異なる体格間での状態価値推定誤差に基づく視野推定能力学習

Acquisition of View Estimation between Different Bodies based on State Value Estimation

島田皓樹 Kouki Shimada 大阪大学 Osaka University
高橋泰岳 Yasutake Takahashi 大阪大学 Osaka University
浅田稔 Minoru Asada 大阪大学/JST ERATO Osaka University/JST ERATO

Abstract: Estimation of a caregiver's view is one of the most important capabilities for a child to understand the behavior demonstrated by the caregiver, that is, to infer the intention of behavior and/or to learn the observed behavior efficiently. We hypothesize that the child develops this ability in the same way as behavior learning motivated by an intrinsic reward, that is, he/she updates the model of the estimated view of his/her own during the behavior imitated from the observation of the behavior demonstrated by the caregiver based on minimizing the estimation error of the reward during the behavior. From this view, this paper shows a method for acquiring such a capability based on a value system from which values can be obtained by reinforcement learning. Experiments with simple humanoid robots show the validity of the method, and the developmental process parallel to young children's estimation of its own view during the imitation of the observed behavior of the caregiver is discussed.

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．本論文において他者の観察を通した行動理解とは，子どもが他者の行為を認識し，行為を行う中で得られる報酬を推測し，そして自分自身でその報酬が得られる行為を再現できることを意味する．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，その行動を自身で行うときの視野を推定し，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．

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 [1]．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 [2]．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．

模倣学習の従来研究の多く (例えば [3, 4]) は観察される動作を模倣するために，観察者は提示者の動作の軌跡を世界座標系や提示者自身の関節空間へ変換し，ジェスチャーや行動を模倣している．一方 Asada et al. [5, 6] はエピポーラ幾何に基づく視野復元の手法を提案している．また Yoshikawa et al. [7] は自身の視野に基づいた他者の視野推定の学習手法を提案している．これらの視野変換の学習は幾何学的拘束と提示者の姿勢推定に基づいており，基本的に模倣による行動学習とは独立している．

工学的には視野変換は 3 次元空間内での平行移動・回転・拡大縮小といった座標変換に加えて，3 次元空間の情報を 2 次元情報に変換する投影変換の組み合わせで構成される [8]．この視野変換のパラメータの推定は，二つの画像における複数の対応点がわかりさえすれば，その間の視野変換行列のパラメータは最小二乗法等で推定でき，幼児に見られる視野推定能力の発達を説明できるモデルになるとは考えにくい．

Meltzoff が提唱している Like Me 仮説 [2] を強化学習の視点から捉えれば，相手の行動を見て相手が受け取っているであろう報酬を自分自身の経験に照らし合わせて予測していると考えられる．感覚と運動および報酬のマッピングはグローバル座標系ではなく自己中

心座標系に基づいていると考えられるので，他者行動を理解・模倣するためには模倣する際の自己の視野を推定する能力が必要である．これまでの議論から，他者行動の観測から自身の感覚に写像し，報酬を推定する為の視野推定能力を獲得させるため，報酬に基づく行動学習と同様に，報酬の推定誤差を視野推定モデルにフィードバックすることにより発達するという仮説が考えられる．

そこで島田ら [9] 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提案し，その具体的な手法をヒューマノイドシミュレータを用いた実験により検証した．しかし，その報告では学習者と養育者の体格が同一であるという条件で検証していた．本報告では視野推定能力を獲得する前の幼児と養育者を想定し，観測者は提示者の体の 3 分の 2 の大きさで重さも異なる条件下で検証する．

2　TD 誤差に基づく視野推定

環境中に行動の提示者と観察者がいると仮定する．観察者は提示者の行動を観察し，提示者が行動中に受け取る報酬と状態価値の予測をするために，提示された行動を自身が実行しているときの視野を推定する．観察情報を自己の感覚に写像し，報酬と状態価値を推定することで他者の行動を理解するための視野推定を目的とするため，提示者自身の視野を正確に推定するのではなく，あくまで提示された行動を自分で行っているときの自分の視野を推定する．ここで観察者と提示者の位置関係が決まれば，視野推定のための変換パラメータは与えられたタスクに依存しない．しかしその視野変換能力は自身の身体や行動を基に学習し，そこで獲得した視野変換能力を基に他者行動を観察し，自己の視野空間にマップすることで，より深い行動理解ができると考えられる．従って，想定される視野推定能力の為の発達過程は以下の通りになる．

1. 自己の試行錯誤による行動学習
2. 報酬推定のための視野推定能力学習
3. 自己感覚系へのマッピングに基づく他者行動の理解と模倣

自己の試行錯誤による行動獲得は従来から多く強化学習の分野で研究がなされている．本論文では (2) の観察による視野変換能力獲得に焦点を当てる．視野変換能力が獲得された後は，この能力を基に観察者にとって新規である提示行動の観察を通して自己の視野空間にマップすることで，提示行動の理解および模倣が可能となる．

2.1　実験設定

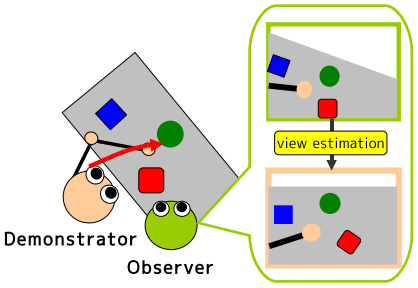


図 1: Scenario of the experiment

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，

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

と定義される．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．

$$V(s_t) \leftarrow V(s_t) + \alpha \Delta V(s_t) \quad (3)$$
$$\Delta V(s_t) = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (4)$$

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 2(a) に示す．強化学習の詳細やその応用については，文献 [10] や [1] 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

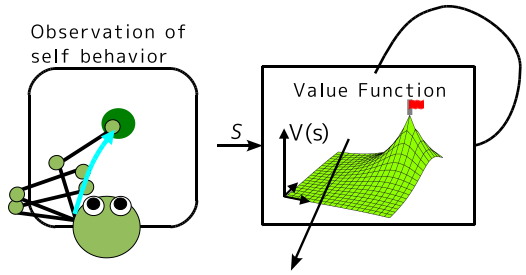


図 2: Update of state values based on TD error through trial and error

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. [11] 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

他者行動の観察情報から自己の状態価値へマッピングするためには観察者が観察しているときの視野から自身が同じ行動を実行するときの視野への変換，すなわち視野推定が必要である．ここでは以下の仮定を置く:

- 通常の強化学習における仮定と同じように，観測中の状態遷移はマルコフ過程である．
- 提示される行動は観察者にとって既に獲得済みである．すなわち，自身で獲得した行動の状態価値は更新されない．
- 提示者の行動は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，提示された動作がどの物体にリーチングする動作であったかを正確に分類できる．
- 観察者と提示者の位置は視野推定パラメータの推定中は固定する．

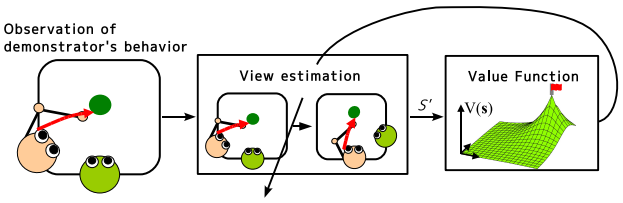


図 3: Update of view transformation parameter from the self to the demonstrator based on the TD error

TD 誤差に基づく視野推定行列のパラメータ更新の枠組みを 2(b) に示す．2(a) に示した状態価値推定のための TD 誤差フィードバックとは異なり，TD 誤差を視野推定行列のパラメータの更新に用いている．

まず，観察者は自己の視野からセンサ情報 ${}^{o}\boldsymbol{x}$ を取得する．観察者が提示された行動を模倣したときに取得されるセンサ情報 ${}^{d}\boldsymbol{x}$ は推定行列 ${}^{d}\boldsymbol{T}_{o}$ を用いて

$$ {}^{d}\boldsymbol{x} = {}^{d}\boldsymbol{T}_{o}\ {}^{o}\boldsymbol{x} \tag{5} $$

で表される．なお，添え字の o および d はそれぞれ観察時 (observation) および実行時 (demonstration) を表す．推定行列の成分であるパラメータ ϕ_{ij} は推定される TD 誤差 $\Delta\hat{V}_t$ によって更新される．

$$ \phi_{ij} \leftarrow \phi_{ij} - \beta \frac{\partial |\Delta V_t|}{\partial \phi_{ij}} \tag{6} $$

ただし，i, j 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

$$ \Delta\hat{V}_t = \hat{r}_t + \gamma \hat{V}_{t+1} - \hat{V}_t \ , \tag{7} $$

で表現される．ただし，簡単のため

$$ \begin{aligned} \hat{r}_t &= r(\hat{s}_t) \\ \hat{V}_t &= V(\hat{s}_t) \\ \hat{s}_t &\leftarrow F^{hash}({}^{d}\boldsymbol{x}_t) \end{aligned} $$

と表記する．$\hat{s}_t$，$\hat{r}_t$，$\hat{V}_t$ はそれぞれ観察者が提示された行動を模倣したときに得られると推定される状態，報酬，状態価値である．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 (6) における偏微分の項に対して式 (8) の数値微分を行い近似した値を用いる．

$$ \frac{\partial |\Delta\hat{V}_t|}{\partial \phi_{ij}} \rightarrow \frac{|\Delta\hat{V}_t({}^{d}\boldsymbol{x}_t|^{\phi_{ij}+\delta\phi_{ij}})| - |\Delta\hat{V}_t({}^{d}\boldsymbol{x}_t|^{\phi_{ij}-\delta\phi_{ij}})|}{2\delta\phi_{ij}} \tag{8} $$

ここで，${}^d\boldsymbol{x}_t|^{\phi_{ij}+\delta\phi_{ij}}$ と ${}^d\boldsymbol{x}_t|^{\phi_{ij}-\delta\phi_{ij}}$ は視野推定行列のパラメータ ϕ_{ij} を $\delta\phi_{ij}$ 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 (9) で表す．

$$\begin{pmatrix} {}^dx \\ {}^dy \\ 1 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} {}^ox \\ {}^oy \\ 1 \end{pmatrix} \quad (9)$$

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部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．ここで観測者と提示者はそれぞれ視野推定能力を獲得する前の幼児と養育者を想定し，観測者は提示者の体の 3 分の 2 の大きさ，つまり腕の長さも含むリンク長がすべて 3 分の 2 の長さを持ち，同じリーチング動作を行う際の軌道と速度が異なる．また，観察者と提示者の身長差からカメラの高さが異なり，観察者は提示者よりも近くで行動を観察するため，提示者の動きが拡大されて見える状況で，スケール変動に対する頑強性も持たなくてはならない．

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．それぞれの物体へのリーチングが達成された状態に正の報酬 (+1) が与えられ，それ以外の状態では 0 の報酬が与えられる．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．

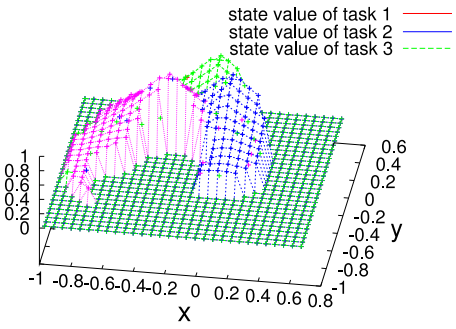


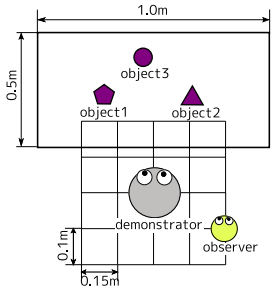
図 4: Three state value functions for reaching behavior to each of three objects

5 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．

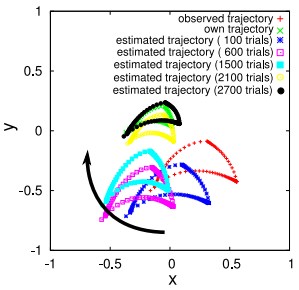
次に観察者は提示者の行動を観察する．提示者は様々な初期位置から手の平を動かし観察者が事前に獲得した物体へのリーチング行動を提示し，観察者は近くで観察することにより獲得済みの状態価値関数から，状態価値を推定し視野推定行列を更新する．視野推定行列は単位行列に初期化している．また，式 6 の更新率は $\beta = 0.01$ に固定する．

推定した視野推定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，観察者が提示された行動を自分で行ったときの自身の視野での軌道を推定し，それを図に描画する．もし視野推定行列が真の値に近ければ，推定した軌道が観察者が同じ行動を実行したときの軌道と一致する．

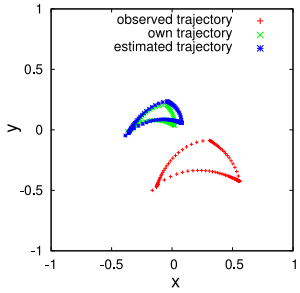
6, 7, 8, 9 に実験結果を示す．6, 7 では，それぞれ (a) に観察者，提示者の位置関係の様子，(b) に視野推定行列の推定過程（視野推定の履歴），(c) に視野推定パラメータ学習後の推定結果を示す．図 6(b) は観察者が提示者の後 0.1m，右 0.3m の位置で推定した視野推定行列の履歴，図 7(b) はテーブルの回りで観察者が提示者から 105 度移動した位置で推定した視野推定行列の履歴を表している．それぞれ 2700 試行の観察を通して視野推定行列のパラメータを更新させた．共に観察回数が増えるにつれ，推定される軌跡が真の値に近



(a) experimental set-up

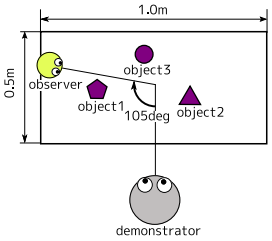


(b) changes of trajectory on estimated view

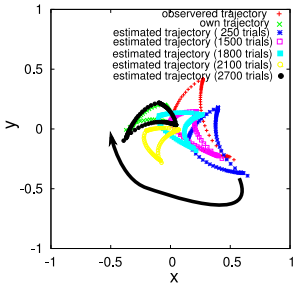


(c) final estimation

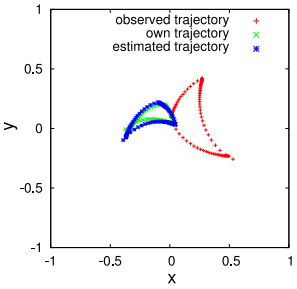
図 5: Estimation of view of self during imitation of observed behavior: Observer posture is parallel to the demonstrator.



(a) experimental set-up



(b) changes of trajectory on estimated view



(c) final estimation

図 6: Estimation of view of self during imitation of observed behavior: Observer orients to the demonstrator with 105 degree rotation.

づいており，提示者の視野を十分に推定していると言える．同様の実験を観察者を様々な位置に配置した場合の結果を 8 に示す．観察者の位置を提示者に対し平行に移動させた時の視野推定結果の様子と推定の成功失敗の分布を図 8(a) 及び 9(a) に示す．25 箇所のうち 20 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 0.2m の位置又は後方 0.2m での位置がほとんどであった．また，観察者の位置をテーブルの回りに移動させた場合（図 8(b) 及び 9(b) 参照），提示者から 105 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．

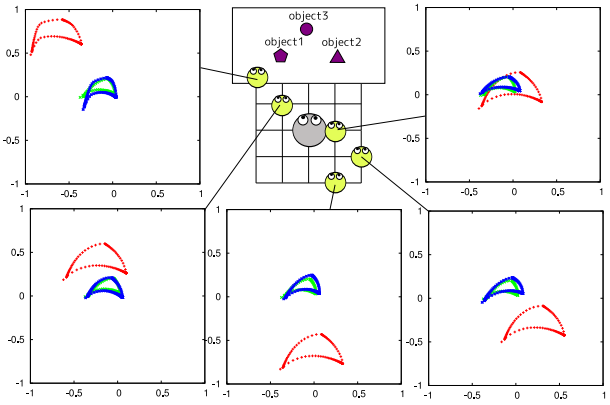
4 まとめ

本論文では，視野推定の能力は報酬に基づく行動学習と同様に，報酬の推定誤差を視野推定モデルの更新に利用することにより視野推定能力が発達するという仮説を提案し，検証した．Meltzoff が提唱している Like Me 仮説を強化学習の視点から捉え，相手の行動を見て相手が受け取っているであろう報酬を自分自身の経験に照らし合わせて予測していると仮定した．感覚と運動のマッピングはグローバル座標系ではなく自己中心座標系に基づいているので，他者行為を理解・模倣するためには視野推定能力が必要である．他者行動の観測から自身の感覚に写像し，報酬を推定する為の視野推定能力を獲得させるため，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た．具体的には，強化学習における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た．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．本論文中の実験では観察者と提示者の位置は視野推定パラメータの推定中は固定していたが，視野推定パラメータに観察者と提示者の位置関係の情報を反映させる枠組みに拡張することで，観察者と提示者の位置関係が変動する環境下でも視野推定能力が獲得できるものと考えられる．

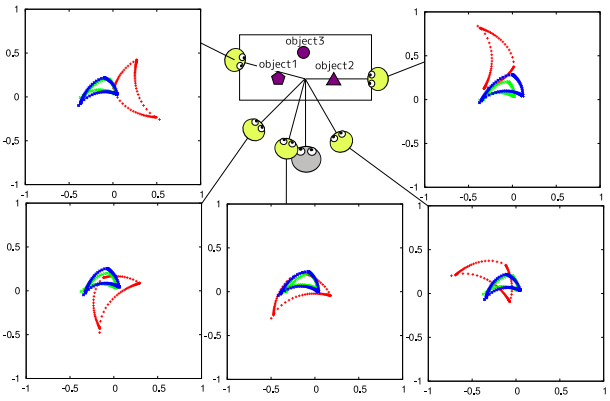
参考文献

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

[2] Andrew N. Meltzoff. 'like me': a foundation for

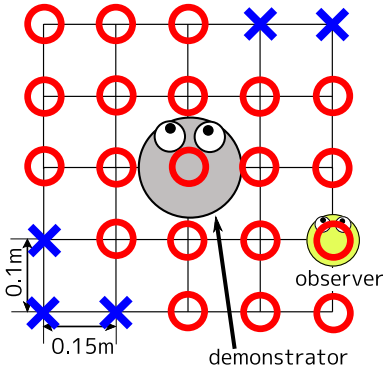


(a) parallel translation

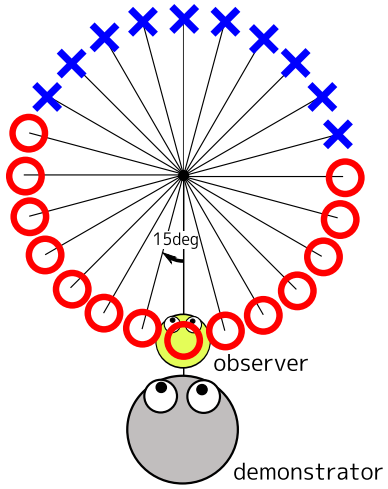


(b) rotational translation

図 7: Final estimated trajectories in various postures. Red, green, and blue curves indicate trajectories in case of observation, estimation, and imitation.



(a) parallel translation



(b) rotational translation

図 8: Success or failure of view estimation in various posture: Circle and cross indicate success and failure, respectively

social cognition. *Developmental Science*, 10(1):126–134, 2007.

[3] Stefan Schaal, Auke Ijspeert, and Aude Billard. Computational approaches to motor learning by imitation, 2004.

[4] Tetsunari Inamura, Yoshihiko Nakamura, and Iwaki Toshima. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, 23(4):363–377, 2004.

[5] Minoru Asada, Yuichiro Yoshikawa, and Koh Hosoda. Learning by observation without three-dimensional reconstruction. In *Proceedings of Intelligent Autonomous Systems (IAS-6)*, pages 555–560, 2000.

[6] Yuichiro Yoshikawa, Minoru Asada, and Koh Hosoda. View-based imitation learning by conflict resolution with epipolar geometry. In *Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1416–1427, 2001.

[7] Yuichiro Yoshikawa, Minoru Asada, and Koh Hosoda. Developmental approach to spatial perception for imitation learning: Incremental demonstrator's view recovery by modular neural network. In *Proceedings of the 2nd IEEE-RSA International Conference on Humanoid Robot*, pages 107–114, 2001.

[8] 田村 秀行. **コンピュータ画像処理**. オーム社, 12 2002.

[9] 島田 皓樹, 高橋 泰岳, and 浅田 稔. 価値システムに基づく視野推定. In **第 *26* 回日本ロボット学会学術講演会予稿集**, volume CD-ROM, pages 1N1–04, Sep 2008.

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

[11] 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.