

状態価値に基づく 人間行為観察と自己行動学習の循環的発達

Mutual Development of Behavior Acquisition and Recognition based on State Value

田村 佳宏	高橋 泰岳	浅田 稔
Yoshihiro Tamura	Yasutake Takahashi	Minoru Asada
大阪大学	大阪大学	大阪大学, JST ERATO
Osaka University	Osaka University	Osaka University, JST ERATO

Abstract: Both self-learning architecture (embedded structure) and explicit/implicit teaching from other agents (environmental design issue) are necessary not only for one-shot behavior learning but more seriously for life-time behavior learning. This paper presents a method for a robot to understand unfamiliar behavior shown by others 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). That is, the state value update can be accelerated by observation without real trials and errors while the learned values enrich the recognition system since it is based on estimation of the state value of the observed behavior. The validity of the proposed method is shown by applying it to a dynamic environment where a human and two robots play soccer.

1 はじめに

ロボット工学における一つの目標として、人間と共生できるロボットの開発が挙げられる。ただし共生環境においては、人間の様々な要求に答えるため、ロボットはその生涯を通して行為を獲得し続けていく必要がある。環境変動に対する適応性の観点から試行錯誤を通して自身で行為獲得する強化学習¹⁾のロボットへの適用研究が多く行われている。強化学習は試行錯誤を通して自律的に報酬の期待値を最大化する行動則を獲得する枠組である。しかし、様々な行為を自分自身の経験のみで学習するには、膨大な学習時間を要する。

一方で近年、神経生理学において自己の行為実行時と他者の行為観察時でほぼ同じ活性パターンを示すミラーニューロンの存在を示唆する実験が報告されている²⁾。これは自己の行動学習と他者の行為推定とが相互に強く関連している可能性を示していると考えられる。実際、他のロボットや人間との共生を行う環境(マルチエージェント環境)下では、自身の試行錯誤のみで行為を獲得する必要は無く、むしろ他者の行っている未経験の行為の観察を通して行為を獲得する方が現実的である。他者行為の観察により行動学習を行うことで学習が加速し、自分自身の経験のみで学習するよりも効率の良い学習ができる^{3, 4, 5)}。しかし、教示行為が明示的に示されない場合でも、観察者側が自律的に観察した行為を認識し、自身の行為獲得にフィードバックしていくことが望ましい。

Takahashi et al.⁶⁾は、観察者が予め提示される行為の状態価値を予め獲得している場合に、この状態価値を利用して他者行為のロバストな認識が可能であることを示した。これは自己行為の状態価値から行為獲得と行為認識の両方が導かれるということを示している。田村ら⁷⁾は、状態価値を利用し、他者行為認識

と行動学習を循環させることで、行為理解が効率的に安定して発達することをコンピュータシミュレーションで示した。これはロボットが分類が明示的でない行為を自分で分類し、その行為の学習に観察経験を活かすことで学習が加速するというものである。さらに循環構造になっていることから、自身の学習経験が増えるにつれて、分類の精度も向上していくので、学習が加速し続ける。しかし、このシミュレーションでは観察する他者の行為は実演者が一つずつ分けて行っており、複数の行為を連続且つランダムに行った場合の提案手法の有効性は確認されていない。

本論文では、人間が複数の行為を連続且つランダムに行った場合であっても、認識と学習の循環により行為理解が効率的に安定して発達することを示す。

2 状態価値に基づく行為認識と行為学習

2.1 概要

環境中には学習者としてのロボット、そして実演者としての人間が存在する。行動発達において、自分と他者との間の同等性の認識が必要であることが示唆されており⁸⁾、これを強化学習の枠組で捉えると、学習者は実演者の行為の報酬や状態価値を推定することになると考えられる。また、学習者にとって実演者の機構や行動の種類、出力しているモータ情報、得ているセンサ情報等は未知であり、実演者は学習者に対して明示的な教示信号等を出さずに、行為を実行すると仮定する。

Fig.1に我々の提案する手法の簡単な概念図を示す。学習者はまず他者行為を認識し、次にモデル規範型の強化学習を用いて行動学習をする。そしてまた他者行為の認識を繰り返す。Fig.1は一つの行為のみの概念図であるが、認識及び学習する行為は複数あるため、学習者は複数の行為学習器を使って認識、そして行動

学習を行う。 $V(s)$ と $\hat{V}(s)$ はそれぞれ自分自身の（行動）経験に基づいて更新された状態価値、観察を通して推定された状態価値である。目標状態に辿り着いた時だけ正の報酬を受け取る場合、状態価値の直観的な意味は目標状態への近さとなる。そのため、ある目標状態に向かって行動を行うとき、その状態遷移系列の状態価値は概ね上昇する傾向にある。よって実演者の行為を観察して得られた状態遷移系列を学習者自身の持つ状態価値関数によって状態価値に写像し、状態価値が上昇した行為学習器を選択することで、実演者からの明示的な教示無しで自律的に行為を分類し、自身の行為獲得にフィードバックをかけることができる。しかし、実演者の動作系列が学習者の実現可能なものである保証が無い場合、観察で得られた状態価値関数をそのまま自身の行動価値関数として使うことはできない。そこで、自身の経験度合に応じて、観察で得られた状態価値関数の値と自身の状態価値関数の値とを比較し、大きい方の値をその状態における状態価値として状態価値関数を更新することでこの問題に対処する。これにより、他者行為の観察を通さない場合よりも行為認識、行為獲得の性能が早く向上し、行為理解が効率的に安定して発達することを示す。

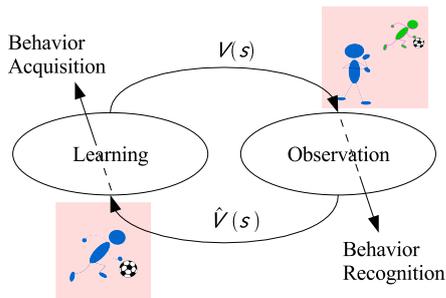


Fig.1 Spiral growth through learning and observation of behaviors

2.2 行動学習器

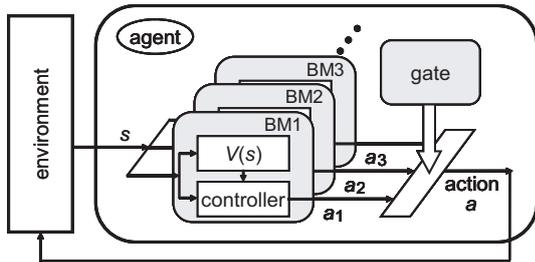


Fig.2 Modular learning system

本研究で用いるモジュール型学習機構は Fig.2 に示すように、行為モジュールとゲートで構成されている。学習者は複数の行為モジュールを持っており、1つの行為モジュールは1つの行為に対応している。行為モジュール (BM: Behavior Module) は環境から状態 s を入力として受け取ると、状態価値 $V(s)$ を基に最適な行動を決定し出力する。各行為モジュールが出力した行動は、学習している行為に応じてゲートによって適切

に選択され、学習者の最終的な行動として出力される。状態価値関数は状態遷移モデルと報酬モデルによって計算される。状態遷移モデルは、ある状態 s で行動 a を選択し、次状態が s' となる確率である状態遷移確率

$$\hat{P}_{ss'}^a = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (1)$$

の推定器を含む。このモデルは観察者が環境と相互作用することによって構築される。また報酬モデルはある状態 s で行動 a を実行し、次状態 s' に遷移した際に期待される報酬

$$\hat{R}_{ss'}^a = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \quad (2)$$

の推定器を含む。状態遷移確率 $\hat{P}_{ss'}^a$ と報酬 $\hat{R}_{ss'}^a$ が決まると、ある状態 s で、行動 a を取った場合の行動価値関数 $Q(s, a)$ 、及び状態価値 $V(s)$ は、

$$Q(s, a) = \sum_{s'} \hat{P}_{ss'}^a \left[\hat{R}_{ss'}^a + \gamma V(s') \right] \quad (3)$$

$$V(s) = \max_a Q(s, a) \quad (4)$$

で与えられる。ここで γ は減衰係数を表す。つまり、ある状態 s で最大の $Q(s, a)$ をとる行動 a を選択することで最適方策を得るということである。

2.3 他者行為認識

例としてロボットがボールに近づくというタスクを考える。状態 s はロボットとボールの相対位置座標 (s_1, s_2) で構成されるとする。ここでロボットが獲得した最適方策による状態遷移系列、すなわち位置座標の系列は path1 のようなものであるのに対し、他者の行う行為は path2 のような系列であったとする。path1 と path2 は状態遷移として比較すると大きく異なるが、Fig.3 で示すように、状態価値関数 $V(s)$ によって状態価値に写像し、その時間変化を見ると、path1 も path2 も状態価値が上昇するという傾向にあるという点においては同じであると言える。そこで行為認識確信度と呼ぶパラメータを導入し、状態価値の上昇・減少に応じて行為認識確信度の値を変化させることで、行為認識が可能である。

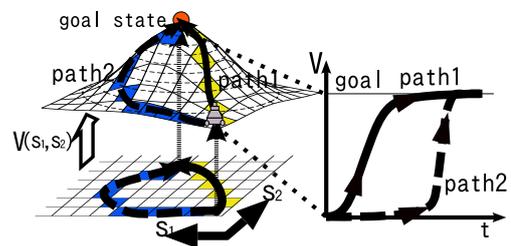


Fig.3 Behavior recognition based on the change of state value

学習者と実演者の視点は観察時では違うため、学習者は環境から自己視点における観測情報を得て、何らかの方法で他者視点における観測情報へと変換することで、他者の状態を推定し、自身の状態価値関数によって状態を状態価値に写像して他者行為の状態価値を推定する。

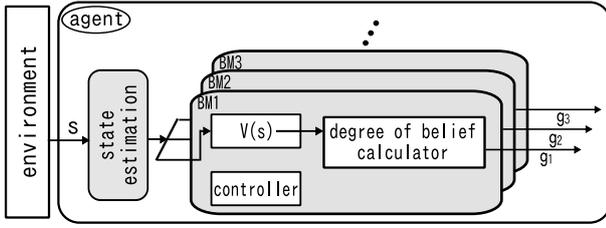


Fig.4 System for behavior recognition

行為認識システム図を Fig.4 に示す．行為認識システムは複数の行為モジュール，状態推定器で構成されている．各行為モジュールは推定された状態を受け取り，自身の持つ状態価値関数によって状態価値に写像し出力する．時間勾配から算出した行動意図の行為認識確度 (reliability) を基に，もっともらしい他者の行為を推定する．各行為モジュール i の行為認識確度 g_i は

$$g_i = \begin{cases} g_i + \beta & (V_i(s_t) - V_i(s_{t-1}) > 0, g_i < 1) \\ g_i & (V_i(s_t) - V_i(s_{t-1}) = 0) \\ g_i - \beta & (V_i(s_t) - V_i(s_{t-1}) < 0, g_i > 0) \end{cases} \quad (5)$$

とする． β は更新度であり，本研究の実験では 0.1 としている．ここで $V_i(s_t)$ は時刻 t の状態 s_t における行為モジュール i の状態価値を表す．よって行為モジュールの状態価値が増え続けるほど，行為認識確度は大きな値となる．最も行為認識確度の大きくなった行為を他者行為として認識する．なお，本実験では全ての行為認識確度の初期値は 0.5 とした．

Fig.5(b), (c) は，Fig.5(a) に示すように実演者 (Cyan) が観察者 (Magenta) にボールを運ぶという行動を取ったときに観察者が推定した状態価値，行為認識確度のグラフである．どちらのグラフにおいてもパス行動の線は行為観察の間中，増加傾向を保っている．またこのパス行動はまずボールに近付き，その後味方に近付くといった構成になっている．そのため，まずボールアプローチの状態価値，行為認識確度が上昇し，その後チームメイトアプローチの状態価値，行為認識確度が上昇している．

2.4 観察による行動学習

2.4.1 他者行為認識

観察により推定される状態価値関数 $\hat{V}^o(s)$ は，時刻 t において推定される実演者の状態を s_t^o とすると式 (6) のようになる．この $\hat{V}^o(s)$ が学習者の状態価値更新のバイアスとして使われる．

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

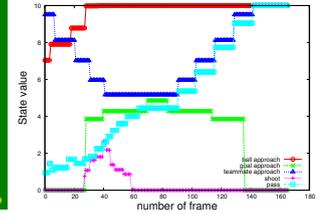
次に学習する行為を推定する 2 つの条件を示す．

- (1) 状態が遷移して行為認識確度が上昇したとき，または行為認識確度が 1 のとき
- (2) (1) 以外の状態で実演者に報酬が与えられたと推定できたとき

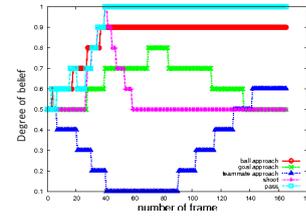
実演者からは明示的な教示が無いため，学習者は 2.4 節で示した行為認識確度によって行為の認識を行う



(a) A behavior of pushing a ball to the magenta player



(b) Estimated Values



(c) Degree of Belief

Fig.5 Sequence of estimated values and reliabilities during a behavior of pushing a ball to the magenta player

が，自身の状態価値が不完全であった場合，1 つめの条件だけでは行為認識が上手くいかない．そこで，2 つめの条件を追加することで行為の推定を行う．

2.4.2 行動学習

自身の学習によって得られた状態価値と観察によって推定された状態価値との比較により，次の状態 s の状態-行動価値関数 $Q(s, a)$ の更新にどちらの値を使用するか決定する．式 (7) に示すように，もしある状態 s において自身の学習によって得られた状態価値 $V(s)$ が観察によって推定された状態価値 $\hat{V}^o(s)$ より大きければ，自身の学習によって得られた状態価値 $V(s)$ を使用し，そうでなければ観察によって推定された状態価値 $\hat{V}^o(s)$ を使用する．これによりたとえ自身の行動学習で状態価値を得ていない状態であっても状態価値を推定することができるため，学習者が行為の目標状態近くの空間を効果的に探索できるようになる．しかし，観察によって推定された状態価値関数は常に学習者にとって適切な値を持つ状態価値関数とは限らない．そこで式 (8) に示すように，その状態を経験した回数 n を使用し． n の値が大きければ大きいほど推定された状態価値を減衰させ，値を小さくし，自己の学習経験による状態価値を優先させるようにする． η は経験回数による減衰係数であり，本実験では 0.9 とした．

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

ただし

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

$$\hat{V}^o(s) = \eta^n \hat{V}^o(s) \quad (8)$$

3 実験

実験はコンピュータシミュレーションで行った。シミュレーションで想定しているロボットはロボカップ中型リーグに出場しているロボットである。このロボットは移動機構として全方位移動機構、視覚センサとしてロボットの上部に全方位カメラ、正面部に通常のカメラを備えている。このためロボットは、常に周囲の物体を認識することが可能であり、2次元平面上においてどの方向にも並進及び回転することができる。

環境中には学習者、実演者(人間)があり、オブジェクトとしてボールが1つ、そしてゴールが2つ存在する。Fig.6のように、学習者は実演者である人間の動きを観察する。実演者は複数の行為をランダムに連続して一定時間行い、学習者に観察させる。

また比較実験の初期条件を合わせるため、フィールドを3×3の9個の領域に分割し、ロボットやボールはそれらの領域の中心に配置することで実験を行う。



Fig.6 An example of observation

獲得する行為及び学習に必要な状態変数を Table.1 に示す。また状態変数の説明は Fig.7 に示す。

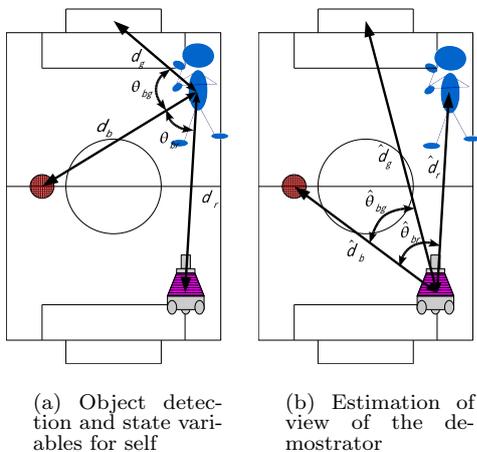


Fig.7 Estimation of view of the demonstrator

Table 1 List of behaviors 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}

3.1 シミュレーション

3.1.1 実験手順

人間行為の観察を通して学習することが、行為獲得と行為認識の性能にどれくらい効果があるのかを調べるため、次の2通りの条件で実験を行う。

1. 他者行為の観察無し

- (1) 自分自身の経験のみで行動学習を一定フレーム間行う
- (2) 行為獲得の成功率を評価する
- (3) 行為認識の成功率を評価する
- (4) 最初に戻る

2. 他者行為の観察有り

- (1) 人間行為の観察を 900 フレーム間行う
- (2) 観察により推定した状態値を使用し、行動学習を一定フレーム間行う
- (3) 行為獲得の成功率を評価する
- (4) 行為認識の成功率を評価する
- (5) 最初に戻る

フレーム数は各行為によって異なるが、観察の有無に関わらず1ループでのフレーム数は両者同じにすることで、全体の学習時間が同じになるようにしている。これらの条件で、それぞれ10回繰り返し性能評価を行った。なお、実演者は学習者に対して明示的な提示を一切せずに行為を実行する。そのため学習者は自身の持つ全ての学習器を同時に走らせ、状態値を推定していく。

行為獲得、行為認識の性能を評価するために、行為成功率、行為認識成功率という評価を導入する。行為成功率はロボットを100%最適で行動をとらせ、目標状態に辿り着くかどうかをロボットとボールが取りうる全ての配置において調べることで算出した。行為認識成功率は、認識判断が行われたとき、その認識が正解であるかどうかをロボットとボールが取りうる全ての配置において調べることで算出した。ここで認識判断を行う基準は行為認識確信度が0.7以上であるかどうかとする。また、行為認識を行うことによる利点の一つとして、できるだけ早期にその行為を認識することができれば、その行為に応じた行為を取ることができるといことが挙げられる。そのため行為認識成功率とは別に行為認識期間率という評価を導入し、どれだけ期間その行為が認識がされているかを調べる。行為認識期間率は、各配置で行為認識確信度が0.7以上になっている時間の割合を計算し、それらの平均をとることで算出した。

3.1.2 実験結果

シミュレーションによるシュート、パスに関する行為獲得率、行為認識成功率、行為認識期間率を Fig.8 ~ 10 に示す。グラフには 10 試行分の平均値と標準偏差をプロットしてある。

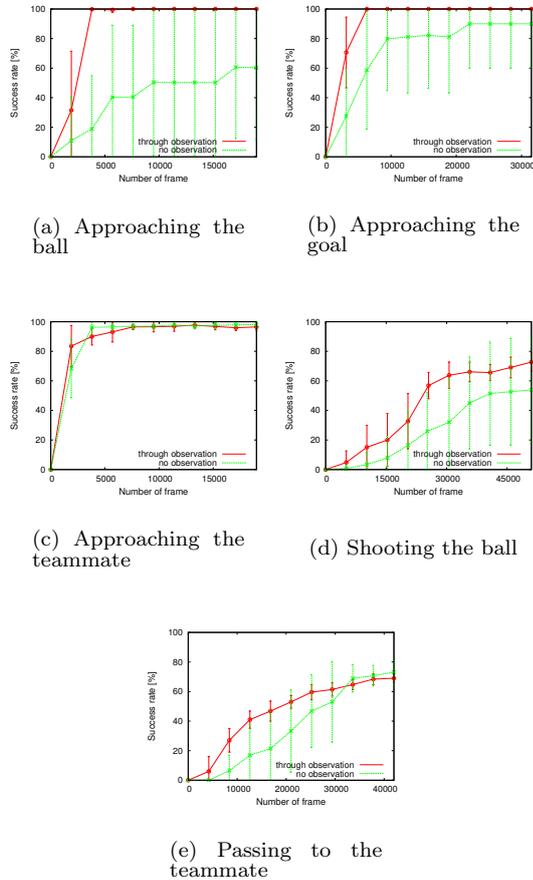


Fig.8 Success rate of the behaviors during learning with/without observation of demonstrator's behavior

Fig.8 ~ 10 より、どの行為においても他者行為の観察を通した学習の方が、平均値の傾きが急になっており、標準偏差の値も小さい。よって、人間行為の観察を通した学習の方が行為成功率、行為認識成功率、行為認識期間率が安定して早く発達している。

実演者の行為は複数の行為を連続させたものであるため、ロボットにとっては常に最適な行動のように観察されるとは限らない。しかし、その場合は自身の経験を優先させるようになっているため、ロボットは自身にとって有益な行為のみを観察することができ、今回のように行為理解が効率的に安定して発達させることができると言える。

4 おわりに

本研究では、強化学習における状態価値に基づいた行為獲得、人間行為認識の循環により、行為理解が効率的に安定して発達する手法を提案した。RoboCup 中型機リーグに出場しているロボットを想定したシミュレータに適用し、実演者が複数の行為を連続且つラン

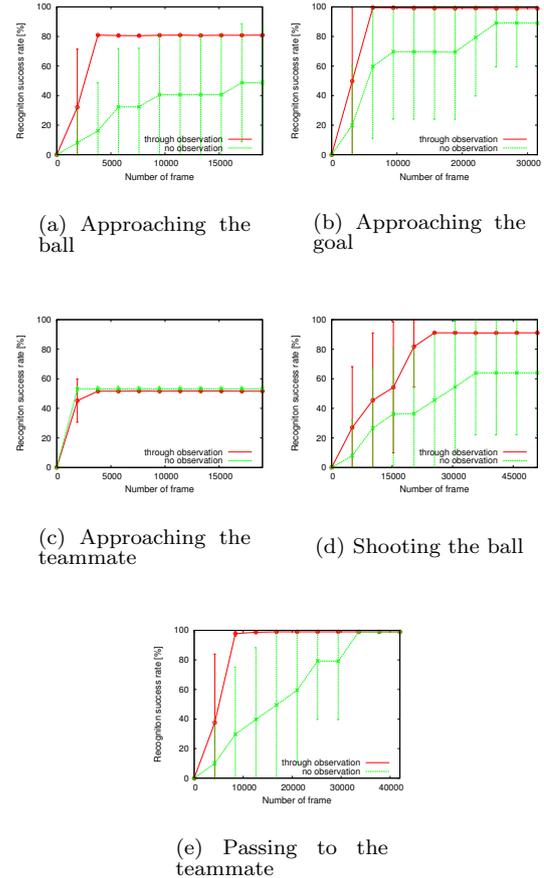


Fig.9 Recognition success rate of the behaviors during learning with/without observation of demonstrator's behavior

ダムに行った場合であっても他者行為の観察により行為獲得・他者行為認識が加速され、行為理解が安定して発達していくことを確認した。

謝辞

本研究の一部は (財) 栢森情報科学振興財団の支援を受けた。

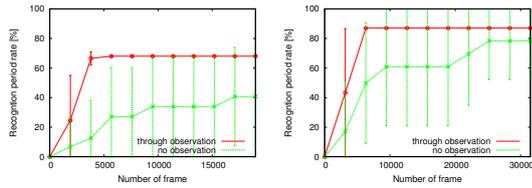
参考文献

- [1] Richard S.Sutton and Andrew G.Barto. 強化学習. 森北出版株式会社, 2000.
- [2] V.Gallese and A.Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive Sciences*, Vol. 2, No. 12, pp. 493-501, 1998.
- [3] Steven D.Whitehead. Complexity and cooperation in q-learning. In *Proceeding Eighth International Workshop on Machine Learning (ML91)*, pp. 363-367, 1991.
- [4] B.Price and C.Boutillier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 2003.

E-mail: yasutake@ams.eng.osaka-u.ac.jp

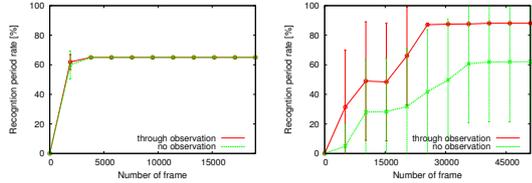
浅田 稔

E-mail: asada@ams.eng.osaka-u.ac.jp



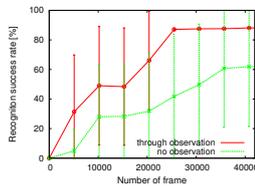
(a) Approaching the ball

(b) Approaching the goal



(c) Approaching the teammate

(d) Ahooting the ball



(e) Passing to the teammate

Fig.10 Recognition period rate of the behaviors during learning with/without observation of demonstrator's behavior

- [5] Darrin C. Bentivrgna, Christopher G. Atkeson, and Gorden Chenga. Learning tasks from observation and practice. *Robotics and Autonomous Systems*, Vol. 47, pp. 163–169, 2004.
- [6] Y.Takahashi, T.Kawamata, M.Asada, M.Negrello. Emulation and behavior understanding through shared values. In *Proceeding of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3950–3955, Oct 2007.
- [7] 田村佳宏, 高橋泰岳, 浅田稔. 価値システムに基づく他者行為観察と自己行動学習の循環的発達. 第24回ファジィシステムシンポジウム講演論文集, pp. TD1–4, Sep 2008.
- [8] Andrew N.Meltzoff. 'like me': a foundation for social cognition. *Developmental Science*, Vol. 10:1, pp. 126–134, 2007.

連絡先

田村 佳宏

E-mail: yoshihiro.tamura@ams.eng.osaka-u.ac.jp

高橋 泰岳