
2) 1 munite for behavior learning by self-trials with observed experience
3) evaluation of self-behavior performance
4) evaluation of behavior recognition performance
5) goto 1.

Both learners attempt to acquire behavior listed in Table I. The human player shows a set of the behavior within 1 minute and the learning robot does not know which behavior the human player is taking. In both experiments, the learner follows ϵ-greedy method; it follows the greedy policy with 80% probalibity and takes a random action else. Performance of the behavior execution and recognition of observed behavior during the learning time is evaluated every 15 learning episodes. The performance of the behavior execution is the success rate of the behavior while the learner, the ball, and the teammate are placed at a set of pre-defined positions. The one of the behavior recognition is the average length of period in which the recognition reliability of the right behavior is larger than 70% during the observation. The soccer field area is divided 3 by 3 and the center of the each area is a candidate of the position of the ball, the learner, or the teammate. The performances are evaluated in all possible combinations of the positions.

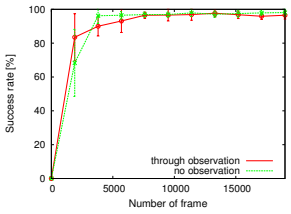
B. Performance of Behavior Learning and Recognition

Fig.6 shows the success rates of the behavior and their variances during learning in cases of learning with/without value update through observation. The success rates with value update of all kinds of behavior grows more rapidly than the one without observation feedback. Rapid learning is one of the most important aspect for a real robot application. The success rate without value update through observation sometimes could not reach the goal of the behavior at the

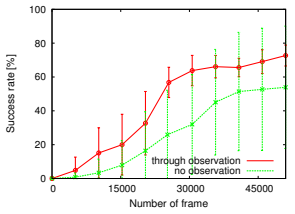


(a) approaching the ball

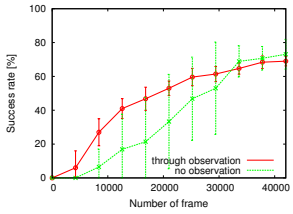
(b) approaching the goal



(c) approaching the teammate

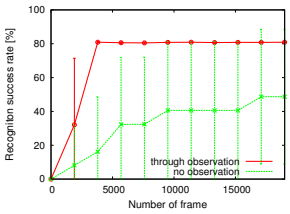


(d) shooting the ball

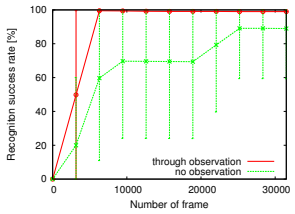


(e) passing to the teammate

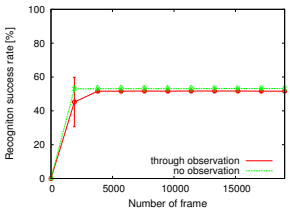
Fig. 6. Success rate of the behavior during learning with/without observation of demonstrator's behavior



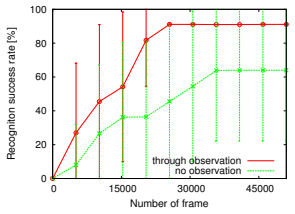
(a) approaching the ball



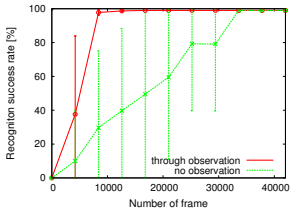
(b) approaching the goal



(c) approaching the teammate



(d) shooting the ball



(e) passing to the teammate

Fig. 7. Recognition performance of the behavior during learning with/without observation of demonstrator's behavior

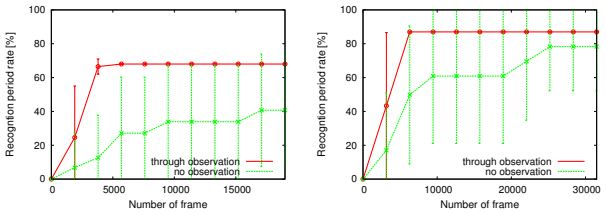
beginning of the learning because there is no bias to lead the robot to learn appropriate actions. This is the reason why the variances of the rate is big. On the other hand, the system with value update through observation utilizes the observation to bootstrap the learning even though it cannot read exact actions of observed behavior.

Recognition performance and recognition period rate of observed behavior and their variances are shown in Figs.7 and 8, respectively. "Recognition period rate" of observed behavior is introduced here to evaluate how long the observer can recognize the observed behavior as a correct one. For example, the recognition period rate is 85% here ,that means, the period in which the reliability of passing behavior is over 70% is 85% during the observation. They indicate a similar aspect with the ones of the success rates. The performance of the behavior recognition depends on the learning performance. If the learning system has not acquired data enough to estimate state value of the behavior, it cannot perform well. The learning system with value update with observed behavior rapidly enables to recognize the behavior while the system without value update based on the observation has to wait to realize a good recognition performance until it estimates good state value of the behavior by its own trial and error. Those figures show the importance of learning through interaction between behavior acquisition and recognition of observed behavior.

IV. CONCLUSION

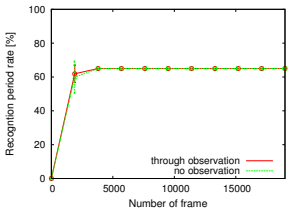
Above, values are defined as behavior, which are defined by the achieved goals. The learning robot uses its own value functions to recognize what the human player will do. Preliminary investigations in a similar context have been done by Takahashi at el. [5] and they showed that the learning robot successfully acquires and recognizes behavior showned by the other teammate under a condition in which the instruction data is segmented by a coach even though he/she does not categolize nor inform them to the learner. This paper shows that the robot can understand unfamiliar behavior shown by a human instructor through the collaboration between behavior acquisition and recognition of observed behavior. The human demonstrator shows a set of behavior without any interuption so that the robot is not informed the segments of the observed behavior. The proposed method shows practical performance of learning and recognizing the observed behavior under a dynamic environment where one robot and a human demonstrator play soccer.

As future work, the behavior recognition system can be extended to generate internal rewards for cooperative/competitive behavior. If the observer can estimate the value of the observed behavior, it might be possible to recognize the other's intention, therefore the observer not only imitate the observed behavior but also learn cooperative/competitive behavior for the demonstrator according to the estimated values. Internal rewards should be also

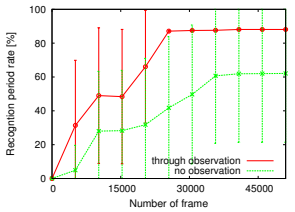


(a) approaching the ball

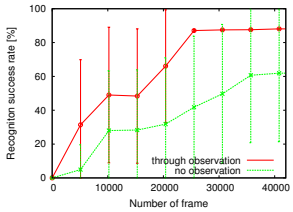
(b) approaching the goal



(c) approaching the teammate



(d) shooting the ball



(e) passing to the teammate

Fig. 8. Recognition period rate of the behavior during learning with/without observation of demonstrator's behavior

developed through observation while we assume that the reward models of the behavior are shared among robots or other entities. Related works have proposed many types of internal/intrinsic rewards for learning acceleration or decomposition of long time-scale tasks. By cooping with the proposed methods, developmental behavior acquisition will be achieved on real robots in our daily life.

V. ACKNOWLEDGMENTS

This work is partially supported by Kayamori Foundation of Informational Science Advancement.

REFERENCES

[1] S. D. Whitehead, "Complexity and cooperation in q-learning," in *Proceedings Eighth International Workshop on Machine Learning (ML91)*, 1991, pp. 363–367.

[2] B. Price and C. Boutilier, "Accelerating reinforcement learning through implicit imitatione," *Journal of Articial Intelligence Research*, 2003.

[3] D. C. Bentivegna, C. G. Atkeson, and G. Chenga, "Learning tasks from observation and practice," *Robotics and Autonomous Systems*, vol. 47, pp. 163–169, 2004.

[4] 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.

[5] Y. Takahashi, Y. Tamura, and M. Asada, "Behavior development through interaction between acquisition and reco gnition of observed behaviors," in *Proceedings of 2008 IEEE World Congress on Computational Intellige nce (WCCI2008)*, June 2008, pp. 1518–1528.

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

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

[8] R. Jacobs, M. Jordan, N. S, and G. Hinton, "Adaptive mixture of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[9] S. P. Singh, "Transfer of learning by composing solutions of elemental sequential tasks." *Machine Learning*, vol. 8, pp. 323–339, 1992. [Online]. Available: citeseer.nj.nec.com/singh92transfer.html

[10] S. Whitehead, J. Karlsson, and J. Tenenberg, "Learning multiple goal behavior via task decomposition and dynamic policy merging," in *ROBOT LEARNING*, J. H. Connell and S. Mahadevan, Eds. Kluwer Academic Publishers, 1993, ch. 3, pp. 45–78.

[11] Y. Takahashi, T. Kawamata, and M. Asada, "Learning utility for behavior acquisition and intention inference of other agent," in *Proceedings of the 2006 IEEE/RSJ IROS 2006 Workshop on Multi-objective Robotics*, Oct 2006, pp. 25–31.