状態価値に基づく
他者行為観察と自己行動学習の循環的発達

○ 田村佳宏 (阪大)　高橋泰岳 (阪大)　浅田稔 (JST ERATO, 阪大)

Mutual Development of Behavior Acquisition and Recognition based on State Value

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

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 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 carry the boxes.

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，様々な行為を自分自身の経験のみで学習には，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する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. 状態価値に基づく行為認識と行為学習

2·1 概要

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 [7]，これを強化学習の枠組で捉えると，学習者は実演者の報酬モデルが自分の持つ報酬モデルと同じであると推定することになる．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．学習者は他者行為を認識した後，モデル規範型の強化学習を用いて行動学習し，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，そして行動学習をする．$V(s)$ 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するとき，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ため，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．

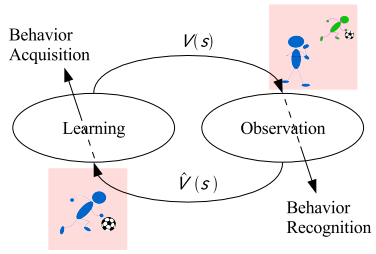


Fig.1 Spiral growth through learning and observation of behaviors

2·2　行動学習器

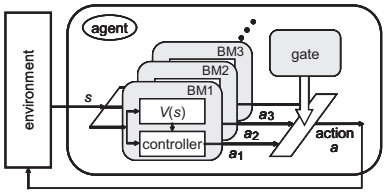


Fig.2 Modular learning system

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

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

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

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

の推定器を含む．状態遷移確率 $\hat{\mathcal{P}}^a_{ss'}$ と報酬 $\hat{\mathcal{R}}^a_{ss'}$ 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，

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

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

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

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．

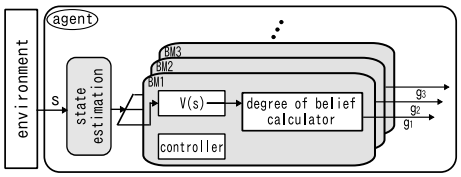


Fig.3 System for behavior recognition

行為認識システム図を Fig.3 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為認識確信度 (degree of belief) 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 の行為認識確信度 g_i は

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

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，行為認識確信度は大きな値となる．最も行為認識確信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信度の初期値は 0.5 とした．

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^o_t 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．

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

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信頼度が上昇したとき，または行為認識信頼度が 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

2·4.2 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．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．

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

ただし

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

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

3. 実験

実験はコンピュータシミュレーションで行った．環境中には学習者，実演者がおり，ボックスが 2 つ，そしてゴールエリアが 2 つ存在する．学習者は Fig.4 のようにボックスを指定された領域へ運ぶといった行為を観察，学習する．獲得する行為及び学習に必要な状態変数を Table.1 に示す．d_r，d_b，d_m，d_c はそれぞれロボットの掌と赤のボックス，青のボックス，赤紫の領域，青緑の領域との距離を表している．

また，Fig.5 に他者行為認識の例として，実演者が赤いボックスを赤紫の領域に置くという行為 (Fig.4 参照) を取ったときに観察者が推定した状態価値，行為認識確信度の遷移の様子を示す．Case1，Case2，Case3，Case4 はそれぞれ赤いボックスを赤紫の領域に置く，青いボックスを青緑の領域に置く，赤いボックスを青緑の領域に置く，青いボックスを赤紫の領域に置くという行為を表している．どちらのグラフにおいても Case1 の線は行為観察の間中，増加傾向を保っている．

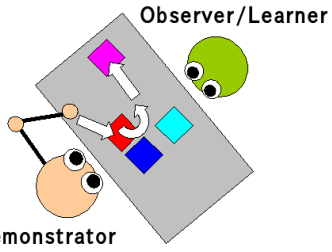
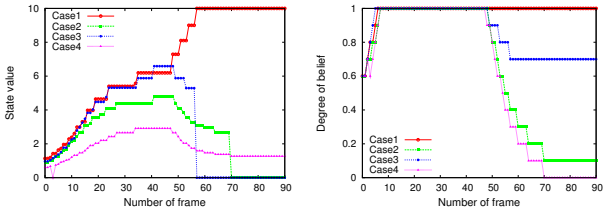


Fig.4 Scenario of the experiment

Table 1 List of behavior learned by self and state variables for each behavior

Behavior	State variables
Putting red box on magenta area	d_r and d_m
Putting red box on cyan area	d_r and d_c
Putting blue box on magenta area	d_b and d_m
Putting blue box on cyan area	d_b and d_c



(a) Estimated values　(b) Behavior recognition degree of belief

Fig.5 Sequence of estimated values and behavior recognition degree of belief during a behavior of putting red box on magenta area

3·1 実験手順

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．

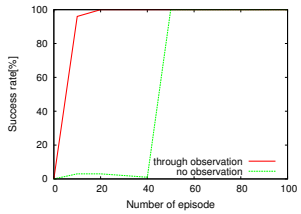
1. 他者行為の観察無し
   (1) 自分自身の経験のみで行動学習を 10 回行う
   (2) 行為獲得・認識の成功率を評価する
   (3) 最初に戻る
2. 他者行為の観察有り
   (1) 他者行為の観察を 2 回行う
   (2) 観察により推定した状態価値を使い，行動学習を 8 回行う
   (3) 行為獲得・認識の成功率を評価する
   (4) 最初に戻る

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．

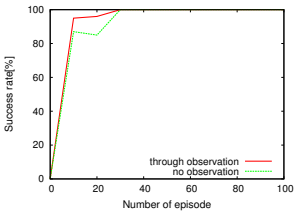
行為獲得，行為認識の性能を評価するために，行為成功率，行為認識成功率という評価を導入する．行為成功率はロボットを 100 %最適で行動をとらせ，目標状態に辿り着くかどうかをロボットのアームが取り得る全ての配置において調べることで算出した．行為認識成功率は，認識判断が行われたとき，その認識が正解であるかどうかをロボットのアームが取り得る全ての配置において調べることで算出した．ここで認識判断を行う基準は行為認識確信度が 0.7 以上であるかどうかとする．また，行為認識を行うことによる利点の一つとして，できるだけ早期にその行為を認識することができれば，その行為に応じた行為を取ることができるということが挙げられる．そのため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．

3·2 実験結果

シミュレーションによる行為獲得率，行為認識成功率，行為認識期間率を Fig.6〜8 に示す．

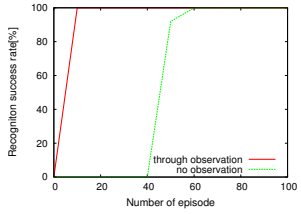


(a) Putting red box on magenta area

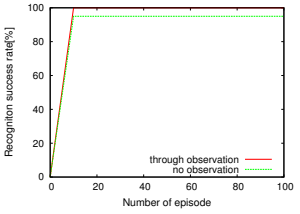


(b) Putting blue box on cyan area

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



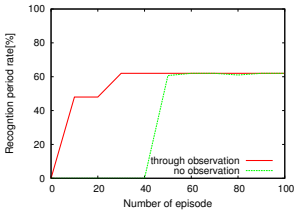
(a) Putting red box on magenta area



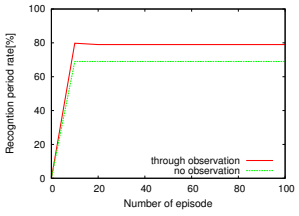
(b) Putting red box on cyan area

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

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行為成功率，行為認識成功率，行為認識期間率が早く発達している．しかし，青いボックスを青緑の領域に置く行為には，あまり観察の効果が現れていない．



(a) Putting red box on magenta area



(b) Putting blue box on cyan area

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

これは赤いボックスを赤紫の領域に置くという行為に比べてボックスとゴールとなる領域との距離が短く難易度が低いためである．このことから観察を通した行動学習・他者行為認識はタスクの難易度が高くなればなるほど効果を発揮すると言える．

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．本手法を上半身ヒューマノイドシミュレータに適用し，その有効性を確認した．

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.Boutilier. Accelerating reinforcement learning through implicit imitatione. *Journal of Artificial Intelligence Research*, 2003.

[5] Darrin C. Bentivrgna, Christopher G. Atkeson, and Gorden Chenga. Learning tasks from observation and practie. *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] Andrew N.Meltzoff. 'like me':a foundation for social cognition. *Developmental Science*, Vol. 10:1, pp. 126–134, 2007.