価値システムに基づく視野推定

○島田皓樹 (阪大院)　　高橋泰岳 (阪大院)
浅田稔 (JST ERATO, 阪大院)

View Estimation based on Value System

*Kouki SHIMADA (Osaka Univ.), Yasutake TAKAHASHI (Osaka Univ.)
Minoru ASADA(JST ERATO, Osaka Univ.)

Abstract— Estimation of a caregiver's view is one of the most important capabilities for a child in understanding the behavior demonstrated by the caregiver, that is, to infer the intention of behavior and/or to learn the observed behavior efficiently. This paper shows a method for acquiring such a capability based on a value system from which values can be obtained by reinforcement learning. Temporal difference error (hereafter TD error: estimation error of the state value) is utilized not only for behavior learning but also for caregiver view estimation. Experiments with simple humanoid robots show the validity of the method, and the developmental process parallel to young children's view estimation from self to other through the observation of the caregiver's behavior is discussed.

Key Words: view estimation, reinforcement learning, state value, value system

1. はじめに

子どもにとって養育者の視野を推定することは行動理解，行動意図の推定，行動獲得の際に重要である．子どもが養育者の行動を観察し，その行動を理解し模倣しようとするとき，養育者からの視野を推定し，養育者の手の平や物体の軌跡から自身の行動表現へのマッピングをする必要がある．ここで視野推定の能力は報酬に基づく行動学習と同様に，報酬の推定誤差を小さくするように自己から養育者への視野推定モデルを更新させることにより発達するという仮説を提案し，検証する．

子どもによる行動学習は食べ物を食べる喜び，おもちゃで遊ぶ楽しさなどの動機付けによるものと考えられる．目標に到達すると報酬を感じ，報酬を得るために試行錯誤を繰り返すことにより行動学習を行っている．期待される報酬と実際に受け取った報酬の推定誤差と脳のドーパミンニューロンの活性と強い相関があるという報告がされており [1]，この推定誤差が子どもの行動学習に影響していると考えられる．このシステムをモデル化したものが強化学習である [2]．

強化学習はシングルエージェントやマルチエージェント環境における行動学習の手法として多くの研究がされている [3]．学習者は先験的な知識を必要とせず，試行錯誤を繰り返すことで最適な行動を獲得する．最適行動の獲得（状態から行動へのマッピングの学習）だけではなく，状態価値と呼ばれる将来期待される報酬の減衰総和の獲得にも用いられる．状態価値の推定誤差は TD(Temporal Difference) 誤差と呼ばれ，学習者は TD 誤差に基づいて状態価値や行動を更新する．Meltzoff は他者の行動，意図，信条を理解するために自分自身の経験を用いる Like me 仮説を唱えている [4]．この仮説を強化学習の立場からとらえた場合，学習者は他者の行為を観察しているとき，その行為の報酬及び状態価値を推定していると考えられる．

そこで本論文では，強化学習の枠組みにおける TD 誤差に基づく視野推定の手法を提案し，報酬の推定誤差を小さくするように自己から養育者への視野推定のパラメータを更新させることにより，視野推定能力が発達するという仮説を立て検証する．提案手法は視覚システムにおいて先験的な拘束を必要としない．ヒューマノイドシミュレータを用いた実験により，この手法の有効性を示す．

2. TD 誤差に基づく視野推定

2·1 実験設定

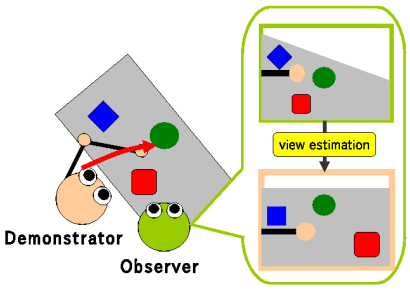


Fig.1 Scenario of the experiment

Fig.1 に実験の外観を表す．2 体のプレイヤーが数個の物体があるテーブルの前にいる．プレイヤーは 2 体とも強化学習に基づき物体へのリーチングの行動を獲得し，状態価値を持つ．行動学習の後，片方が提示者となり物体にタッチし，もう片方が観察者となり提示者の視野を推定する．

2·2 状態価値

エージェントがある状態 s_t を起点として方策 π に従って報酬 r_t を受け取るときの状態価値 $V(s_t)$ は，

$$V(s_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots]$$

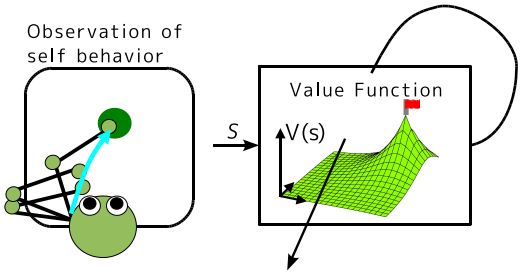


Fig.2 Update of state values based on TD error through trial and error

$$= E[r_t] + \gamma V(s_{t+1}) \ .$$

と定義される．また，状態価値 $V(s_t)$ は学習率 $\alpha(0 \leq \alpha \leq 1)$ を用いて以下のように更新される．

$$V(s_t) \leftarrow V(s_t) + \alpha \Delta V(s_t)$$
$$\Delta V(s_t) = r_t + \gamma V(s_{t+1}) - V(s_t)$$

このとき求められる状態価値の推定誤差 $\Delta V(s_t)$ を TD 誤差と呼ぶ．方策に従って行動をすることで状態が変化し，報酬が与えられることで状態価値が更新される．この状態価値の更新の様子を **Fig.**2 に示す．強化学習の詳細は Sutton et al. の書籍 [2] やロボット学習に関する書籍 [3] を参照されたい．

2·3　自己の状態価値に基づく行為理解

強化学習により状態価値が獲得・更新され，その結果，与えられたタスクを達成するための最適行動（状態から行動へのマップ）が得られる．状態価値の直感的な意味はゴール状態への近さであり，ゴールに近づけば状態価値も大きくなる．このことから観察によって推定した他者の状態価値が上昇したら，他者がゴール状態に近づいていることが認識できると考えられる．Takahashi et al. [5] は獲得済みの自身の状態価値関数に基づき他者の状態価値を推定することで相手の行動を認識するシステムを提案した．

観察中に自身の獲得した状態価値に基づき他者の行動を認識する場合は，エージェントは次の手続きをふむ：

1. 提示者の行動を観察する
2. 提示者からの視野を推定する
3. 推定した視野に基づき状態価値を推定する
4. 推定した状態価値の遷移に基づき他者の行動を認識する

2·4　TD 誤差に基づく視野推定パラメータの更新

観察した他者行動から状態価値へマッピングするためには自己視野から他者視野への視野変換が必要である．ここでは以下の 2 つの仮定を設定する：

- 他者の行為観察による行動学習は行わない．すなわち，自身で獲得した状態価値は更新されない．
- 提示者の行為は常に観察者自身が事前に獲得したリーチング行動であり，最終的にリーチングする物体からさかのぼって，どの物体にリーチングする動作であるかを正確に分類できる．

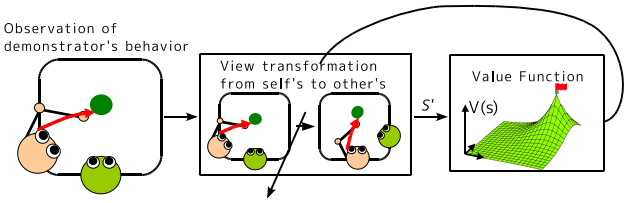


Fig.3 Update of view transformation parameter from the self to the demonstrator based on the TD error

TD 誤差に基づく視点変換行列のパラメータ更新の枠組みを **Fig.**3 に示す．まず，観察者は自己の視野からセンサー情報 ${}^o\boldsymbol{x}$ を観察により獲得する．提示者の視野から獲得されるセンサー情報 ${}^d\boldsymbol{x}$ は変換行列 ${}^d\boldsymbol{T}_o$ を用いて

$${}^d\boldsymbol{x} = {}^d\boldsymbol{T}_o \, {}^o\boldsymbol{x}$$

で表される．変換行列の成分であるパラメータ ϕ_i は推定される TD 誤差 $\Delta\hat{V}_t$ によって更新される．

$$\phi_i \leftarrow \phi_i - \beta \frac{\partial |\Delta V_t|}{\partial \phi_i} \tag{1}$$

ただし，i はパラメータのインデックス，$\beta(0 \leq \beta \leq 1)$ は更新率である．ここで，推定される TD 誤差は

$$\Delta\hat{V}_t = \hat{r}_t + \gamma \hat{V_{t+1}} - \hat{V}_t \ , \tag{2}$$

で表現される．ただし，簡単のため

$$\hat{r_t} = r(\hat{s_t}) \ , \quad \hat{V_t} = V(\hat{s_t}) \ , \quad \hat{s_t} \leftarrow F^{hash}({}^d\boldsymbol{x}_t)$$

と表記する．ここで，F^{hash} とはセンサ情報ベクトル ${}^d\boldsymbol{x}_t$ を状態 $s \in \boldsymbol{S}$ へと変換するハッシュ関数である．

今回，式 (1) における偏微分の項に対して式 (3) の数値微分を行い近似した値を用いる．

$$\frac{\partial |\Delta\hat{V}_t|}{\partial \phi_i} \rightarrow \frac{|\Delta\hat{V}_t({}^d\boldsymbol{x}_t|^{\phi_i + \delta\phi_i})| - |\Delta\hat{V}_t({}^d\boldsymbol{x}_t|^{\phi_i - \delta\phi_i})|}{2\delta\phi_i} \tag{3}$$

ここで，${}^d\boldsymbol{x}_t|^{\phi_i+\delta\phi_i}$ と ${}^d\boldsymbol{x}_t|^{\phi_i-\delta\phi_i}$ は視野推定行列のパラメータ ϕ_i を $\delta\phi_i$ だけ増減した時に推定される提示者のセンサ情報ベクトルである．この線形近似をにより値を求めることで状態価値関数の不連続性の問題を避けている．

2·5　視野推定

エージェントはそれぞれ頭にカメラを持ち，カメラ画像上での物体の位置情報を獲得する．観察者の自己視野からの情報を ${}^o\boldsymbol{x} = ({}^ox, {}^oy)$，提示者の視野からの情報を ${}^d\boldsymbol{x} = ({}^dx, {}^dy)$ としたとき，視点変換行列の成分 $\boldsymbol{\phi}$ を用いて式 (4) で表す．

$$\begin{pmatrix} {}^dx \\ {}^dy \\ 1 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} {}^ox \\ {}^oy \\ 1 \end{pmatrix} \tag{4}$$

視野変換行列は観察者と提示者の位置関係に依存する．以下の実験では観察者と提示者の位置は視野変換パラメータの推定中は固定する．

3. ヒューマノイドロボットを用いた実験

3·1 ヒューマノイドシミュレータ

環境中には 2 体のヒューマノイドロボットとテーブルの上には色の付いた箱が存在する．ヒューマノイドロボットは 2 自由度で腕を動かすことが出来る．手の平と物体はロボットの頭にあるカメラにより認識される．テーブルの大きさ，観察者と提示者の位置関係は **Fig.**6(a) の通りである．以下の実験ではロボットの視野中心を ${}^{o}\boldsymbol{x}$ の原点とした．

3·2 視野推定実験

まず，観察者は強化学習における状態価値を物体へのリーチングを行うことで獲得する．カメラ画像上における手の平の x, y 座標を状態変数とし，それぞれ 30 分割して状態空間を作る．観察者はそれぞれの物体へのリーチング行動毎に状態価値関数を用意する．観察者の手の平は状態空間において接触できる全ての位置から一つの物体へのリーチングを行い，その行動に対応した状態価値を学習する．

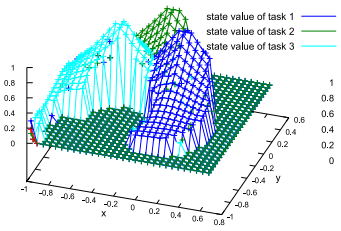


Fig.4 State value functions: the learner reaches for 3 objects

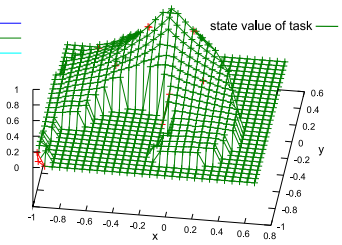


Fig.5 State value function: the learner reaches for 1 object

Fig.4 は 3 つの物体にリーチングする行動それぞれの状態価値関数を示している．この状態価値関数の形は円錐状ではなく，緩急のある勾配の山の形をしている．これは腕の長さや関節角度の影響で可動領域や動作速度に制限があるためである．

次に観察者は提示者の行動を観察する．提示者は様々な初期位置から手の平を動かし観察者が事前に獲得した物体へのリーチング行動を提示し，観察者は近くで観察することにより獲得済みの状態価値関数から，状態価値を推定し視点変換行列を更新する．視野変換行列は単位行列に初期化している．また，式 (1) の更新率は $\beta = 0.01$ に固定する．

価値システムに基づく視野推定を評価するために以下の 2 つの場合について実験を行った．

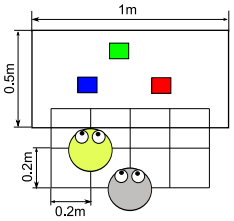
- Case 1: 提示者が 3 つの物体にリーチングをする場合．観察者と提示者は同じ身体を持つ．
- Case 2: 提示者が 1 つの物体にリーチングをする場合．観察者と提示者は同じ身体を持つ．

工学的な観点から視野変換行列を推定するためには 2 点 (ここでは到達点に対応) 以上の情報が必要である．Case 1 は画像上の与えられた位置情報が視野変換に十分であるときの実験で，Case 2 は与えられた位置情報が視野変換に不十分であるときの実験である．

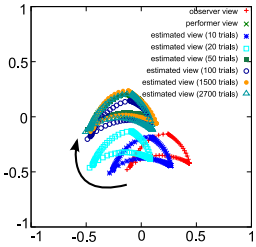
推定した視点変換行列を評価するために，いずれの実験でも提示者が 3 つの物体に一つずつリーチングする様子を観察者に提示し，観察者が提示者の視野での軌道を推定し，それを図に描画する．もし視点変換行列が真の値に近ければ，推定した軌道が提示者からの視野の軌道と一致する．

3·2.1 Case 1: 3 物体へのリーチング

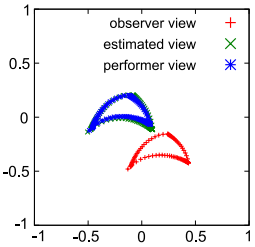
Case 1 では，提示者が 3 つの物体にリーチングを行い，同じ大きさの身体を持つ観察者が視野を推定する．**Fig.**6, **Fig.**7 に実験結果を示す．**Fig.**6 では，それぞれ (a) に観察者，提示者の位置関係の様子，(b) に視点変換行列の推定過程（視野推定の履歴），(c) に視野推定パラメータ学習後の推定結果を示す．**Fig.**6(b) は観察者が提示者の前 0.2m，左 0.2m の位置で推定した視点変換行列の履歴を表している．推定される軌跡が真の値に近づいており，提示者の視点を十分に推定していると言える．同様の実験を様々な位置に配置させて行った場合（**Fig.**7(a) 参照），25 箇所のうち 21 箇所において推定される軌跡が真の値に近づいていた．視野推定が失敗した箇所は観察者の位置が提示者の手の平が見える限界である前方 0.4m の位置がほとんどで，前 0.4m，左 0.4m の箇所，前 0.4m，左 0.2m の箇所，前 0.4m，右 0.4m の箇所，後 0.4m，左 0.4m の箇所であった．また，観察者の位置をテーブルの回りに移動させた場合（**Fig.**7(b) 参照），提示者から 120 度移動した位置からでも視野の推定ができ，広い範囲で正しく推定できることが示された．



(a) experimental set-up



(b) changes of the view transform parameters

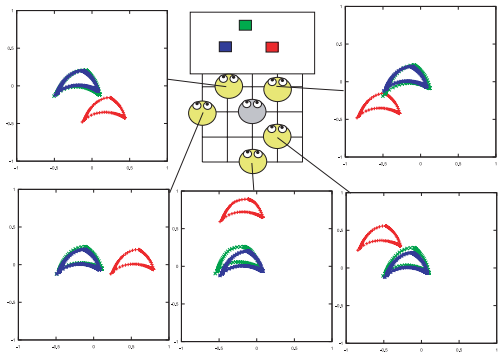


(c) final result

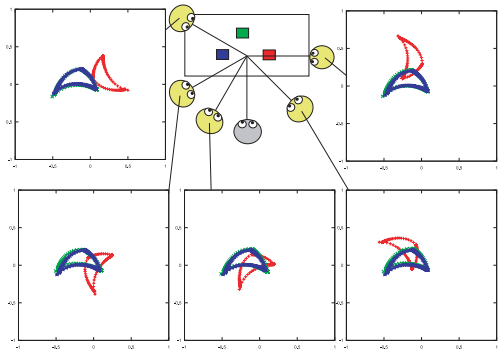
Fig.6 Estimation of the demonstrator's view: The observer posture is parallel to the demonstrator.

3·2.2 Case 2: 1 物体へのリーチング

Case 2 では提示者が 1 つの物体にリーチングを行い，同じ大きさの身体を持つ観察者が視野を推定したときの場合の結果を表す．**Fig.**5 で表される状態価値関数を用いて視野推定を行った．その結果を **Fig.**8 に示

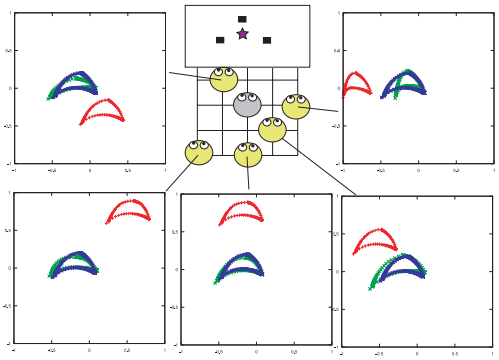


(a) parallel posture

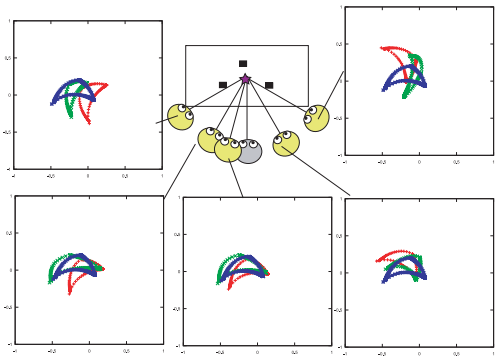


(b) rotated posture

Fig.7 Estimation results of trajectories in the demonstrator's view while it reaches for 3 objects one-by-one. Blue, red, and green curves indicate the trajectories on the demonstrator's view, observer's view, and one estimated by the observer.



(a) parallel posture



(b) rotated posture

Fig.8 Estimation results of trajectories in the demonstrator's view while it reaches for only one object. Colors indicate in the same way as **Fig.**7

す．図中の星印は観察者がリーチング学習時の目標位置である．観察者が提示者に対して平行移動した場合（**Fig.**8(a) 参照），25 箇所のうち 15 箇所において推定結果が真の値に近づいたが，回転した場合（**Fig.**8(b) 参照）は提示者から 30 度以上移動した位置から観察すると真の値に近づかなくなった．

3·3 考察

提示者が 3 つの物体にリーチングし，観察者がその提示から視野を推定する場合，広い範囲で正しく視野推定が可能であることが示された．また，本手法を用いると Case 2 のように，工学的な観点から視野変換行列を推定するには不十分である 1 点のみの情報からでも，視野を推定することができた．

視点変換行列のパラメータの更新に対して状態価値関数の形状が緩急のある勾配の山の形をしていることが影響している．この形状のために，観察者からの視野と提示者からの視野が異なる場合は推定される TD 誤差が大きいため，視野変換行列のパラメータが真の値に向けて更新されている．

4. まとめ

本論文ではエージェントが強化学習における TD 誤差に基づき視野推定の能力を発達させるという仮説を提案した．状態価値のパラメータが TD 誤差によって更新されるように，視野推定のパラメータも TD 誤差によって更新している．提案手法は視覚システムに関する幾何学的なパラメータの事前知識を必要としない．ヒューマノイドシミュレータを用いた実験によりこの手法の有効性を示した．

参考文献

[1] Jeffrey R. Hollerman and Wolfram Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, (1):304–309, 1998.

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

[3] Jonalthan H. Connell and Sridhar Mahadevan. *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.

[4] Andrew N. Meltzoff. 'like me': a foundation for social cognition. *Developmental Science*, 10(1):126–134, 2007.

[5] 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*, pages 3950–3955, Oct 2007.