

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

Spiral Development of Behavior Acquisition and Recognition based on State Value

○ 田村 佳宏 (阪大) 正 高橋 泰岳 (阪大) 正 浅田 稔 (JST ERATO, 阪大)

Yoshihiro TAMURA, Osaka University, 2-1 Yamadaoka, Suita, Osaka
Yasutake TAKAHASHI, Osaka University.
Minoru ASADA, JST ERATO, Osaka University

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 behaviors shown by others through the collaboration between behavior acquisition and recognition of observed behaviors, 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 updates 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 two robots play soccer.

Key Words: Reinforcement Learning , Behavior Recognition , Value System , Learning by Observation

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、多くのセンサ情報(状態変数)が行動決定に必要とされる複雑な行為を自分自身の経験のみで学習する場合、探索空間が状態変数の数に応じて指数関数的に大きくなり、非現実的な学習時間が合目的な行為獲得までに必要とされる。

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、教示行動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。そこで本稿では、強化学習における状態価値に基づいた行為獲得と他者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ことを示す。ロボットは自身の試行錯誤の経験だけでは無く、観察した他者の行為の報酬を読み取り、他者の行為中の状態価値を推定し、これも自身の状態-行動価値の推定に利用することで自身の行動学習、及び他者行為認識を加速させることができる。

2 提案手法

2.1 基本概念

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ロボット(または人間)が存在し、学習者は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ると仮定する。また、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、提示する行為を実行する。Fig.1に我々の提案する手法の簡単な概念図を示す。学習者は他者行為の認識を行い、モデル規範型の強化学習を用いて、行動学習を行い、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の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の分類をし，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．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，行為獲得の質が早く向上し，行為理解が効率的に安定して発達する．

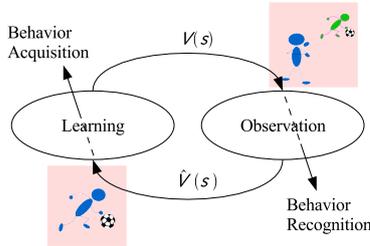


Fig.1 Spiral growth through Learning and Observation of Behaviors

2.2 システム概要

システムの全体図を Fig.2 に示す．1 つの学習器で行動学習と他者行為認識の 2 つを行うことができる．

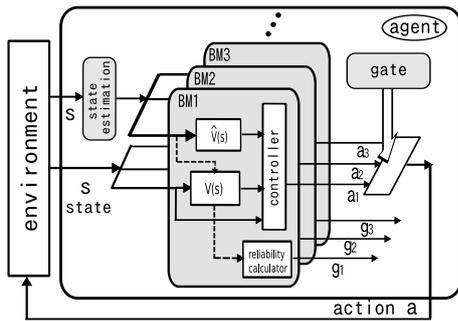


Fig.2 A whole system

2.3 行動学習器

行動モジュール (BM) は環境から状態 s を入力として受け取ると，状態価値 $V(s)$ を基にコントローラにより最適な行動を決定し出力する．各行動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 $s+1$ 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．

2.4 他者行為認識

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，そして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の状態価値を推定する．

状態価値の時間勾配から算出した行動意図の信頼度 (reliability) を基に，もっともらしい他者の行為の推定を行う．行動モジュール i の信頼度 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，本稿の実験では 0.1 としている．ここで $V_i(s_t)$ は状態 s_t における行動モジュール i の状態価値を表す．よって行動モジュールの状態価値が増え続けるほど，信頼度は大きな値となる．最も信頼度の大きくなった行為を他者行為として認識する．

2.5 観察による行動学習

2.5.1 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作成のバイアスとして使われる．

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

次に学習する行為の推定を行う 2 つの条件を示す．実演者からは明示的な教示が無いため，学習者は 2.4 で示した信頼度によって行為の認識を行うが，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．

1. 状態が遷移して，信頼度が上昇したとき，または信頼度が 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．

2.5.2 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．

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

ただし

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

3 実験

実験はシミュレーション，および実機で行った．実機による実験は RoboCup 中型機リーグに出場しているロボットを使用した．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．シミュレーションでも実機と同じロボットを想定している．環境中には学習者，実演者があり，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．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．

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.3 に示す．

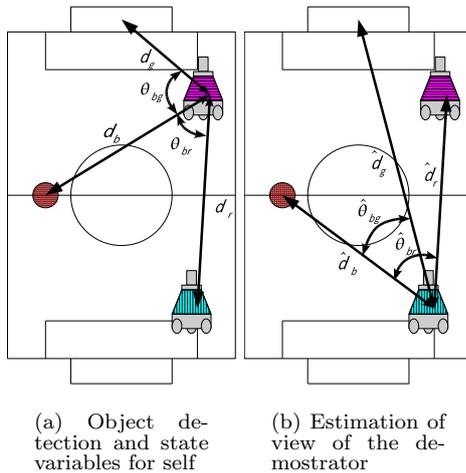


Fig.3 Estimation of view of the demonstrator

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 実験手順

3.1.1 シミュレーション

他者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パフォーマンス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

- 自分自身の経験のみで行動学習を15回行う
- 行為獲得のパフォーマンスを評価する

- 行為認識のパフォーマンスを評価する
- 最初に戻る

2. 他者行為の観察有り

- 他者行為の観察を5回行う
- 観察により推定した状態価値を使って，行動学習を10回行う
- 行為獲得のパフォーマンスを評価する
- 行為認識のパフォーマンスを評価する
- 最初に戻る

これらの条件を1試行とし，10試行行った．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，状態価値を推定していく．

行為獲得，行為認識のパフォーマンスを評価するために，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する．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する．ここで認識判断を行う基準は信頼度が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動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，どれだけの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，各配置で信頼度が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 実機による実験

提案手法が実機に適用できるかどうかを調べるため，観察を10回行い，約30分学習させ，ロボットが行為獲得，行為認識できているかを確認する．

3.2 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.4~6 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.4~6より，シュート，パス共に観察を通した方が各パフォーマンスが良くなっている．また，実機による実験のシュート，パスに関する行為獲得，行為認識のパフォーマンスを Fig.7.8 に示す．Fig.7.8より，シュート，パス共に行為獲得，行為認識が正しく実行できている．

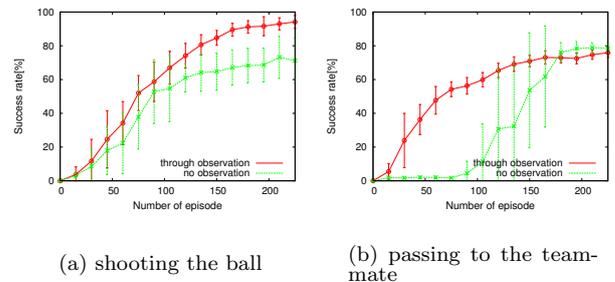


Fig.4 Success rate of the behaviors during learning with/without observation of demonstrator's behavior

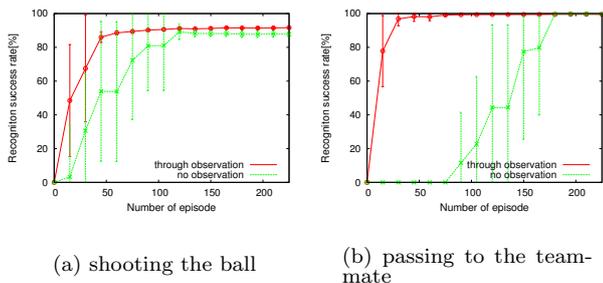


Fig.5 Recognition Success rate of the behaviors during learning with/without observation of demonstrator's behavior

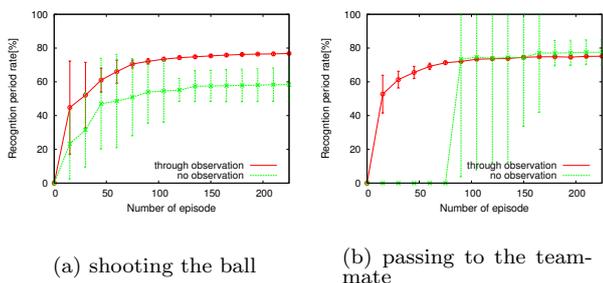


Fig.6 Recognition Period rate of the behaviors during learning with/without observation of demonstrator's behavior

4 おわりに

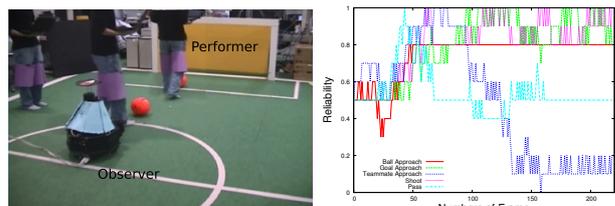
本稿では、強化学習における状態価値に基づいた行為獲得、他者行為認識の循環により、行為理解が効率的に安定して発達する手法を提案した。本手法の有効性を示すために、RoboCup 中型機リーグに出場しているロボットを想定したシミュレータ、及び実機に適用し、行為獲得・他者行為認識が加速され、行為理解が安定して発達していくことを確認した。

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.
- [5] Darrin C. Bentivrgna, Christopher G. Atkeson, and Gordon Chenga. Learning tasks from observation and practice. *Robotics and Autonomous Systems*, Vol. 47, pp. 163-169, 2004.
- [6] Y.Takahashi, T.Kawamata, M.Asada, and 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.



(a) A scene of observing the shoot behavior (b) A scene of executing the acquired shoot behavior

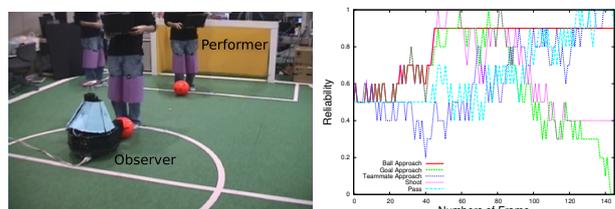


(c) Recognize the other's behavior (d) Reliability of each behavior module

Fig.7 Result of the shoot behavior acquisition and recognition of the real machine



(a) A scene of observing the pass behavior (b) A scene of executing the acquired pass behavior



(c) Recognize the other's behavior (d) Reliability of each behavior module

Fig.8 Result of the pass behavior acquisition and recognition of the real machine