# Incremental Behavior Acquisition Based on Reliability of Observed Behavior Recognition

**Tomoki NISHI**[1], **Yasutake TAKAHASHI**[1], **Minoru ASADA**[1,2]
[1]**Osaka University,** [2]**JST ERATO Asada Synergistic Intelligence Project**
{tomoki.nishi,yasutake,asada}@ams.eng.osaka-u.ac.jp

## Abstract

We propose a novel approach for acquisition and development of behaviors through observation in multi-agent environment. Observed behaviors of others gives fruitful hints for a learner to find a new situation, a new behavior for the situation, necessary information for the behavior acquisition. RoboCup scenario gives us a good test-bed multi-agent environment where a learner can observe behaviors of others during practices or games. It is more realistic, practical, and efficient to take advantages of observation of skilled players than to discover new skills and necessary information only though the interaction of a learner and an environment. The learner automatically detects state variables and a goal of the behavior through the observation based on mutual information. Reinforcement learning method is applied to acquire the discovered behavior suited to the robot. Experiments under RoboCup MSL scenario shows the validity of the proposed method.

## 1 Introduction

Imitation is one of the most significant ability of a person. The young children increase their repertoire of behaviors and keep advancing it through the imitation of the observed behaviors with interest. Meltzoff et al. insists that the ability of imitation has borne important role for understanding intention and feelings of the other person.

In the environment involving robots and/or humans, it is less necessary to discover new tasks/behaviors only with self exploration than through observations of behaviors of others with skills. The latter is also much useful and realistic in practical view. Furthermore, it is a very useful ability to utilize understanding of intention of humans in order to generate cooperative behaviors with them.

Baldwin et al. [1] showed a fact that a young child of 10-11 months already has a segmentation ability of a behavior through some experiments. In addition, they insist that young children seem to have a low level model with continuity of trajectory of the body and segment the observed trajectory where it deviates from the model as the break of the behavior because of the fact that they can do the behavior segmentation even if the observation is the first time for them. Itti and Baldi [2] did the experiment regarding visual features which induce the gaze of a person. A vision image is divided spatially into small regions and temporally short periods and the feature quantity such as difference of color strength of red and green are modelled dynamically in regard to the each regions. Then, they showed that the human gaze is induced by the regions where the changes of parameter values of the model are large. From those insights, segmentation of an observed behavior seems to be done by detecting a remarkable point as a break point of the behavior where parameter values of the model change largely.

We propose a novel approach for incremental acquisition and development of behaviors through detection of remarkable points of observed behaviors. and apply it to our robots. A local linear model is introduced to check continuity of trajectory concerning each sensor value and

a point with a big error of this model is regarded as remarkable point for segmentation of the observed behavior. A new behavior learning module is assigned autonomously to a novel segmented behavior. Reinforcement learning method is applied to acquire the new behavior suited to the robot. Experiments under RoboCup MSL scenario show the validity of the proposed method.

## 2 Related Work

Research regarding imitation through observation has been done so far [3, 4, 5]. Almost conventional work focuses on efficient reproduction of observed behaviors by following trajectory of an observed behavior of a demonstrator. Those proposed imitation methods have applicability limitation as a trajectory of its imitated behavior becomes almost same of the instructed behavior because the imitated behavior is evaluated not by the intention of the behavior but by the similarity of the trajectory itself. The imitation with reproducing the observed trajectory is called mimicry as known as the most primitive imitation.

Expanded definition of imitation of the young child includes this mimicry, emulation, and narrow defined imitation. Emulation is when after observing an action, the observer jumps to conclusions and performs only those actions that will lead it to the goal, without caring about the exact methods of the demonstrator (although observed methods biases future actions). Finally, imitation is the crowning of copying, the sophisticated capability of reenacting sequence of actions to detailed levels, with the agent clearly aiming for the same objective as the demonstrator's.

Capability of emulation is useful for intention recognition because it is important to reproduce the result of the observed behavior but not about the exact trajectory of the observed behavior. It is unrealistic in the real world to acquire precise trajectory of an observed behavior because of the sensor/actuator noises or any possible differences in the parameters of body between the observer and the demonstrator and/or objects. Takahashi et al. [6] proposed a method of emulation that does not use similarity of the trajectory and does infer the intention based on the increase and decrease of achievement of the observed behavior. They showed the validity of the proposed method to infer the intention of other even if the trajectory of observed behavior is different from the one of own behavior of itself.

Reinforcement learning [7] has been studied well for motor skill learning and robot behavior acquisition. It generates not only an appropriate policy (map from sensor outputs to motor commands) to achieve a given task but also an estimated discounted sum of reward that will be received in future while a learning agent is taking an optimal policy. But it is known well that learning time and the required computational resources for a simple application to a real robot tends to be too huge and almost unpractical. One of the potential solutions might be application of so-called "mixture of experts" proposed by Jacobs et al. [8], in which a whole state space is decomposed to a number of areas so that each expert module can produce good performance in the assigned area, and one gating system weights the output of the each expert module for the final system output. This idea is very general and has a wide range of applications [9, 10, 11]. Therefore, emulation can be achieved based on reinforcement learning so that the observed behavior is divided into a number of modules and a reward is given when the result of behavior is reproduced. In general, it is a difficult problem to design appropriate combination of behaviors beforehand and it is desirable to be done autonomously by the observer itself.

Many kinds of modular learning systems with autonomous behavior segmentation mechanisms are proposed so far. Samejima et al. [4] arranged modules of a linear prediction model and a controller of reinforcement learning method as group in parallel, changed those assignment adaptively based on prediction error of the prediction models. Taniguti et al. [12] also proposed a system with a set of reinforcement learning modules in parallel that splits and merges among them based on the prediction error of reinforcement signal. In these systems, the state space, the space which describes the relationship between a learning agent and an environment, and a reward function have to be defined beforehand. Unfortunately, it is also difficult in general to design a state space and a reward function appropriately because it depends on not only behaviors the learner will observe and acquire in future but also the sensors and the motion mechanism equipped on it. We need some mechanism to find an appropriate state space and reward function automatically for each segmented behavior through observation of instructed behaviors.

## 3 Observed Behavior Segmentation based on Remarkable Points

### 3.1 Basic Idea

From the insights of Baldwin et al. [1] and Itti and Baldi [2], it seems to be possible to segment an observed behavior properly at remarkable points where parameter values of a simple model changes largely during the observation as a human tends to look at the points carefully. Since a trajectory of one single behavior tends to have stable direction and speed, then a local linear model can be applied to fit the trajectory. On the other hand, because linearity of a trajectory breaks when continuity of trajectory breaks, the remarkable point can be easily found by measuring change of reliability of the local linear model parameters. We call this measurement as "degree of attention" in this paper. In addition, an important space and a target state in order to explain the observed behavior can be found based on the mutual information between the remarkable points and the state in the space because a remarkable point is also a position of the target state of the behavior.

On the basis of argument above, our method,

1. finds remarkable points of observed behavior based on degree of attention,
2. segments observed behavior based on the remarkable points, and
3. assigns a behavior learning module to each segment of the observed behavior.

If observed behavior can be emulated by an appropriate module in a behavior repertory, then the degree of attention is suppressed so that acquisition of only novel behaviors for the learner can be focused on. An hierarchical learning system integrates a number of small time-scale behaviors so that the observer can emulate a long time-scale behavior. The hierarchical learning system has reinforcement learning modules that acquire purposive action policy through trial and error manner. Another advantage of this hierarchical learning system is efficient re-usability of behaviors learned before and enables the observer to keep learning new behaviors while it accumulates useful ones.

### 3.2 Algorithm Overview

A learner tries to emulate observed behavior and cumulatively acquire behaviors by the procedure below:

1. Observe behaviors of other,
2. Detect remarkable points in the observed behavior
3. If there is at least one remarkable points, then, go to next, else, go to 6.
4. Segment the observed behavior into smaller ones based on the remarkable points,
5. Find a state space and a goal state for the segmented behavior based on degree of attention.
6. If there are more than one observed behavior, then generate another learning module at higher layer to coordinate them.

We adopt a hierarchical learning system proposed by Takahashi et al. [9] Because of limited space of paper, we eliminate the details of the hierarchical learning system and concentrate on the acquisition of new behaviors from the observation.

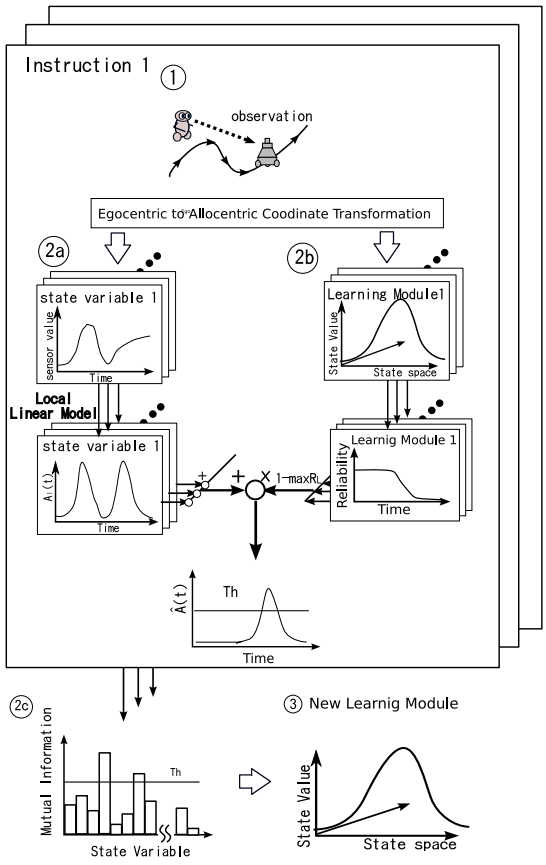



Figure 1: Sketch of Algorithm of New Behavior Detection

Fig. 1 shows a sketch of algorithm of new behavior detection and acquisition. From a number of observation data, degree of attention and reliabilities of existing behavior are calculated. The degree of attention suppressed by the reliabilities is used for selection of state

variables related to a new behavior acquisition based on mutual information. A reinforcement learning modules is assigned with a new generated state space with the selected state variables and a goal state, then, acquires the behavior through trials and errors. The details are explained in following sections.

### 3.3 Remarkable Point and Degree of Attention

As mentioned about the insights from the work of Baldwin et al. and Itti and Baldi before, we introduce a concept of degree of attention. The degree of attention should be large if the continuity of the trajectory breaks. We adopt a local linear model to detect a remarkable point of the continuity of the observed trajectory.

One dimensional local linear model $x = at + b$ is introduced and we assume that variation of measurement values follows normal distribution, then, variance of error of parameter is presumed as below.

$$\sigma_0^2 = \frac{1}{N-2}(\sum x_i^2 - b\sum x_i - a\sum t_i x_i) \quad (1)$$

$$\sigma_a^2 = \frac{N}{N\sum t_i^2 - (\sum t_i)^2}\sigma_0^2 \quad (2)$$

$$\sigma_b^2 = \frac{\sum t_i^2}{N\sum t_i^2 - (\sum t_i)^2}\sigma_0^2 \quad (3)$$

where N is number of samples during observations. We define a reliability $R_m(t)$ of the parameter of the local linear model $m$ at time $t$ as follows:

$$\sigma^2(t) = \sqrt{(\sigma_a^2(t))^2 + (\sigma_b^2(t))^2} \quad (4)$$

$$R_m(t) = \begin{cases} -\sigma^2(t) & \text{if} \quad \sigma^2(t) < 1 \\ & \text{else} \end{cases} \quad (5)$$

This reliability has a high value if the observed data has good linearity. Then we apply one dimensional local linear model on each state variable and define the degree of attention $A(t)$ at of time $t$ as total sum of the changes of the reliabilities of all models. That is,

$$A(t) = \sum_{m \in M} |R_m(t+1) - R_m(t)| \quad . \quad (6)$$

We define the remarkable point where the degree of attention $A(t)$ is larger than a threshold $k$ as the part with big change of the parameters of the all models.

### 3.4 Suppression of Degree of Attention

If a behavior has been already assigned before, then the assignment of a new behavior module should be suppressed even if the degree of attention is high. State value $V(t)$ which is utilized with reinforcement learning represents the closeness to a goal of the behavior. If a demonstrator follows to a policy of the behavior, the state value keeps rising, while it shows a movement in the opposite direction, then, the state value tends to decrease. Reliability that has higher value when the state value is rising and lower else is introduced here. The degree of attention is suppressed by this reliability (Fig.1 2b)

The reliability $R_l(t)$ of a learning module $l$ at of time $t$ is defined as follow:

$$R_l(t) = \frac{1}{1 + \exp(-k_1 e(t))} \quad (7)$$

$$e(t) \mathrel{+}= \begin{cases} 0 & \text{if} \quad e(t-1) > k_2 \\ & \text{or} \quad e(t-1) < -k_2 \\ V(t) - V(t-1) & \text{else,} \end{cases} \quad (8)$$

where $k_1$ and $k_2$ are a gradient factor of sigmoid function and maximum value of $e(t)$, respectively. Initial value of the reliability is 0.5, that is, $e(0) = 0$ in this paper.

This reliability $R_l(t)$ can evaluate how the observed behavior follows the policy of the module. In other words, if the reliability of this learning module is high, this means that the observed behavior has been already acquired in advance. Then degree of attention is suppressed as below on the basis of the reliability of the existing learning modules. The suppressed degree of attention $\hat{A}(t)$ is calculated as

$$\hat{A}(t) = (1 - \max_{l \in L} R_l(t))A(t) \quad , \quad (9)$$

where $\max_{l \in L} R_l(t)$ is maximum of reliabilities of all existing learning module acquired.

### 3.5 Selection of a state space and a goal state

Fig. 2 shows a diagram of selection of state space and a goal state for learning a new behavior based on the suppressed degree of attention $\hat{A}(t)$. The mutual information $I(X;Y)$ is information gain of the phenomenon $Y$ when the phenomenon $X$ is observed and shows depth of the relation of two phenomena. A new state space for an observed behavior is selected as the space with most mutual information gain between the phenomena "$\hat{A}(t)$ is higher than a threshold" and the one "The state $s$ takes place in the state space $S$". Then, a new behavior module is assigned to the state space. The concrete procedure of selection of a state space and a goal state is shown below:

1. Create a histogram $H_1$ of appearance frequency of state visited through observed behaviors

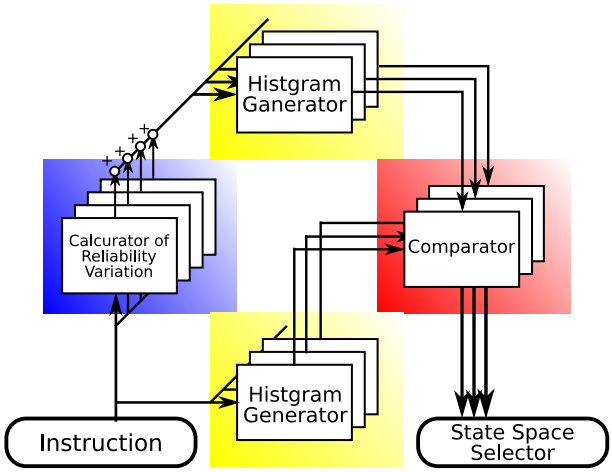



Figure 2: Sketch of a system which selects state space

2. Create a histogram $H_2$ of appearance frequency of state where the $\hat{A}(t)$ is larger than a threshold

3. Calculate the mutual information gain between phenomena "$\hat{A}(t)$ is higher than the threshold" and the one "the state $s$ takes place in the state space $S$" based on the histogram $H_1$ and $H_2$

4. Assign a new behavior learning module with the state space and the goal state which appear most frequently in the histogram if the mutual information gain is larger than a threshold

## 4 Experiments

### 4.1 Experimental setup

The proposed method was verified with real robots under a scenario of RoboCup MSL. A robot has one omnidirectional vision system and detects objects around it in all direction simultaneously. It also has a omnidirectional vehicle to go to all direction and turn around on floor. Additionally, a kick mechanism is attached on the robots. There are a passer, who shows instruction behaviors, a receiver, a ball, two goals, and an observer (learner) in the environment.

### 4.2 Passing Behavior Observation and Acquisition

In our experiments, the learner observed a passing behavior 41 times from different viewpoints. 41 times is experimentally enough for the task. Fig. 3(a) and (b) show an example of the observation situation and a sequence of major sensor values during the observation respectively. Red, blue and the green lines in figure Fig. 3(b) indicate distance between the ball and the receiver, relative angle between the ball and the receiver from the viewpoint of the passer, distance between the ball and the passer, on the image of the observer, respectively. These values are normalized accordingly. From this figure, the passer starts dribbling from around 150th step toward the receiver and kicks a ball to the receiver at approximately 220th step. Then, the receiver received the ball at around 230th step.

Degree of attention $\hat{A}(t)$ is calculated through all 41 times observations. Fig. 4(a) shows a sequence of degree of attention during the observation of the instruction. $\hat{A}(t)$ is calculated with all observed behavior and the mutual information between each state variable and the region where the $\hat{A}(t)$ is over than a threshold 0.2 is graphed in Fig. 4(b). Two new behavior modules, LM1 and LM2, that have state spaces and goal states where are highly related with a space where $\hat{A}(t)$ is larger than a threshold based on mutual information is assigned. Table 1 shows the state space and the goal state. LM1 acquired a behavior of approaching a ball while LM2 acquired a behavior of turning around the ball and facing them in front of the body. Fig. 5(a) and (b) show examples of the acquired behaviors by learning modules LM1 and LM2. Whole pass behavior is acquired by integrating these behavior modules using a hierarchical reinforcement learning mechanism. One example of the acquired passing behavior is shown in Fig. 5(c).

### 4.3 Shooting Behavior Observation and Acquisition

As the second instruction, the learner observed shooing behavior 41 times, again. The relationship between a state space and a region which $\hat{A}(t)$ is larger than a threshold was calculated based on mutual information, again. Fig. 6(a) shows an example behavior acquired by a new module with the new state space. Fig. 7(a) and (b) show a sequence of degree of attention $\hat{A}(t)$ during the observation and the mutual information between each state variable and the region where the $\hat{A}(t)$ is over than a threshold, respectively. A behavior of approaching a ball is necessary for this observed behavior and this behavior has been already acquired as LM1 through passing behavior, then, LM1 does not need additional learning stage. A new behavior module LM3 is generated with an appropriate state space and a goal state shown in Table 1. The shooting behavior is acquired as the integration of LM1 and LM3 with hierarchical reinforcement learning mechanism. One example of the

Table 1: List of state variables and goal state in acquired learning modules

| learning module | state variable | center of goal state |
|---|---|---|
| LM1 | distance to ball | 0.05 |
| | angle to ball | 0.00 |
| LM2 | angle between ball and receiver | 0.00 |
| | angle to ball | 0.00 |
| | distance to ball | arbitary |
| LM3 | angle between ball and goal | 0.00 |
| | angle to a ball | 0.00 |
| | distance to a ball | arbitrary |

(a) behavior of LM1 (b) behavior of LM2 (c) integrated behavior

Figure 5: Examples of acquired behaviors of LM1 and LM2 and integrated one

behavior is shown in Fig. 6(b).

## 5 Conclusion and Future work

We proposed a novel approach for acquisition and development of behaviors through observation in multi-agent environment. The learner automatically detected state variables and a goal of the behavior through the observation based on mutual information. Experiments under RoboCup MSL scenario showed the validity of the proposed method. The learner observed passing and shooting behaviors and tried to imitate them by incremental skill acquisition such as LM1, LM2, and LM3. Future work will investigate more number of typical behaviors like "obstacle avoidance", "receiving a ball", "interfering the opponents" and so on in RoboCup games.

## 参考文献


[1] Dare A. Baldwin, Jodie A. Baird, Megan M. Saylor, and M. Angela Clark. Infants parse dynamic action. *Child Development*, Vol. 72, No. 3, pp. 709–717, May/June 2001.

[2] Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, June 2005.

[3] Yuichiro Yoshikawa, Minoru Asada, and Koh 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*, pp. 1416–1427, 2001.

[4] Kazuyuki Samejima, Kenji Doya, and Mitsuo Kawato. Mosaic reinforcement learning architecture: Symbolization by p redictability and mimic learning by symbol. *Journal of the Robotics Society of Japan*, Vol. 19, No. 5, pp. 551–556, 2001.

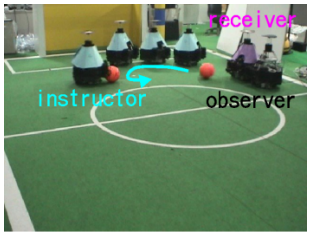
[5] Aude Billard and Maja J. Mataric. Learning human arm movements by imitation: evaluation of a biologically inspired connectionist architecture. In *Robotics and Autonomous Systems*, Vol. 941, pp. 1–16, 2001.

[6] Y.Takahashi, 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*, pp. 25–31, 2006.

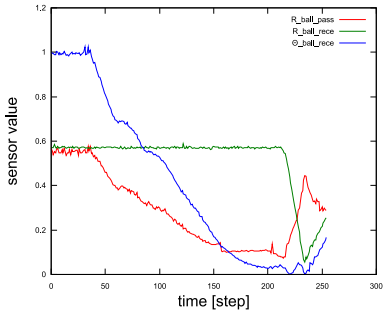
[7] R. Sutton and A. Barto. *Reinforcement Learning : An Introduction*. MIT Press, 1998.

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

[9] Y.Takahashi and M.Asada. Multi-controller fusion in multi-layerd reinforcement learning. In




(a)




(b)

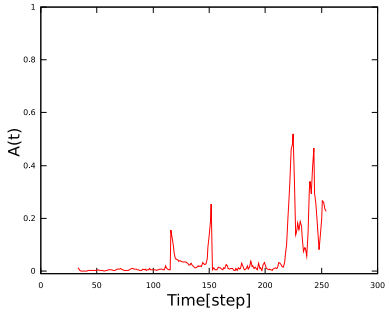
Figure 3: (a)Instruction of passing behavior, (b)Example sequence of the major sensor values ;((red line): distance between receiver and ball, (blue line): angle between receiver and ball, and (green line): distance between passer and ball)

*IEEE/RSJ International Conferrence on Multisensor Fusion and Integration for Intelligent Systes(MFI2001)*, pp. 7–12, 2001.

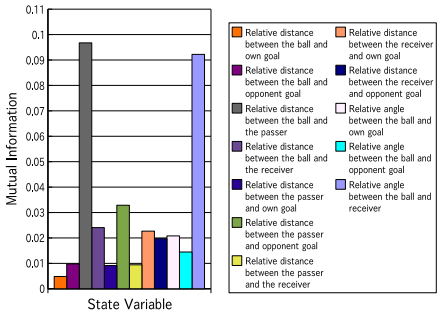
[10] Jun Morimoto and Kenji Doya. Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. In *Proceedings of International Conference on Machine Learning*, pp. 623–630, 2000.

[11] Satinder P. Singh. The efficient learning of multiple task. *Neural Information Processing Systems*, Vol. 4, pp. 251–258, 1992.

[12] T. Taniguchi and T. Sawaragi. Incremental acquisition of behavioral concepts through social interactions with a caregiver. In *Artificial Life and Robotics (AROB 11th '06)*, 2006.

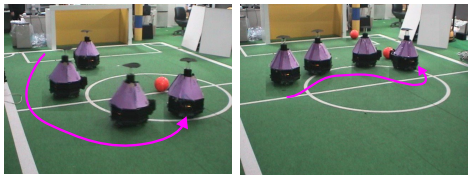



(a)



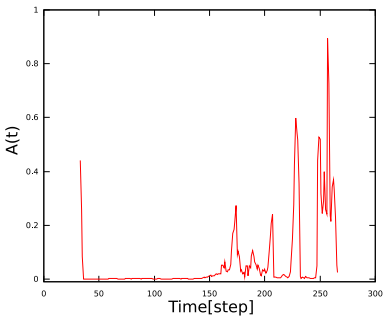

(b)

Figure 4: (a) Sequence of degree of attention during observation and (b) Mutual information in each state variable

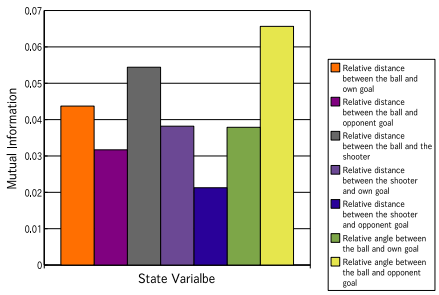


(a) behavior of LM3 (b) integrated behavior

Figure 6: Example of acquired behavior of LM3 and integrated one for shooting a ball




(a)




(b)

Figure 7: (a) Sequence of degree of attention during observation of shooting behavior and (b) Mutual information in each state variable