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

Mutual Development of Behavior Acquisition and Recognition based on Value System

高橋 泰岳 †， 田村 佳宏 †， 浅田 稔 ††， † 大阪大学大学院工学研究科 †† 大阪大学大学院工学研究科/科学技術振興機構 ERATO 浅田共創知能システムプロジェクト 大阪府吹田市山田丘 2-1 Yasutake

Takahashi † Yoshihiro Tamura † and Minoru Asada †† † Graduate School of Engineering, Osaka University †† Graduate School of Engineering, Osaka University/JST ERATO Asada Synergistic Intelligence Project Yamadaoka 2-1, Suita, Osaka, Japan

本論文では，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する．自身の試行錯誤の経験のみによる学習では獲得する行為が複雑になればなるほど多大な探索空間や莫大な学習時間が必要になる問題が強化学習による行為獲得には存在する．他者行為を観察し学習対象の行為の状態価値を推定し，それを自己の行動学習にフィードバックすることで行動学習を加速可能である．しかし，観測した他者行為を自己の行動学習に利用するためには，他者がどの行為を行っているのかを認識しなくてはならない．一方で，自己の行為の状態価値を基に他者の行為認識をロバストに行えることが先行研究によって示されている．行動学習と他者行為認識を交互に繰り返すことで，行為獲得を通した行為理解が効率的に安定して進められる．本手法の有効性を検証するため，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，及び実機に本手法を適用し，本手法の有効性を示す．

☒ 価値システム，行為理解，模倣，強化学習

Abstract

Both self-learning architecture and explicit/implicit teaching from other agents are necessary not only for learning beahvior for a task 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 interaction 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 updates can be accelerated by observation without real trial and error 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 Value system, Behavior recognition, Imitation, Reinforcement learning

1 はじめに

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．強化学習とは，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，実環境で活動するロボットが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的な行為獲得までに必要になる．この問題を解決するために様々な研究がなされている．学習の効率化という点から，例えば取り扱うタスクをサブタスクに分解し，個々に学習した行為を利用して複合的なタスクを行わせる研究[2, 3, 4, 5, 6, 7, 8, 9]や他者が学習した経験や学習結果を共有する手法[10, 11]などがある．

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[12]．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，すなわち，行動学習のモジュールは行動の学習のためだけに使われるのではなく，他者の行為認識のためにも使用され，また行為認識は行動学習のためのバイアスとなっている可能性を示していると考えられる．本論文において「認識」とはロボットが観察した行為を自身の行為の一つに分類することを示す．実際，他のロボットや人間との共生を行う環境下では，自身の試行錯誤のみで行為学習する必要はなく，むしろ他者の行っている行為の観察を通して未学習行為を獲得する方が現実的である．他者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可能性がある[10, 13, 14]．そのために観察した他者の行為を理解する必要があるが，観察した行為を理解するということは，単純に実演者のエンドエフェクタや関節の軌道を追従することを意味するのではなく，他者の意図（観察された行為の目的・目標）を読み取ることや，目的を達成する方法を自分自身で見つけられることを意味する．これを強化学習の枠組で捉えると，観察を通して，観察した行為の報酬を読み取り，状態価値を推定することになる．

他者行為を観察する際，他者がある一つのタスクを遂行する行為のみを行っているのではなく，複数のタスクを順次，あるいは複合的に遂行していることが多いと考えられる．また，複数の行為を同時に学習する方が逐次的に一つ一つ別々に学習するよりも効率がよい．複数のタスクを学習する手法としてモジュール型学習機構を用いることが多い．Jacobs と Jordan[15]は複数の学習モジュールを用い，各学習モジュールの出力をゲートで重み付けしたものをシステム全体の出力とする Mixture of Experts と呼ばれる学習システムを提案している．各学習モジュールの状況に対する適応度を重みとすれば，広く適用できる[16, 17, 18, 19]．同様の考え方で環境のダイナミクスの変動に対応可能な学習アルゴリズムも提案されている[20, 21]．鮫島ら[22]や Haruno ら[23]は非線形・非定常なタスクの制御則をモジュール構造を用いて学習させる MOSAIC を提案している．

他者の行動予測に関して研究として Nagayuki ら[24]は観察者が他者の動作系列を観察して記憶し，他者の行為のモデルとして獲得する手法を提案し，観察者自身の行動の決定の際に，他者の行為モデルを用いて予測することで，自身のタスク達成に最適な行動を選択している．Tohyama ら[25]は，他者の行為が単一ではなく複数ある場合に対応して，観察者が他者の行為モデルを複数持つことを提案している．その複数のモデルの中から，観測された他者の状態遷移に対して尤度の高い他者の行動モデルを他者の意図とみなし，予測に用いている．また，Inamura ら[26]は隠れマルコフモデル (HMM) を用いてヒューマノイドにおける運動パターンの認識および生成を行なっている．行為認識においては，事前の学習によって，複数の行動を HMM を用いたシンボルとして表現し，実際に観測によって得られた他者の関節角に基づく行動要素系列に対して，尤度の高い HMM を他者の行為として認識している．鮫島ら[27]は前述の MOSAIC を用い観察によって他者の行動の行為認識を実現している．他者の行動の観察から他者の状態を推定し，次状態を各学習器の予測器毎に予測させ，次の時刻において実際に観測された状態と，予測された状態を比較し，最も誤差の小さい予測をした学習器を他者の行為と見なす．しかし，同じ行為でも複数の状態遷移系列が多く存在する場合，予測器の予測する状態の遷移系列と実際の状態遷移の系列を比較する鮫島らの手法では，それぞれの状態遷移系列が同一の行為を示していると認識できない．これは相手の行為を状態遷移系列によって分類する手法の欠点である．片山ら[28]は他者（ユーザ）間の類似性をもとに，他者に対する複数の行為学習を高速化している．この手法は他者の行為を参考にしているものの，その行為の模倣は行っていないため，目的が異なる．

高橋ら[29]は学習済みの行為の状態価値を用いて，他者の行為の認識を行う手法を提案した．提示された行為の状態遷移系列が異なる場合でも比較手法と比べてロバストに認識できること示した．これはつまり，

状態価値という一つの基準によって，行為獲得と行為認識の両方が導けるということを示している．つまり，他者行為を観察し，学習対象の行為の状態価値を推定し，それを自己の行動学習にフィードバックすることで行動学習を加速可能である．しかし，観測した他者行為を自己の行動学習に利用するためには，他者がどの行為を行っているのかを認識しなくてはならない．未知の行為に対しては状態価値が推定できないため，学習初期では他者行為の観察情報を自己の行動学習に生かせないが，行動学習と他者行為認識を交互に繰り返すことで，行為獲得を通した行為理解が効率的に安定して発達できると考えられる．

そこで本論文では，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ことを示す．ロボットは自身の試行錯誤の経験だけではなく，観察した他者の行為の報酬を読み取り，他者行為の状態価値を推定し，それを利用することで自身の行動学習を加速させる．Fig.1 に我々の提案する手法の概念図を示す．$V(s)$ 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．自己行動学習によって得られた状態価値 $V(s)$ を利用して他者行為を認識し，他者行為の観察によって得られた推定状態価値 $\hat{V}(s)$ を自己の行動学習にフィードバックする．この循環を通すことで，自己の行動学習を加速し，他者の行為認識の性能も上げることができる．

以下では，第２章において強化学習と状態価値に基づく他者行為認識について，第３章において提案手法について述べ，第４章で提案手法の有効性を検証するために行った実験について，最後に第５章で結言を述べる．

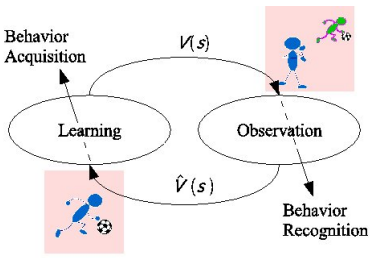


Fig. 1　Interaction between learning and observation of behavior

2　強化学習と状態価値に基づく他者行為認識

ここでは強化学習と Takahashi ら [30] によって提案された自己行為の状態価値に基づく他者行為認識手法について述べる．

21　強化学習に基づく行為学習

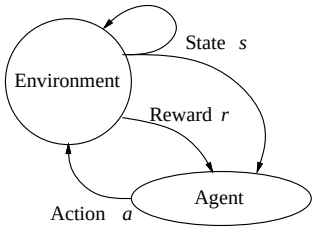


Fig. 2　Agent-environment interaction

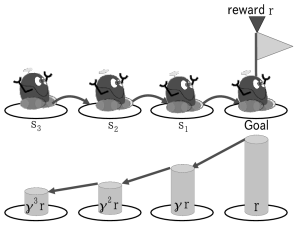


Fig. 3　Sketch of state value propagation

Fig.2 に強化学習の概念図を示す．ロボットは環境から状態 $s \in \mathcal{S}$($\mathcal{S}$ は可能な状態の集合) を観測する．環境はマルコフ過程に従うと仮定し，ロボットは現状態 s_t においてある方策 π に基づき行動選択し，次状態に遷移し，報酬 r_{t+1} を受け取る．状態価値 V^π は方策 π に従って行動選択しているときに将来にわたって受け取るだろう報酬の減衰和の期待値であり，以下の式で表される．

$$V^\pi = E\left\{\sum_{t=1}^{\infty} \gamma^t r_t\right\} \tag{1}$$

Fig.3 にロボットがゴール状態に止まったときに正の報酬を，それ以外の状態では 0 の報酬を受け取った場合の状態価値の模式図を示す．状態価値はロボットが正の報酬を受け取る場所で最も高い値を持ち，それより遠い状態には減衰された値が伝搬される．

合目的的な行為を行う方策 π に従うことで状態価値が上がる傾向にある．強化学習ではこの状態価値がより高くなるように方策を修正し，修正した方策に対応する状態価値を推定し直す手続きを繰り返し踏むことで，方策を改善する．本論文では獲得された方策 π に従う行動の系列を行為と呼ぶ．より詳しい説明は Sutton and Barto の著書 [1] や学習ロボットのサーベイ [2] に詳しい．

ここでモデルベースの強化学習を導入する．学習モジュールは状態遷移モデルを持ち，このモデルを用いて状態・行動価値関数を推定する．状態遷移モデルはある状態 s において行動 a を選択したときに次状態 s' に遷移する確率 $\hat{\mathcal{P}}^a_{ss'}$ を出力する．

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

また状態遷移モデル以外に報酬モデル $\hat{\mathcal{R}}_s$ も保持する．

$$\hat{\mathcal{R}}(s) = E\{r_t | s_t = s\} \qquad (3)$$

学習者が経験した状態・行動・報酬の系列から単純に確率と平均を計算することで状態遷移モデルと報酬モデルを推定する．この状態遷移モデルと報酬モデルから，ある状態 s における行動 a の価値，行動価値関数 $Q(s,a)$，が以下で与えられる．

$$Q(s,a) = \sum_{s'} \hat{\mathcal{P}}^a_{ss'} \left[\hat{\mathcal{R}}(s') + \gamma V(s') \right] \qquad (4)$$

$$V(s) = \max_a Q(s,a)\ , \qquad (5)$$

ただし γ は減衰係数である．

22　モジュール型行動学習機構

複数の行為を観察/学習/実行するためにモジュール型学習機構を採用する．多くのモジュール型学習機構が今まで提案されている [15, 16, 3] が，本研究で用いるモジュール型学習機構は Fig.4 に示すように，行為モジュール（Behavior Module:BM）とゲートで構成する．学習者は複数の行為モジュールを持っており，一つの行為モジュールは一つの行為に対応する．行為モジュールは環境から状態 s を入力として受け取ると，状態価値関数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選択され，学習者の行動として出力される．

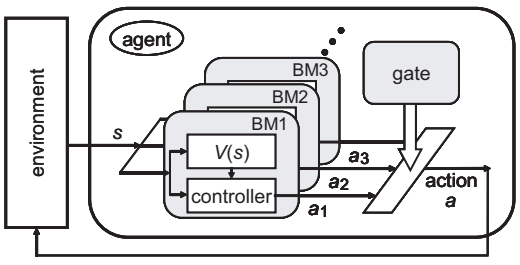


Fig. 4　Modular learning system

23　状態価値に基づく他者行為認識

例として Fig.5 のように，エージェントがボールに近付くタスクを行なう場合を考える．状態 s はエージェントの位置座標 (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な方策による状態遷移系列，すなわち位置座標の系列は path1 に示す系列であるのに対し，他者の行う行為は path2 に示す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.5 で示すように，状態価値関数 $V(s_1, s_2)$ 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傾向にある点においては同じであるといえる．ここで確信度と呼ぶパラメータを導入し，状態価値の上昇・減少に応じてその確信度の値を変化させていけば，同じ行為として認識できる．

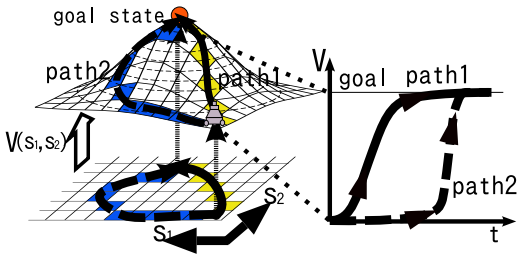


Fig. 5　Behavior recognition based on the change of state value

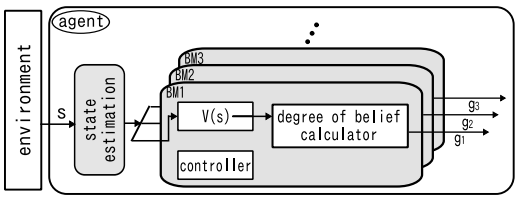


Fig. 6　System for behavior recognition

行為認識システム図を Fig.6 に示す．行為認識システムの中には複数の行為モジュール，そして他者状態推定器（state estimation）が存在する．一つの行為モジュールは一つの行為に対応し，学習者は複数の行為を同時に認識する．各行為モジュールは推定された他者の状態を受け取り，自身の持つ状態価値関数によって状態を状態価値に写像し，出力する．それらの状態価値は確信度計算機（degree of belief calculator）に渡され，状態価値の時間変化を見ることで．他者の行為を推定する．

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，それを三次元再構成することで他者視点における観測情報へと変換し，他者の状態，及び状態価値を推定する．確信度計算機は状態価値の勾配から算出した行為認識の確信度 (degree of belief) を基に，もっともらしい

他者の行為の推定を行う．行為モジュール i の確信度 g_i は

$$g_i = \begin{cases} g_i + \beta & \text{if } V_t - V_{t-1} > 0 \text{ and } g_i < 1 \\ g_i & \text{if } V_t - V_{t-1} = 0 \\ g_i - \beta & \text{if } V_t - V_{t-1} < 0 \text{ and } g_i > 0 \end{cases}, \tag{6}$$

とする．ただし，β は更新度であり，本論文中の実験では 0.1 としている．β の値を小さくとると確信度が変化が鈍り，必要な認識時間が長くなるが，より正確に認識できる．大きくとった場合，確信度の変化が敏感になり，認識に必要な時間は短くなるが，認識性能は悪化する．この値は適切に設定する必要があり，ここでは経験的に決定した．ここで $V_i(s_t)$ は状態 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，確信度は大きな値となる．各学習器の状態遷移が行われたときに確信度の更新が行われる．

3 他者行為観察と自己行動学習による循環的発達

他者行為観察と自己行動学習による循環的発達のシステムについて述べる．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とする．また，実演者は学習者に対して明示的な情報提供をせず，なんらかの行為を実行する．

状態価値の直観的な意味は目標状態への近さである．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ときのみ状態遷移モデル，行動価値関数を更新させれば，得られる状態価値関数は自身の状態価値関数よりも良くなる可能性がある．しかし，実演者がどのような行動をしているかは学習者には判断できないため，観察で得られた推定状態価値 $\hat{V}^o(s)$ をそのまま自身の行動価値 $Q(s,a)$ として使うことはできない．そこで，未学習領域で観察で得られた推定状態価値 $\hat{V}^o(s)$ を利用することを考える．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 の行動価値関数 $Q(s,a)$ の更新にどちらの値を使用するかを決定する．

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

$V'(s)$ は次式で定義する．

$$V'(s) = \begin{cases} V(s) & \text{if } V(s) > \eta^n \hat{V}^o(s) \\ \eta^n \hat{V}^o(s) & \text{else} \end{cases} \tag{8}$$

$0 < \eta < 1$ は減衰係数，n は学習者が状態 s を経験した回数，$\hat{V}^o(s)$ は観察データによる推定状態価値関数であり，次式で定義する．

$$\hat{V}^o(s) = \sum_{s'} \hat{\mathcal{P}}^o_{ss'} \left[\hat{\mathcal{R}}(s') + \gamma \hat{V}^o(s') \right] \tag{9}$$

ここで，$\hat{\mathcal{P}}^o_{ss'}$ は観察時の状態 s から次状態 s' への状態遷移確率である．式 (7) に示すように，もしある状態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 $\hat{V}^o(s)$ に自身の経験回数を割り引いた値より大きければ，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の経験数に応じた減衰値 $\eta^n \hat{V}^o(s)$ を使用する．これにより，未学習の領域で状態価値が低く見積もられている状態においても，観察によって得られた推定状態価値を利用することで，状態価値学習が加速するため，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最適な値を持つ状態価値関数とは限らない．そこで，その状態を経験した回数に応じて推定された状態価値を割り引き，その状態を経験した回数が多ければ多いほど，観察による状態価値の学習の影響を減らせることで，自己の学習経験による状態価値を優先する．

4 観察を通した行為学習

第 3 章の提案メカニズムを RoboCup 中型機リーグに出場しているサッカーロボット (Fig.7) に適用し，シミュレーション，実機での実験により提案手法の有効性を検証する．

41 実験設定

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．全方位カメラは全方位ミラーと単眼カメラを組み合わせたものを使い，色情報を使った単純な画像処理により周りのゴール，チームメイト及びボールを毎秒 30 フレームで認識する．今回使用したフィールドサイズ（約 5m x 8m）では，ロボットとオブジェクトがピッチ上にある場合は十分認識できると仮定した．このためロボットは，常に周囲の物体を認識できる．全方位異動機構により 2 次元平面上においてどの方向にも並進及び回転できる．シミュレーション実験も実機による実験と同じように，

RoboCup 中型機リーグのフィールド，それに出場しているロボットを想定したものである．環境中には学習者，実演者がおり，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．またフィールドを 3×3 の 9 個の領域に分割し，試行開始時に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.8 参照）．

Fig. 7　A whole image of robot

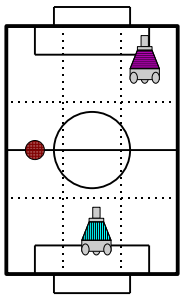


Fig. 8　Predefined positions on the soccer field for experiments

411　タスクと学習時の行動

学習，認識に使用したタスクは以下の 5 つである．今回の実験において，シュートとはゴールの位置までドリブルしていくことであり，パスはチームメイトの位置までドリブルしていくことであると定義する．

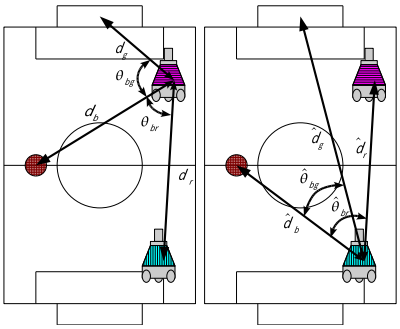
- ボールへのアプローチ
- ゴールへのアプローチ
- チームメイトへのアプローチ
- ゴールへのシュート
- チームメイトへのパス

ロボットの学習時の行動としては以下に示す 6 つである．

- ボールに近づく
- ゴールに近づく
- チームメイトに近づく
- ボールから離れる
- ボールを中心に時計回りに回転
- ボールを中心に反時計回りに回転

412　状態変数

各タスクにおける状態変数は Fig.9 に示すように，自己行動学習時は Table.1，他者行為観察時は Table.2 のようにした．



(a) Object detection and state variables for self　(b) Estimation of view of the demostrator

Fig. 9　Estimation of view of the demostrator

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}

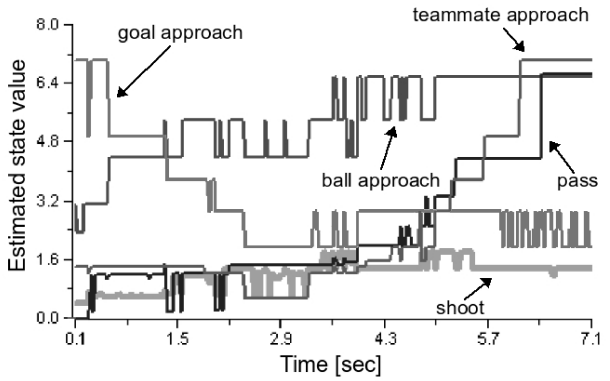
Table 2　List of observed behaviors and state variables for each behavior

Behavior	State variables
Approaching a ball	$\hat{d}_b$
Approaching a goal	$\hat{d}_g$
Approaching the teammate	$\hat{d}_r$
Shooting a ball	$\hat{d}_b$, $\hat{d}_g$, $\hat{\theta}_{bg}$
Passing a ball	$\hat{d}_b$, $\hat{d}_r$, $\hat{\theta}_{br}$

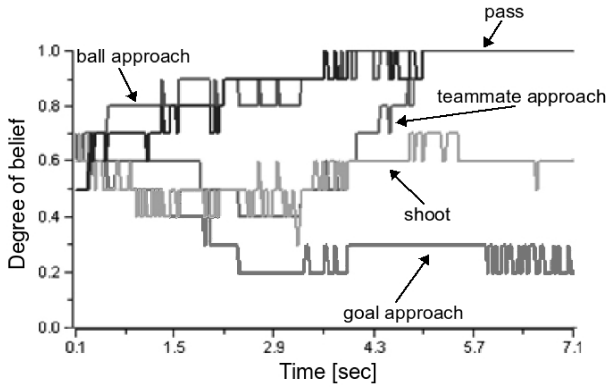
42　自己の状態価値に基づく行為認識

行為観察に基づく行動学習を行う前に，Takahashi ら [30] に基づき，各行為モジュールが観察している行為の状態価値を推定できることを示す．学習者は実演者の行為を観察し，自信のセンサ情報から実演者が取得しているであろうセンサ情報を簡単なマッピングを使って推定する．Fig.9 にその推定方法を示す．学習者は色づけされたオブジェクトを全方位視覚を使って

認識し，自己中心の世界座標系における距離と方向を算出する．そこから実演者の位置に原点を移動させることで実演者の視野を推定し，実演者の状態を算出する．その推定した状態を用いて各行為モジュールはそれぞれの状態価値関数に写像し，状態価値の時系列から確信度を算出し続ける．



(a) Estimated Values



(b) Degrees Of Belief

Fig. 10 Sequence of estimated values and degrees of belief during a behavior of pushing a ball to the magenta player

実演者が観察者にパス行為を見せたときの推定状態価値と確信度の時系列を Figs.10(a) ，(b) に示す．チームメイトへのパス行為の推定状態価値が上昇傾向にあることがわかる．この行為はボールへのアプローチ，チームメイトへのアプローチ行為を含んでいるため，これらの行為の推定状態価値も観察時の前半および後半で上昇していることがわかる．すべての確信度の初期値は 0.5 に設定し，推定状態価値の遷移によって確信度が更新される．推定状態価値の値そのものが小さくても，その値が上昇していれば確信度はあるステップ幅で上昇するため，適切な時間内に行為認識が行える．チームメイトへのパス行為の確信度は行為観察時の中程で 1.0 に到達しているのがわかる．本実験では確信度が高い複数の行為をすべて認識するとする．つまり，パス行為はボールアプローチとチームメイトへのアプローチ行為を含み，これらの確信度も上がっているため，これらの行為も同時に認識する．

43 性能比較実験

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，シミュレーション上で次の２通りの条件で実験を行う．

- 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１５回行う
  2. 獲得行為のパフォーマンスを評価する
  3. 行為認識のパフォーマンスを評価する
  4. １に戻る
- 他者行為の観察有り
  1. 他者行為の観察を５回行う
  2. 観察により推定した状態価値を利用し，行動学習を１０回行う
  3. 獲得行為のパフォーマンスを評価する
  4. 行為認識のパフォーマンスを評価する
  5. １に戻る

これらの条件を 1 試行とし，10 試行行った．両条件で学習に利用する訓練データの数を同数にすることで妥当な比較が行えると考えたので，他者行為の観察無しの条件では 15 回の行動学習に対し，他者行為の観察ありの条件では他者行為の 5 回の観察に加え 10 回の行動学習を行うことで，両条件での訓練データの数を同数にした．なお，実演者には

1. ボールへのアプローチ
2. ゴールへのアプローチ
3. チームメイトへのアプローチ
4. ゴールへのシュート
5. チームメイトへのパス

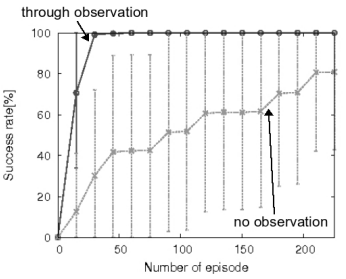
の順番で行動させ，学習者にはその順番を一切知らせずに観察させる．そのため学習者は自身の持つ全ての学習器を同時に走らせ，状態価値を推定していく．実演者の各試行が終了した際は，その実演者と観察者，ボールの位置をフィールド上でランダムに配置し，次の試行を開始する．行動学習時の行動選択はその状態

において最大の行動価値を持つ行動を 80%の確率でとり，20%の確率でランダムに行動選択をする．獲得行為のパフォーマンスを評価する際は常に最大の行動価値を持つ行動を選択する．経験不足で行動価値が推定されず，最大の行動価値を持つ行動を選択できない場合は，ランダムに行動を選択する．

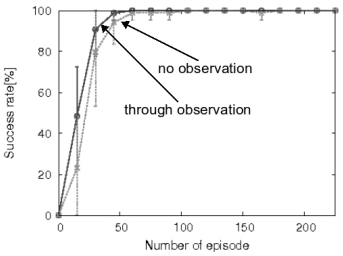
獲得行為，行為認識のパフォーマンスを評価するために，行為成功率，行為認識成功率と呼ぶパラメータを導入する．行為成功率はロボットをその時点で最適と推定される行動で行動させ，目標状態に辿り着くかどうかをロボットとボールが取りうる全ての配置において調べることで算出する．行為認識成功率は，認識判断が行われたとき，その認識が正解であるかどうかを，ロボットとボールが取りうる全ての配置において調べることで算出する．ここで認識判断を行う基準は観測終了時に確信度が 0.7 以上の行為の存在の有無とする．また，行為認識を行うことによる利点の一つとして，できるだけ早期にその行為を認識することができれば，その行為に応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呼ぶパラメータを導入し，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．行為認識期間率はロボットとボールが取りうる全ての配置においてで確信度が 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．

各タスクにおける獲得行為率，行為認識成功率，行為認識期間率を Fig.11〜13 に示す．Fig.11 はそれぞれの時点での学習結果を用い，全状況から試行したときのタスク成功率を示している．学習時や観察時の試行の初期配置はランダムに行い，一連の実験を 10 回繰り返したときの平均値と標準偏差を示している．観察を通した行動学習時の成功率はどの行為に関しても観察を行わないものよりも良い値を示している．行動学習の加速化は実ロボットへの適用時に最も重要な項目の一つであり，提案システムが有効であることを示している．観察情報を利用しない場合，学習初期では全くタスクを達成できない状況が続く場合がある．これは学習初期ではランダムな行動選択をとる傾向があり，そのため偶然にタスクを達成する確率が低いためである．そのため成功率の分散が大きい．一方で観察情報を自己の行動学習にフィードバックしている方は学習が加速し，かつ学習傾向も安定している．

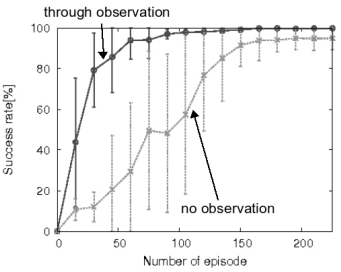
行為認識成功率と行為認識期間率を Figs.12 と 13 にそれぞれ示す．これらは行為成功率と同じような傾向を示す．つまり，自己の行為モジュールを利用して実演者の行為を認識しているため，行為成功率が高ければ行為認識成功率と行為認識期間率も同様に高い．



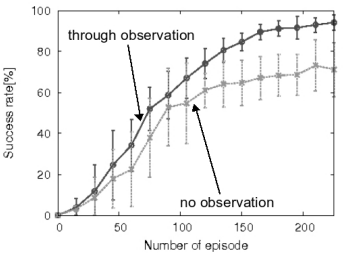
(a) approaching the ball



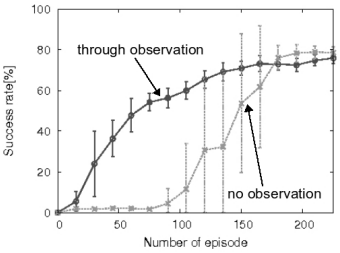
(b) approaching the goal



(c) approaching the teammate

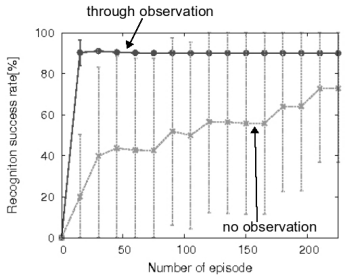


(d) shooting the ball

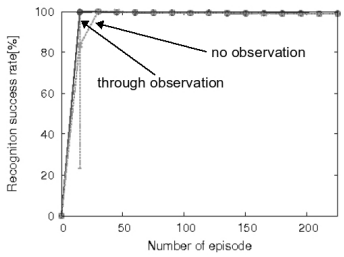


(e) passing to the teammate

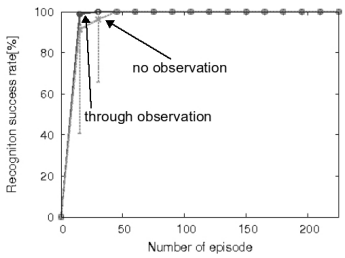
Fig. 11　Success rate of the behavior during learning with/without observation of demonstrator's behavior



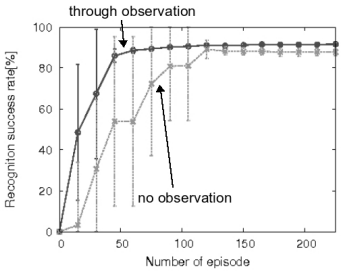
(a) approaching the ball



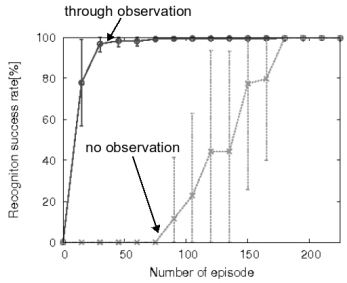
(b) approaching the goal



(c) approaching the teammate

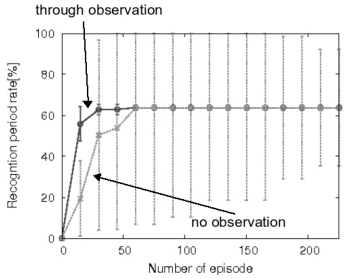


(d) shooting the ball

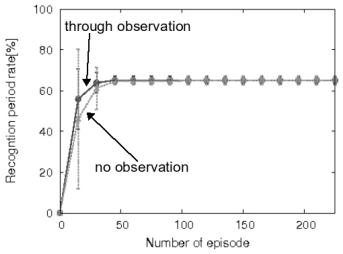


(e) passing to the teammate

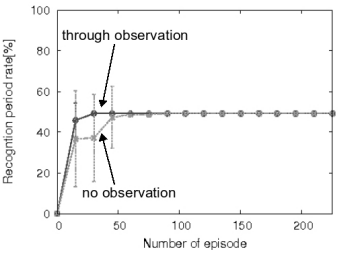
Fig. 12 Recognition performance of the behavior during learning with/without observation of demonstrator's behavior



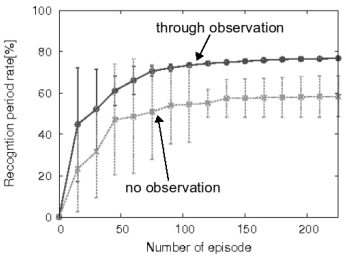
(a) approaching the ball



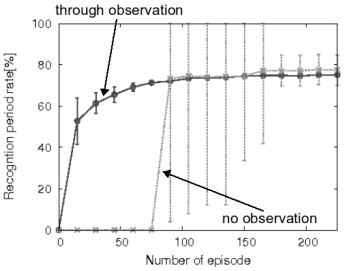
(b) approaching the goal



(c) approaching the teammate



(d) shooting the ball



(e) passing to the teammate

Fig. 13 Recognition period rate of the behavior during learning with/without observation of demonstrator's behavior

逆に行為認識成功率と行為認識期間率が高ければ，その分自己の行動学習にフィードバックできる情報が増えるため，行動学習が加速する．このように自己の行動学習と他者の行為認識を循環させることで，行動学習および行為理解が十分加速されることが示された．

どのタスクにおいても他者行為の観察を通した学習の方が，傾きが急になっており，標準偏差の値も小さい．よって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，行為認識期間率が安定して早く発達していると言える．しかし，ゴールアプローチのタスクだけは他のタスクに比べて観察を通した行動学習の効果が著しく小さい．今回，シミュレーション実験において，ロボットは常にゴールの方向を向いた状態で配置しており，まっすぐ進むだけで目標状態に辿り着く単純なタスクとなっている．さらに，ロボットが学習時に取る行動の中にはゴールから離れるといったゴールアプローチのタスクに対して必ず負の要素になるような行動は入れていない．このような単純なタスク，行動条件においては，自身の学習経験のみで十分に行為を獲得できるため，観察を通して良い状態価値関数を推測してもメリットは殆ど無い．よって，ゴールアプローチのタスクに関しては観察を通した行動学習の効果は小さくなる．

提案手法を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，30 分学習させ，ロボットの獲得行為および行為認識の性能を確認する．獲得行為，行為認識の性能については，獲得行為についてはその時点で最適と推定される行動を選択させ，目標状態に辿り着くかどうかを確認し，行為認識については人間の行為を見せ，それが正確に認識できているかどうかを確認する．今回はシュートとパスの行為について実機による実験を行う．

シュート，パスに関する獲得行為，行為認識のパフォーマンスを Fig.14，15 に示す．Fig.14(a), 15(a) は人間の実演者がロボットにそれぞれシュートとパスを提示している一例を示している．それらの行為の観察をもとに行為獲得・他者行為認識の循環を通して学習した行為の一例を Fig.14(b), 15(b) に示している．また，学習後に人間の実演者がシュート行為とパス行為を提示し (Fig.14(c),Fig.15(c))，その行為の認識結果をそれぞれ Fig.14(d),15(d) に示す．例えば Fig.14(d) では早い段階でボールアプローチ，シュートの確信度が上がり，ゴールアプローチは一旦確信度が下がるが最終的に高い値に収束している．シュートにはボールアプローチとゴールアプローチを伴うため，これら 3 つの行為の確信度が高い値に収束する．その他のチームメイトへのアプローチやパスは，たまたまボールがチームメイトの近くにあるため一旦上昇するが，最終的に離れていっているので最終的な確信度が低くなっているのがわかる．Fig.15(d) でもタスクがシュート行動からパス行動に，ゴールアプローチがチームメイトアプローチに変わるが，同様の傾向が見られる．従って，どちらのタスクにおいても，獲得行為，行為認識が実現できており，提案手法は実機にも適用できると言える．

5 おわりに

本論文では，高橋らによっていて案された手法 [29] を拡張し，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，獲得行為・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．今回の実験設定では実演者が提示した行動軌道が学習者が学習すべき行動に近かったため，行為獲得・他者行為認識の循環を通した性能が安定し，循環を通さない場合に比べて良い結果が得られた．提示された行為が学習者にとって実現不可能な場合や，学習者の他者行為認識の性能が悪かった場合は，今回の実験の様な良い結果は得られない可能性がある．しかし，式 (8) で自身の経験によって得られた状態遷移モデルから他者行為認識によって得られた推定状態価値の重みを下げることによって観察に依存しない行動学習の形になっているため，一時的に学習の遅れがあっても最終的には目的の行動を学習可能ではある．今後の研究では学習者にとって実現可能な行動軌道を学習初期に推定し，実現不可能なデータについては棄却する枠組みに拡張することが考えられる．また，良い初期方策を得ることが難しく，他者行動の観察なしでは学習が難しいケース，例えばより難しいダイナミクスを考慮したタスクに対して，本手法を適用し有効性を確認する予定である．

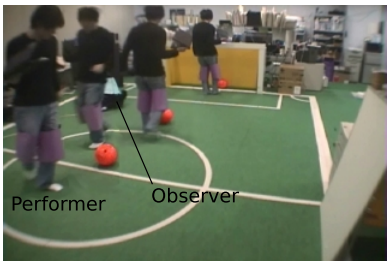
謝辞

本研究の一部は栢森情報科学振興財団の助成を受けて遂行された．

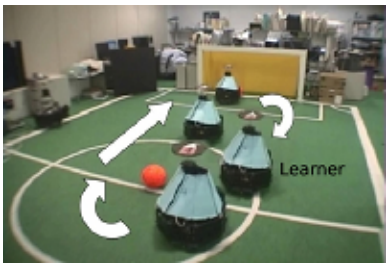
参考文献

1) R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

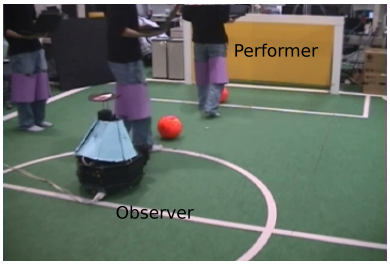
2) Jonathan H. Connell and Sridhar Mahadevan. *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.



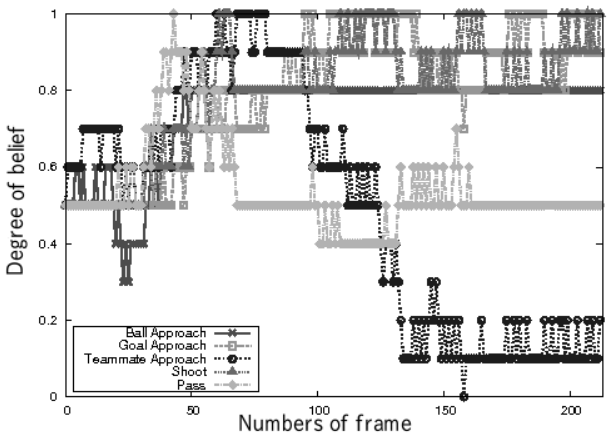
(a) A scene of observing the shoot behavior



(b) A scene of executing the acquired shoot behavior

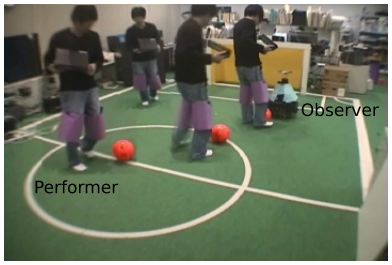


(c) Recognition of the demonstrator's behavior



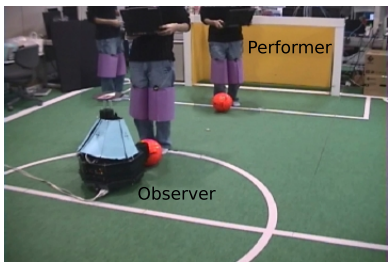
(d) Degree of belief each behavior module

Fig. 14 Result of the shoot behavior acqusition and recognition of the real machine

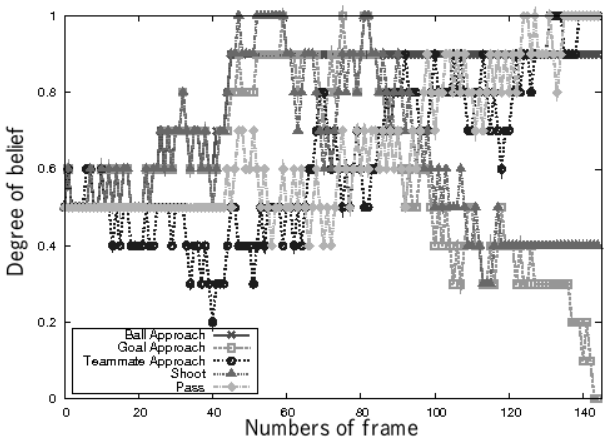


(a) A scene of observing the pass behavior

(b) A scene of executing the acquired pass behavior



(c) Recognition of the demonstrator's behavior



(d) Degree of belief of each behavior module

Fig. 15 Result of the pass behavior acqusition and recognition of the real machine

3) Steven Whitehead, Jonas Karlsson, and Josh Tenenberg. Learning multiple goal behavior via task decomposition and dynamic policy merging. In Jonathan H. Connell and Sridhar Mahadevan, editors, *ROBOT LEARNING*, chapter 3, pp. 45–78. Kluwer Academic Publishers, 1993.

4) Leslie Pack Kaelbling. Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the Tenth International Conference on Machine Learning*, 1993.

5) Peter Stone and Mamuela Veloso. Layered approach to learning client behaviors in the robocup soccer server. *Applied Artificial Intelligence*, Vol. 12, No. 2-3, 1998.

6) Yasutake Takahashi, Koichi Hikita, and Minoru Asada. Incremental purposive behavior acquisition based on self-interpretation of instructions by coach. In *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 686–693, Oct 2003.

7) Stefan Elfwing, Eiji Uchibe, Kenji Doya, and Henrik I. Christensen1. Multi-agent reinforcement learning: Using macro actions to learn a mating task. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. CD–ROM, Sep 2004.

8) Tomoki Nishi, Yasutake Takahashi, and Minoru Asada. Incremental behavior acquisition based on reliability of obser ved behavior recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 70–75, Oct 2007.

9) Yasutake Takahashi, Kentaro Noma, and Minoru Asada. Efficient behavior learning based on state value estimation of self and others. *Advanced Robotics*, Vol. 22, No. 12, pp. 1379–1395, 2008.

10) Steven D. Whitehead. Complexity and cooperation in q-learning. In *Proceedings Eighth International Workshop on Machine Learning (ML91)*, pp. 363–367, 1991.

11) Matthew E. Taylor, Peter Stone, and Yaxin Liu. Transfer learning via inter-task mappings for temporal difference learning. *Journal of Machine Learning Research*, Vol. 8, No. 1, pp. 2125–2167, 2007.

12) 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.

13) Bob Price and Craig Boutilier. Accelerating reinforcement learning through implicit imitatione. *Journal of Articial Intelligence Research*, 2003.

14) Darrin C. Bentivegna, Christopher G. Atkeson, and Gordon Chenga. Learning tasks from observation and practice. *Robotics and Autonomous Systems*, Vol. 47, pp. 163–169, 2004.

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

16) Satinder Pal Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, Vol. 8, pp. 323–339, 1992.

17) Satinder P. Singh. The effeicient learnig of multiple task sequences. In *Neural Information Processing Systems 4*, pp. 251–258, 1992.

18) Jun Tani and Stefano Nolfi. Self-organization of modules and their hierarchy in robot learning problems: A dynamical systems approach. Technical report, Technical Report: SCSL-TR-97-008, 1997.

19) J. Tani and S. Nolfi. Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. *Neural Networks*, Vol. 12, No. 7-8, pp. 1131–1141, 1999.

20) Klaus-Robert Muller, Kohlmorgen Jens, and Pawelzik Klaus. Analysis of switching dynamics with competing neural networks. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, Vol. E78-A, No. 10, pp. 1306–1315, Oct 1995.

21) P. Hartono and S. Hashimoto. Temperature switching in neural network ensemble. *Journal of Signal Processing*, Vol. 4, No. 5, pp. 395–402, 2000.

22) 鮫島和行, 銅谷賢治, 川人光男. 強化学習 mosaic: 予測性によるシンボル化と見まね学習. 日本ロボット学会誌, Vol. 19, No. 5, pp. 551–556, 2001.

23) Masahiko Haruno, Daniel M. Wolpert, and Mitsuo Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, Vol. 13, pp. 2201–2220, 2001.

24) Yasuo Nagayuki, Shin Ishii, and Kenji Doya. Multi-agent reinforcement learning: An approach based on the other agent's internal model. In *ICMAS*, pp. 215–221, 2000.

25) Tohyama S., Omori T., Oka N., and Morikawa K. Identification and learning of other's action strategies in cooperative task. In *Proc. of 8-th International Conference on Artificial Life and Robotics(AROB8th'03)*, pp. 40–43, 2003.

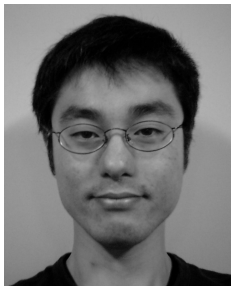
26) Tetsunari Inamura, Yoshihiko Nakamura, and Iwaki Toshima. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, Vol. 23, No. 4, pp. 363–377, 2004.

27) 鮫島和行, 杉本徳和. モジュール強化学習と意図. 人工知能学会誌, Vol. 20, No. 4, pp. 441–448, 7 2005.

28) 片上大輔, 大村英史, 安村禎明, 新田克己. 社会的インタラクションに基づくマルチユーザ学習エージェント（mula）. 日本知能情報ファジィ学会誌, Vol. 17, No. 3, pp. 340–350, 2005.

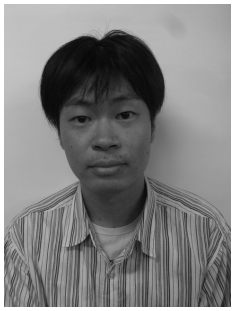
29) 高橋泰岳, 河又輝泰, 浅田稔. 自己の価値に基づく他者行為理解. 日本知能情報ファジィ学会誌, Vol. 21, No. 3, pp. 381–391, Jun 2009.

30) Yasutake Takahashi, Teruyasu Kawamata, Minoru Asada, and Mario Negrello. Emulation and behavior understanding through shared values. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3950–3955, Oct 2007.



著者紹介

高橋 泰岳 （たかはし　やすたけ）[正会員] 1994 年大阪大学大学院工学研究科博士前期課程修了．2000 年同大学博士後期課程中退，同年同大学大学院工学研究科知能・機能創成工学専攻助手．2006 年 6 月より 2007 年 9 月までドイツ Fraunhofer IAIS 客員研究員．2009 年 7 月より現在，福井大学大学院工学研究科知能システム工学専攻講師．博士（工学）ロボカップ中型機リーグや知能ロボットの行動獲得に関する研究に従事．人工知能学会，日本ロボット学会，知能情報ファジィ学会などの会員．田村 佳宏 （たむら　よしひろ）



[非会員] 2008 年大阪大学工学部応用理工学科卒業．現在大阪大学大学院工学研究科知能・機能創成工学専攻博士前期課程．ロボットの行動学習・行為発達に関する研究に従事．浅田 稔 （あさだ　みのる）[非会員]

1982 年大阪大学大学院基礎工学研究科後期課程修了．1995 年大阪大学工学部教授．1997 年大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る．2005 年より JST ERATO 浅田共創知能システムプロジェクト研究総括．認知発達ロボティクスの研究に従事．文部科学大臣賞 (2001) など受賞多数．