Learning Utility for Behavior Acquisition and Intention Inference of Other Agent

Yasutake Takahashi, Teruyasu Kawamata, and Minoru Asada*
Dept. of Adaptive Machine Systems,
Graduate School of Engineering, Osaka University
Yamadaoka 2-1, Suita, Osaka, 565-0871, Japan
**JST ERATO Asada Synergistic Intelligence Project*
{*yasutake, kawamata, asada*}*@er.ams.eng.osaka-u.ac.jp*

***Abstract*— Neurophysiology revealed the existence of mirror neurons in brain of macaque monkeys and they shows similar activities during executing an observation of goal directed movements performed by self and other. The concept of the mirror neurons/systems is very interesting and suggests that behavior acquisition and the inferring intention of other are related to each other. That is, the behavior learning modules might be used not only for behavior acquisition/execution but also for the understanding of the behavior/intention of other.**

We propose a novel method not only to learn and execute a variety of behaviors but also to infer the intentions of others supposing that the observer has already estimated the utility (state values in reinforcement learning scheme) of all kinds of behaviors the observed agent can do. The method does not need a precise world model or coordination transformation system to deal with view difference caused by different viewpoints. This paper shows an observer can infer an intention of other not by precise object trajectory in global/egocentric coordinate space but by estimated utility transition during the observed behavior.

I. Introduction

Recent robots in real world are required to perform multiple tasks, adapt their behaviors in an encountered multi-agent environment, and learn new cooperative/competitive behaviors through the interaction with other agents. Reinforcement learning has been studied well for motor skill learning and robot behavior acquisition in single/multi agent environments. However, it is unrealistic to acquire a various behaviors from scratch without any instruction from others in real environment because of huge exploration space and enormous learning time. Therefore, importance of instructions from others has been increasing, and in order to understand the instructions, it is necessary to infer their intentions to learn purposive behaviors.

Understanding other agent intention is also a very important issue to realize social activities, for example, imitation learning, cooperative/competitive behavior acquisition, and so on. Recently, many researchers have studied on methods of other agent's behavior recognition system. For example Schaal et al. [3] proposed a motor learning method through imitation of teacher's behaviors. They assume that a learner can observe all state variables and their trajectories in global coordinate system of the environment or the joint space of the others and the learner imitates manipulative tasks or gestures. Doya et al. [2] proposed to estimate intention of other agent for imitation learning and/or cooperative behavior acquisition based on multi-module learning system. These typical approaches assume the detailed knowledge of the task, the environment, and the others (their body structure and sensor/actuator configuration) based on which they can transform the observed sensory data of the others' behaviors into the global coordinate system of the environment, or an egocentric parameter space like the joint space of the others to infer their intentions. However, such an assumption seems unrealistic in the real world and brittle to the sensor/actuator noise(s) or any possible changes in the parameters. Furthermore, there are a variety of behaviors for achieving a certain task. The variety will be caused by constraints of body or environments or experiences received so far. It is almost impossible to cover all variation of behaviors even for one certain tasks. In other words, it is very difficult to infer others' intentions based only on these geometric parameters.

These two issues, behavior acquisition/execution and recognition of intention of other, have been discussed independently. However, neurophysiology recently revealed the existence of mirror neurons/systems in brain and they shows similar activities during executing an observation of goal directed movements performed by self and another one. We do not discuss about this mirror neurons/system here in details, however, the concept of the mirror neurons/system is very interesting and suggests that behavior acquisition/execution and the inferring intention of other are related to each other. That is, the behavior learning modules might be used not only for behavior acquisition but also for the recognition of the behavior/intention of other.

Reinforcement learning generates not only an appropriate behavior (action map from states to actions) to achieve a given task but also an utility of the behavior, an estimated discounted sum of reward value that will be received in future while the robot is taking the optimal policy. We call this estimated discounted sum of reward "state value." This utility roughly indicates closeness to a goal state of the given task, that is, if the agent is getting closer to the goal, the utility becomes higher. This suggests that the observer may understand which goal the agent likes to achieve if the utility of the corresponding task is going higher. The relationship

between an agent and objects such that the agent gets close to the object or the agent faces to a direction is much easier to understand from the observation, and therefore such qualitative information should be utilized to infer what the observed agent likes to do. The information might be far from precise ones, however, it keeps qualitative information and we can estimate well the temporal difference of the utility. If the observer can estimate the utility of each behaviors of the other, it might be possible to recognize the other's intention, therefore the observer not only imitate the observed behavior but also cooperative/competitive behaviors according to the recognized intention.

We propose a novel method not to only learn/execute a variety of behaviors but also to infer the intentions of others supposing that the observer has already estimated the utility (state values in reinforcement learning scheme) of all kinds of behaviors the observed agent can do. The method does not need a precise world model or an accurate coordination transformation system to cope with the problem of view dependency. We apply the method to a simple RoboCup situation where the agent has kinds of tasks such as navigation, shooting a ball into a goal, passing a ball to a teammate, and so on, and the observer judges which goal the agent is now achieving from the observation with estimated utilities. The preliminary experiments are shown and future issues are discussed.

II. Behavior Learning Based on Reinforcement Learning

In this section, reinforcement learning scheme, state value (utility) function and modular learning system for various behavior acquisition/execution are briefly explained.

A. Acquisition of Behavior Utility

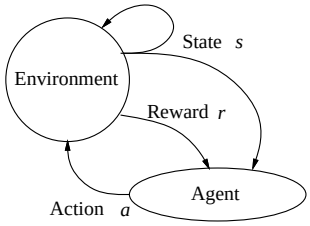


Fig. 1. A basic model of agent-environment interaction

Fig.1 shows a basic model of 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 (utility), discounted sum of the reward received over time under execution of policy π, will be calculated as follows:

$$V(s) = \sum_{t=0}^{\inf} \gamma^t r_t \quad . \tag{1}$$

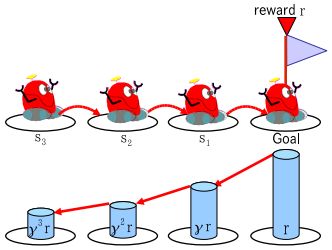


Fig. 2. Sketch of state value propagation

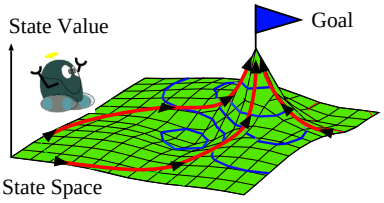


Fig. 3. Sketch of a state value function

Figs.2 and 3 show sketches of a state value function where a robot receives a positive reward when it stays at a specified goal while zero reward else. The state value will be highest at the state where the agent receives a reward and discounted value is propagated to the neighbors states (Fig.2). As a result, the state value function seems to be a mountain as shown in Fig.3. The state value becomes bigger and bigger if the agent follows the policy π. The agent updates its policy through the interaction with the environment in order to receive positive rewards in future.

B. Modular Learning System

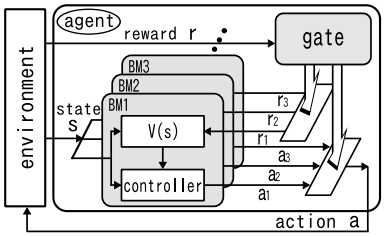


Fig. 4. Modular Learning System

In order to learn/execute a number of behaviors simultaneously, we adopt a modular learning system. Jacobs and Jordan [1] proposed a mixture of experts, in which a set of the expert modules learn and the gating system weights the output of each expert module for the final system output. Fig.4 shows a sketch of such a modular learning system. We prepare a number of behavior modules each of which has already acquired a state value (utility) function for one goal-oriented behavior. A learning module has a controller that calculates an optimal policy based on the utility function. Gating module selects one output from a module according to the agent's intention.

III. Intention Inference by Behavior Utilities

We assume that the observer has already acquired a number of behaviors based on a reinforcement learning method. Each behavior module can estimate utility at arbitrary time t to accomplish the specified task. Then, the observer watches the performer's behavior and maps the sensory information from an observer viewpoint to the agent's one with a mapping of state variables. The behavior modules estimate the utility of the observed behavior and the system selects ones that are increasing their utilities.

A. Basic Idea of Intention Inference

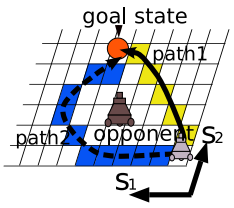


Fig. 5. Sketch of different behaviors in a grid world

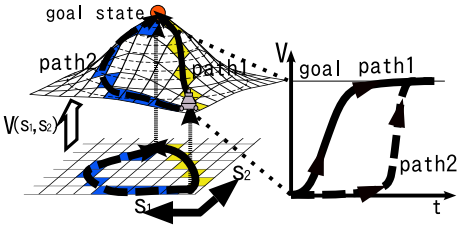


Fig. 6. Inferring intention by the change of state value

Fig.5 shows an example task of navigation in a grid world. There is a goal state at the top center of the world. An agent can move one of the neighbor grids every one-step. It receives a positive reward only when it stays at the goal state while zero else. There are various optimal/suboptimal policies for this task as shown in Fig.5. If one tries to match the action that the agent took and the one based on a certain policy in order to infer the agent's intention, he or she has to maintain various optimal policies and evaluate all of them in the worst case.

On the other hand, if the agent follows an appropriate policy, the utility is going up even if it is not exactly optimal one. Fig.6 shows that the utility becomes larger even if the agent takes different paths. We can regard that the agent takes an action based on one policy when the utility is going up even if it follows various kind of policies.

This indicates a possibility of robust intention recognition even if several appropriate policies can exist for the current task. An agent tends to acquire various policies depending on the experience during learning. The observer cannot practically estimate the agent's experience beforehand, therefore, it needs a robust intention recognition method provided by the estimation of utilities.

B. Intention Inference under Multiple Candidates

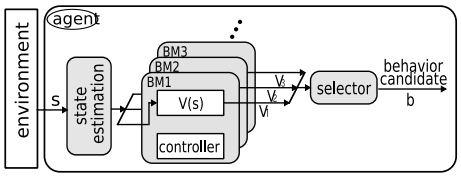


Fig. 7. Behavior inference diagram

At intention inference stage, the system uses the same behavior modules as shown in Fig.7. While an observer watches an behavior of a performer, the system estimates the relationship between the agent and objects such as rough direction and distance of the objects from the agent. Then, each behavior module estimates the utility based on the estimated state of the performer and sends it to the selector. The selector watches the sequence of the utilities and selects a set of possible behavior modules of which utilities are going up as a set of behaviors the performer is currently taking. As mentioned in II-A, if the utility goes up during a behavior, it means the module seems valid for explaining the behavior. The goal state/reward model of this behavior module represents the intention of the agent.

Here we define reliability g that indicates how much the intention inference would be reasonable for the observer as follow:

$$g = \begin{cases} g+\beta & \text{if } V(s_t) - V(s_{t-1}) > 0 \text{ and } g < 1 \\ g & \text{if } V(s_t) - V(s_{t-1}) = 0 \\ g-\beta & \text{if } V(s_t) - V(s_{t-1}) < 0 \text{ and } g > 0 \end{cases},$$

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. We put another condition in order to keep g value from 0 to 1.

IV. Task and Environment

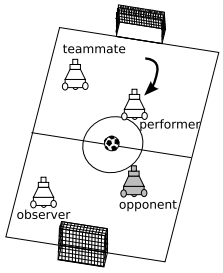


Fig. 8. Environment

Fig. 9. A real robot

Fig.8 shows a situation the agents are supposed to encounter. An agent shows a behavior and the observer estimates the behavior using a set of behavior modules of its own. Fig.9 shows a mobile robot we have designed and built. Fig.10

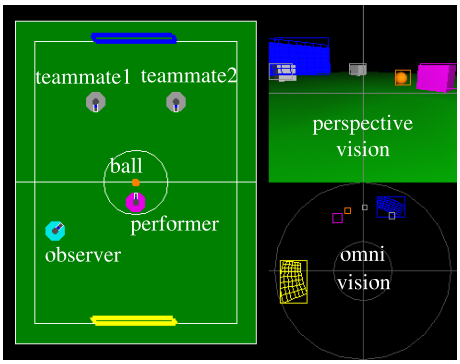


Fig. 10. Viewer of simulator

shows the viewer of our simulator for our robots and the environment. The robot has a normal perspective camera in front of its body. It has an omni-directional camera, however, it doesn't use it, here. A simple color image processing is applied to detect an ball, a performer, and teammates on the image in real-time (every 33ms).

The left of Fig.10 shows a situation the agent encounters while the top right images show the simulated ones of the normal and the bottom right omni vision systems. The mobile platform is an omni-directional vehicle (any translation and rotation on the plane). Table I shows a list of prepared behavior modules and their state variables. The observer has learned the behaviors and its utility estimator based on a reinforcement learning method beforehand.

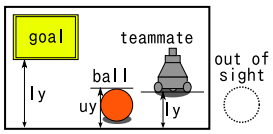


Fig. 11. State variables representing distances to the objects

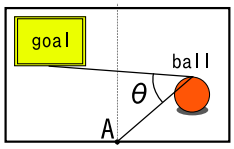


Fig. 12. A state variable θ representing the positional relationship between the objects

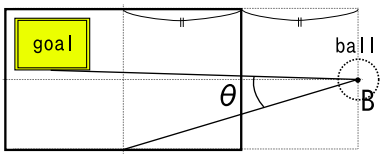


Fig. 13. A state variable θ when one of the objects is out of sight

We prepared the distances of positions of a ball, a goal, and players from the bottom and their relative angles between them on the image of the frontal camera on the robot as state variables. Figs.11 and 12 show examples of those state variables. We divide this state space into a set of region to obtain state id. The space of position value is quantized into 6 subspaces and the space of relative angle between objects into 5 spaces here. Behavior modules define their policy and utility function in this state space.

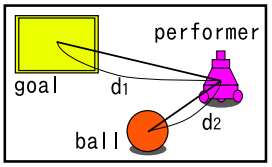


Fig. 14. Estimated state variables representing distances

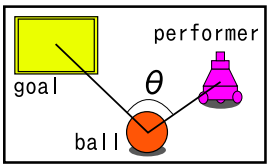


Fig. 15. An estimated state variable θ representing the position relation among objects

When an observer infers an intention of the performer, it has to estimate the state of the performer. Here, we introduce a simple method of state estimation of the performer. Fig.14 shows the estimated distances from the agent and the objects. Fig.15 shows the estimated angle between the objects. The observer uses these estimated state for estimation of utility instead of its own state shown in Figs.11 and 12. These estimated states with this method are far from precise ones, however, it keeps topological information and we can acquire good estimation of temporal difference of utility with this method.

V. Experiments

The observer has learned a number of behaviors shown in Table I before it tried to infer the performer's intention. We gave the observer many experiences enough to cover all exploration space in state space.

A. Same behavior demonstration

After the behavior acquisition, we let the performer play one of the behaviors from Table I and the observer infers which behavior the other is taking. Fig.16 shows an example behavior performed by another agent. The performer showed exactly same behavior that the observer acquired in behavior learning stage, here. The bottom right agent shows "Shoot-Blue" behavior and the top left observer watches the behavior. The observer tries to keep the performer in own perspective during the behavior. Fig.16 (a) shows the sequence of the behavior. Fig.16 (b) and (c) show sequences of estimated utility and reliability of the inferred intentions of the agent, respectively. The green line indicates the behavior of shooting a ball into a blue goal and goes up during the trial. The observer successfully inferred the intention of the performer.

Fig.17 shows an example passing behavior performed by an agent, a sequence of estimated utility of each modules, and a sequence of reliability of inferred intention, respectively. These figures show that the observer can infer the performer's passing behavior (purple line), too. The orange line indicates the reliability of going to a ball behavior and it also goes up during the trial because the agent is continuously approaching to the ball during the trial to pass it to the teammate.

Fig.21 shows an experimental result with real robots. A performer wearing a magenta marker shows a shooting behavior and a observer wearing a cyan marker estimates the behavior successfully.

TABLE I

PREPARED MODULES AND THEIR STATE VARIABLES

Module	State variables
GoToBall	ball position y on the image of perspective camera
GoToYellow	yellow goal position y on the image of perspective camera
GoToBlue	blue goal position y on the image of perspective camera
ShootYellow	ball position y, yellow goal position y, and angle between them θ on the image
ShootBlue	ball position y, blue goal position y, and angle between them θ on the image
PassToTeammate1	ball position y, teammate 1 position y, and angle between them θ on the image
PassToTeammate2	ball position y, teammate 2 position y, and angle between them θ on the image

B. Different behavior demonstration

The observer cannot assume that the performer will take exactly same behavior of the observer even if its intention is same. We prepare other different behaviors for the demonstration of the performer. Fig.18(a) shows an example of the different shooting behavior demonstrated by the performer. The learned behavior by the observer is a zippy motion as shown in Fig.16. On the other hand, the demonstrated behavior is a smoother motion. Therefore, the state transition probability will be different from each other. Figs.18 (b) and 18 (c) show sequences of estimated utility of each modules and a sequence of reliability of inferred intention, respectively. These figures show that the observer can infer the performer's intention of shooting (green line).

C. Comparison with A Typical Approach

In this section, we compare performances between our proposed method and the one based on state estimation using coordinate transformation system and tracing state transition probability. In order to estimate the utility of a behavior module through the observation of the performer, there must be a rough coordinate transformation matrix beforehand. Figs.19 and 20 show a rough sketch of the transformation system. In order to estimate y position on the performer's view image, the observer assumes there are tiles on the floor, maps the positions of an object and the performer, estimates rough distance between them, and maps the distance to the y position on the image of the observer's view. Fig.20 shows a sketch of estimation of direction from the performer to an object. The lower right rectangle shows an example image from the perspective camera and it captures the performer, a ball, and a goal. We assume that it can detect a direction of the performer on the image under a vision system. We put a potential image plane in front of the agent and estimate rough x positions of the objects on the image of the performer's view. Fig.20 shows that the ball is mapped to the left side on the image and the goal to an area of lost to the right side from the camera image. Table II shows the success rate of intention inference of the proposed method and the one with the coordination transformation system. The proposed method shows much better results over the behaviors than the one with the coordination transformation system. "ShootBlueGoal1" indicates a case of inference of shooting behavior identical to the observer's one. "ShootBlueGoal2" indicates a case of inference of shooting behavior but different from the observer's one.

TABLE II

INFERRING INTENTION PERFORMANCES OF THE PROPOSED METHOD AND THE ONE WITH COORDINATION TRANSFORMATION SYSTEM

	Proposed method	Method based on state trans. prob.
ShootBlueGoal1	84%	24%
ShootBlueGoal2	78%	11%
ShootYellowGoal	86%	20%
PassToTeammate1	76%	34%

VI. FUTURE WORK

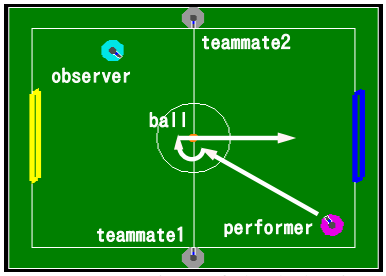
This basic idea can be applied for not only intention inference but also cooperative behavior acquisition. How to define a reward function for cooperative behavior acquisition in multi-agent system is one of the most interesting issues. The proposed method can infer intention of other and estimate the reward/utility of the agent for each step. This indicates that the observer can explore some actions and evaluate how much they will contribute to the other efficiently. Then, it can learn cooperative behavior based on a certain reinforcement learning approach without any heuristic/hand-coded reward function by which it evaluates a reward of itself based on the estimated reward/utility of the other agent.

REFERENCES

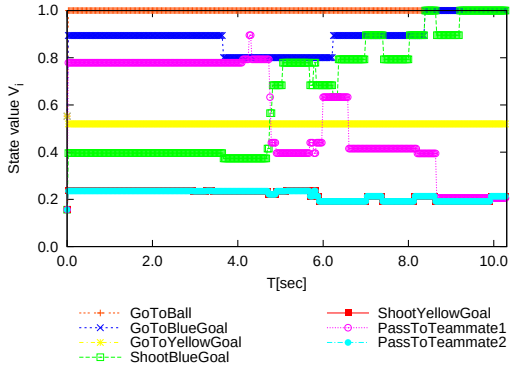
[1] R. Jacobs, M. Jordan, Nowlan S, and G. Hinton. Adaptive mixture of local experts. *Neural Computation*, 3:79–87, 1991.

[2] Doya K., Sugimoto N., Wolpert D.M., and Kawato M. Selecting optimal behaviors based on contexts. In *International Symposium on Emergent Mechanisms of Communication*, pages 19–23, 2003.

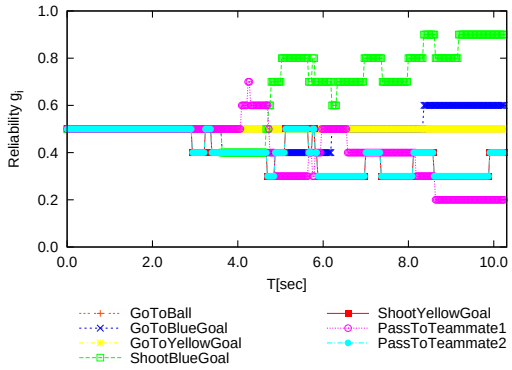
[3] Stefan Schaal, Auke Ijspeert, and Aude Billard. Computational approaches to motor learning by imitation, 2004.



(a) An overview of the behavior

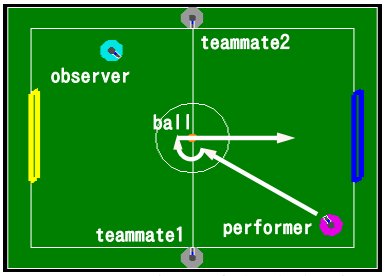


(b) Utility V_i of each behavior module

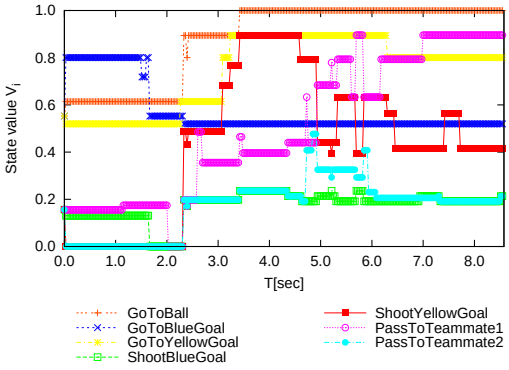


(c) Reliability g_i of each behavior module

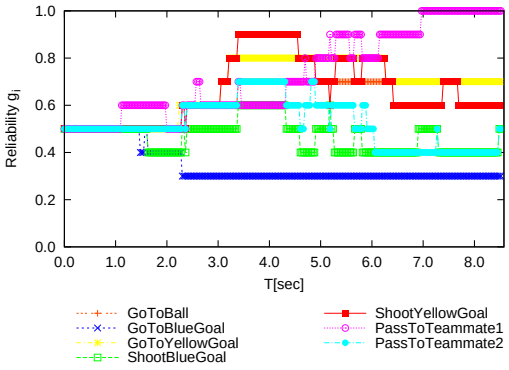
Fig. 16. Result of inferring intention of shooting a ball to the blue goal



(a) An overview of the behavior

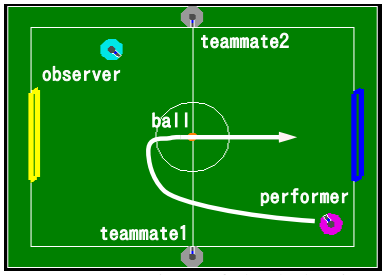


(b) Utility V_i of each behavior module

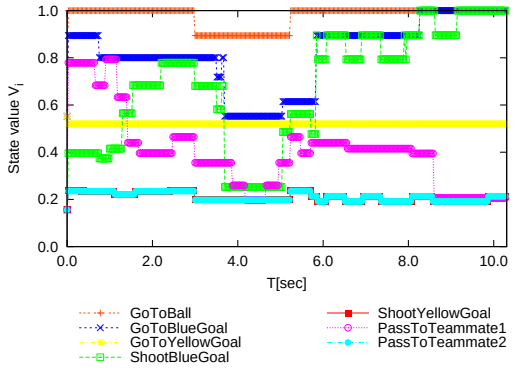


(c) Reliability g_i of each behavior module

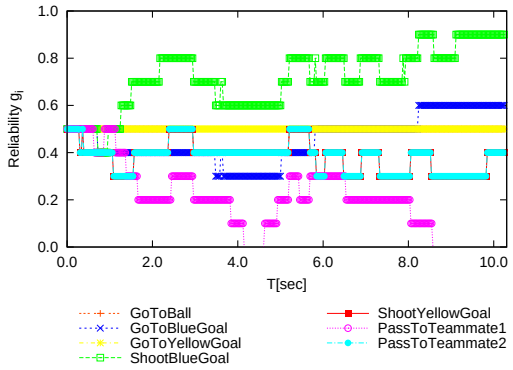
Fig. 17. Result of inferring intention of passing a ball to teammate1



(a) An overview of the behavior



(b) Utility V_i of each behavior module



(c) Reliability g_i of each behavior module

Fig. 18. Result of inferring intention of shooting a ball to the blue goal with different manner

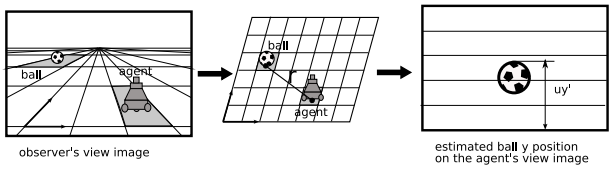


Fig. 19. Estimation of y position of the performer's image

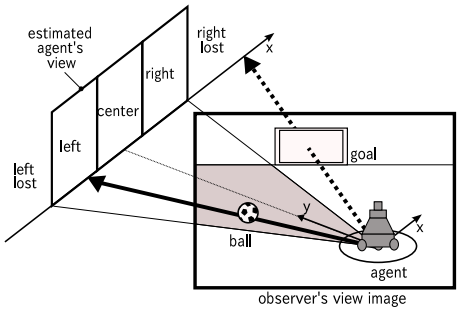
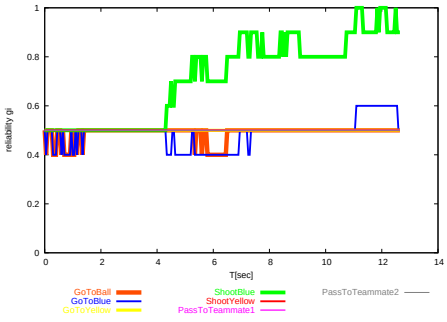


Fig. 20. Estimation of direction from a performer to an object

(a) An overview of the behavior



(b) Reliability g_i of each behavior module

Fig. 21. Experimental result with a real robot system of inferring intention of the performer trying to shoot ball to the blue goal