他者観察を通した注目度に基づく模倣と行為の累増的獲得

Incremental Behavior Acquisition Based on Reliability of Observed Behavior Recognition

○ 西 智樹 (阪大)　正 高橋 泰岳 (阪大, 阪大 FRC)　正 浅田 稔 (阪大, 阪大 FRC)

Tomoki NISHI, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Yasutake TAKAHASHI, HANDAI Frontier Research Center, Osaka University
Minoru ASADA, HANDAI Frontier Research Center, Osaka University

We propose a novel approach for acquisition and development behaviors through observation of behaviors in multi-agent environment. Observed behaviors of others gives fruitful hints to find a new situation, a new behavior for the situation, necessary information for the behavior acquisition. RoboCup scenario gives us a good test-bed multi-agent environment where a learner can observe behaviors of others during practices or games. It is more realistic, practical, and efficient to take advantages of observation of skilled players than to discover new skills and necessary information only though the interaction of robot and environment. The robot automatically detects state variables and a goal of the behavior through the observation based on mutual information. Reinforcement learning method is applied to acquire the discovered behavior suited to the robot. Experiments under RoboCup Middle Size scenario shows the validity of the proposed method.

Key Words: reinforcement learning, multi-layered learning system

1 はじめに

人の特筆すべき能力の一つに模倣がある．幼児は親などの他者の観察から様々な行為に興味を持ち模倣することで，行為のレパートリを増加させ発達していく．また Meltzoff[1] は模倣の能力が他者の意図や心情を理解する上でも重要な役割を担っていると主張している．このように模倣は人が成長し，社会において生活を営む上で極めて重要な能力であると考えられる．

一方近年 AIBO や ASIMO をはじめとする様々なロボットが研究・開発されてきている．しかしながら生活の場において人の代わりに，または人と協調しながら作業を行なうロボットはまだまだ少なく，今後研究・開発していく必要がある．このような環境では環境は流動的であるため設計者が予めロボットの動作を全て埋め込んでおくことは困難であり，ロボットには自律的にその場に適した行為を獲得する能力が求められる．この時，人や他のロボットと共存する環境下では，一からタスクを自身の経験のみで発見する必要性は少なく，むしろ他者の行為から新たな行為を見出す方が現実的であり，有用性も高い．そこで本研究では模倣をロボットにより実現することで，他者行為の観察を通した自律的なタスクの発見と行為の累増的獲得を目指す．

これまでにも他者行為の観察に基づく模倣学習に関する研究は数多く行なわれてきた [2, 3]．その多くは他者行為の軌道を教師とすることにより効率的に行為の軌道を再現するということに重点がおいている．しかしながら，他者行為の軌道から正確に自身が再現すべき軌道を生成するめには様々な仮定や設備が必要であり，工場などの管理された環境以外では実現が難しい．またその他にも軌道を模倣するため，例示された場面とは異なる状況では適切な動作をすることができないという問題もある．

このような軌道そのものを再現することによる模倣は mimicry と呼ばれ，最も原始的な模倣に位置づけられている．幼児の模倣には mimicry の他にも行為の結果を再現する emulation，行為の裏にある意図を再現する imitaiton などがある．emulation による模倣は他者行為の結果を再現することであるため観察から軌道自体を正確に知る必要はなく，また同じ結果に至る異なる軌道も一つの行為として獲得することができる．

行為の目的を報酬というスカラ値を用いて表現し，これに基づき環境との相互作用を通して自律的にタスクへの対処法を獲得する手法として強化学習が注目を集めている [4]．強化学習とは未知の環境中で試行錯誤することにより，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．しかしながら単純な強化学習により複雑な環境やタスクに対する行為を獲得しようとした場合には学習時間や学習に必要な記憶容量が増大し，学習が困難となることがよく知られている．このような問題に対して，行為の自律的な分節化を行ないながら行為を累増的に獲得することを目指した研究に谷口ら [7] の研究がある．

谷口らは強化学習モジュールを並列に配置し，報酬予測モデルとの予測誤差に基づきモジュールを同化及び分化を行なう強化学習システムを提案している．このシステムでは自身と環境との関係を記述する空間である状態空間と学習者が獲得する全ての行為に関する適切な逐次報酬関数を予め設計者が設計しておく必要があるが，これは容易なことではない．そのため学習者が利用する報酬は学習者自身が自律的に設計した報酬関数から生み出されるか、また学習者のタスクの成功または失敗といった第三者でも判断できる基準により与えられることが望ましい．

一方 Baldwin et al.[8] は 10〜11ヶ月の幼児が既に行為の分節化能力を持っているということを実験により示し，幼児は身体の軌道変化の連続性というような低いレベルのモデルを持っており，モデルから逸脱したところを行為の切れ目としているのではないかと主張している．また Itti and Baldi[9] はデータの取得前後でモデルのパラメータが大きく変化した領域をヒトは注視するということを実験により示した．これらの知見から，ヒトが注視しやすいモデルのパラメータが大きく変化する領域を行為の切れ目とすることにより，うまく行為を分節化でき

るのではないかと考えられる．

そこで本研究では他者行為の観察から行為を自律的に分節化，再統合することにより行為の結果の模倣と行為の累増的な獲得を実現する．この時分節化を行なうために各センサ値について軌道変化の連続性を表現する一次元局所線形モデルを導入し，モデルが大きく変化した領域と関連が深い状態空間と目標状態を相互情報量により決定することで自律的にモジュールを構築する．また Takahashi et al.[5] により提案された階層型強化学習機構を用い，ロボットが自律的に構築したモジュールの切替え方を学習することにより emulation が可能となる．さらに提案手法で複数の行為の観察から模倣することで行為を累増的に獲得することができると考えられる．

2 アルゴリズム

2.1 基本アイデア

Baldwin et al.[8] と Itti and Baldi[9] の知見から，ヒトが注視しやすいモデルのパラメータが大きく変化する領域を行為の切れ目とすることにより，うまく行為を分節化できるのではないかと考えられる．軌道変化の連続性とは同じ行為を行なっている間はほぼ同じ方向や速度で変化するということを表しており，これは局所線形モデルによりモデル化できる．また軌道変化の連続性が崩れるということは軌道データの線型性が成り立たなくなるということを表しているため，局所線形モデルのパラメータの信頼度の変化量を測ることにより見付けることができ，この量のことをここでは注目度と呼ぶことにする．また行為の切れ目はある行為が完了した地点であるためその地点と各状態空間の状態との関連性を相互情報量を用いて測ることでその行為を実現する上で重要な空間や目標状態を見付けることが出来る．

そこで本研究では他者行為の観察から計算される注目度を用いて行為の切れ目となり得る領域を見付け，その領域と関連が深い状態空間と目標状態を相互情報量により決定することにより自律的に学習器を構築する．この時すでに獲得済みの学習器により行為が実現できる領域では注目度を抑制することで，あらたな行為を獲得する必要がある領域と関連が深い状態空間と目標状態を決定することができると考えられる．

また行為の結果を再現できた場合に報酬を与えることで，観察に基づき獲得された複数のモジュールの切替え方を階層型学習機構の枠組を用いて学習することにより，試行錯誤を通じて行為の結果を再現する，つまり emulation することが可能となる．さらに提案手法では行為を複数の小さな行為に分節化しながら行為を獲得するため，複数の行為を観察することにより以前獲得した小さな行為を再利用しながら，新たな行為を累増的に獲得していくことができる．

2.2 アルゴリズム概要

観察者は以下の手順を繰り返すことで模倣および行為の累増的獲得を実現する．

1. 例示を観察
2. 注目度に基づき行為と関連性が高い状態空間及び目標状態の決定
3. 関連性が高い状態空間があり
   かつその空間を状態空間として持つ学習器が未獲得の場合は生成，学習し，再び 2 へ
4. 生成した学習器が複数存在する場合には上位層を生成，学習

階層型学習機構については Takahashi et al.[5] の手法を用いるため以下では特に状態空間および目標状態の選択方法について説明する．

2.3 状態空間と目標状態の選択システム

状態変数が多数存在する時，状態空間を構成する状態変数の組合せは無数に存在するが，学習可能な空間であるという仮定をおくと数を限定することができる．この学習可能な状態空間というのは最適性を無視すればタスクとは無関係に決めることが出来る．そのため提案手法では学習可能な状態空間のセットは既知であるとし，この学習可能な状態空間のセットの中でどの状態空間が獲得したい行為を学習する上で適切かを選ぶことで学習モジュールを構築する．以下ではシステムの詳細について説明する．

2.3.1 注目度

一次元局所線形モデル $x = a(t)t + b(t)$ 測定値のばらつきが正規分布に従うと仮定すると，パラメータの誤差の分散は以下のように推定される．

$$\sigma_0^2 = \frac{1}{N-2}(\sum x_i^2 - b\sum x_i - a\sum t_i x_i)$$
$$\sigma_a^2 = \frac{N}{N\sum t_i^2 - (\sum t_i)^2}\sigma_0^2$$
$$\sigma_b^2 = \frac{\sum t_i^2}{N\sum t_i^2 - (\sum t_i)^2}\sigma_0^2$$

このことから局所線形モデル m のパラメータの時刻 t での信頼度 $R_m(t)$ を

$$e(t) = \sqrt{(\sigma_a^2(t))^2 + (\sigma_b^2(t))^2}$$
$$R_m(t) = \begin{cases} 1 - e(t) & if \quad e(t) < 1 \\ 0 & \text{else} \end{cases}$$

と定義する．この式は信頼度が高いほど得られたデータが線形に近いことを示している．そこで各状態変数について一次元局所線形モデルを考え，時刻 t での注目度 $A(t)$ を各モデルにおける信頼度の変化量の総和と定義する．つまり

$$A(t) = \sum_{m \in M} |R_m(t+1) - R_m(t)|$$

で表現される．この時注目度 $A(t)$ が閾値 k よりも大きい領域をモデルが大きく変化した領域とする．

2.4 注目度の抑制

強化学習で利用される状態価値 V は目標状態の近さを表現している．そのため，教示者が学習器のモデルに従って動いている場合には状態価値は上がっていき，また反対に教示者が学習器のモデルに従って動いていない場合には状態価値は減少する傾向にある．そこで学習器モデルの時刻 t での学習器 l の信頼度 $R_l(t)$ を以下のように定義する．但し信頼度の初期値は 0.5 とし，k_1，k_2 はそれぞれシグモイド関数の傾きを決める係数と $x(t)$ の最大値である．

$$R_l(t) = \frac{1}{1 + \exp(-k_1 x(t))}$$
$$x(t) + = \begin{cases} 0 & if x(t-1) > k_2 \\ & \text{or} x(t-1) < -k_2 \\ V(t) - V(t-1) & \text{else} \end{cases}$$

これにより他者の行為がどれくらいこの学習器のモデルに従っているかを評価することが可能となる．つまり他者行為を観察した時に学習器モデルが信頼度が高いということは既にその領域を実現する行為を獲得していると

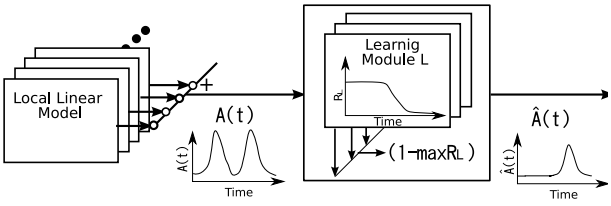


Fig.1 Sketch of the system in order to suppress the degree of attention by learning modules

いうことである．そこで以下のような式により既に獲得済みの学習器の信頼度に基づき注目度を抑制し，その値を $\hat{A}(t)$ とする (Fig. 1)．但し $\max_{l \in L} R_l(t)$ は既に獲得済みの学習器の中での信頼度の最大値である．

$$\hat{A}(t) = (1 - \max_{l \in L} R_l(t))A(t)$$

注目度をこのように抑制することによってまだ行為が獲得されていない領域でのみ注目度が大きくなる．

2.5 状態空間と目標状態の選択

相互情報量 I(X;Y) とは事象 X を観察することで得ることができる事象 Y の情報の量のことであり，互いの事象の関連の深さを評価ことができる．このことから「$\hat{A}(t)$ が閾値よりも大きい」という事象と「状態空間 S で状態 s をとる」という事象との相互情報量が大きい状態空間を選択することで，行為の切れ目に関連した状態空間を持つ学習器を獲得できると考えられる．

3 実験

3.1 実験設定

提案手法の有効性について，ロボットのサッカー大会である RoboCup 中型リーグの環境を模したシミュレータ及び実際のロボットを用いて検証した．ロボットは全方位視覚センサを有しているため常に周囲の物体を認識することが可能であり，全方位移動機構により常に 2 次元平面上においてどの方向にも並進及び回転することが出来る．その他にもロボットにはキック機構が搭載されている．

また状態空間が大きい場合には相互情報量を正確に計算するためには膨大な回数の例示が必要となるため，ここでは 1 次元の状態変数に関して相互情報量を計算し，予め用意した状態空間のセットの中から相互情報量が閾値より大きい状態変数を含む状態空間を選ぶこととした．

3.2 実験結果

3.2.1 パス行為

環境中にはパサー (教示者)，レシーバ，ボール，自陣ゴール，相手ゴールそして観察者 (学習者) が存在し，様々な状況において教示者がレシーバにパスする様子を観察者は 41 回観察した．Fig. 2 に例示と主要なセンサ値の変化の一例を示す．Fig. 2(b) の赤，青，緑線はそれぞれボールとレシーバとの相対距離，パサーを中心としたボールとレシーバの相対角度，ボールとパサーとの相対距離を表している．この図から 150step 付近からパサーはレシーバに向かってドリブルを開始し，約 220step 前後でパサーはレシーバに向かってボールを蹴っていることがわかる．さらには 230step 付近の時刻でレシーバがボールをレシーブしているということも読みとることができる．

41 回の全て例示において $\hat{A}(t)$ を計算し，相互情報量を用いて $\hat{A}(t)$ が閾値よりも大きい領域と関連性の高い空間と目標状態を持った学習器を生成すると Table 1 のようになり，学習の結果獲得された行為の一例を Fig. 4(a),Fig. 4(b) に示す．またこれらの学習器の切替え方を階層型強化学習機構を用いて学習することによりパス行為を実現することができた．その様子の一例を Fig. 4(c) に示す．

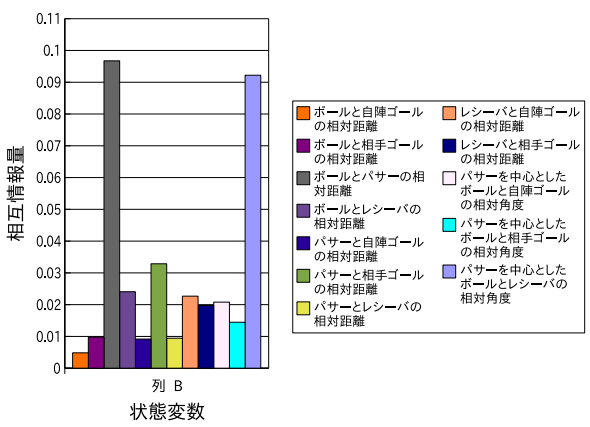


Fig.3 Mutual information in each state variable

Table 1 List of state variables and goal state in aquired learning modules

学習器	状態変数	目標状態の中心値
LM1(新規)	ボールの距離	0.05
	ボールの角度	0.00
LM2(新規)	ボールとレシーバの相対角度	0.00
	ボールの角度	0.00
	ボールの距離	任意

3.2.2 シュート行為

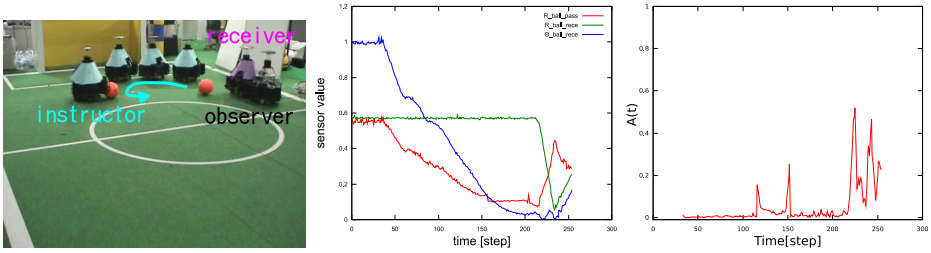
パス行為と同様 41 回の例示を行ない，相互情報量を用いて，$\hat{A}(t)$ が閾値よりも大きい領域と状態空間との関連性を計算した．その結果新たに「ボールと相手ゴールの相対角度」「ボールとの角度」「ボールとの距離」を状態変数に持つ学習器 LM3 が生成された (Table 2) 生成された学習器が学習を行なった結果獲得された行為の一例を Fig. 5(a) に示す．また獲得した学習器の切替え方を階層型学習機構により学習することにより，シュート行為が実現できた．その様子の一例を Fig. 5(b) に示す．

4 結論

パス行為およびシュート行為を観察し，模倣することで下位の行為として LM1，LM2，LM3 をまた上位の行為としてパス行為とシュート行為を累増的に獲得することができ，提案手法の有効性を検証することができた．今後はさらに「障害物を避ける」や「ボールをレシーブする」,「相手をブロックする」といった RoboCup で見られる典型的な行為が本手法により獲得可能であるか検証していく必要がある．

参考文献

[1] Andrew N. Meltzoff. Imitation and other minds: The ”like me” hypothesis. *Neuroscience to Social Science*,

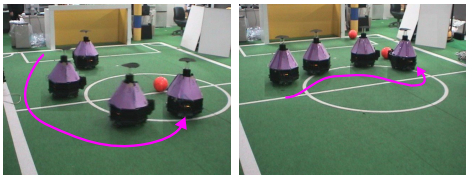


(a) (b) (c)

Fig.2 (a)A instruction, (b)the example of the major sensor values' change;((red line):the relative distance between receiver and ball, (blue line):the relative angle between receiver and ball, and (green line): the relative distance between passer and ball)and (c)Sequence of degree of attention during observation in a pass behavior of other

(a) A behavior in LM1 (b) A behavior in LM2 (c) A behavior in the upper layer

Fig.4 Examples of acquired behavior in LM1,LM2,upper layer module



(a) A behavior in LM3 (b) A behavior in the upper layer

Fig.5 Examples of acquired behavior in LM1,LM2,upper layer module

Table 2 List of state variables and goal state in aquired learning modules

学習器	状態変数	目標状態の中心値
LM1(既存)	ボールの距離	0.05
	ボールの角度	0.00
LM2(既存)	ボールとレシーバの相対角度	0.00
	ボールの角度	0.00
	ボールの距離	任意
LM3(新規)	ボールと相手ゴールの相対角度	0.00
	ボールの角度	0.00
	ボールの距離	任意

Vol. 2, pp. 55–77, 2005.

[2] 吉川雄一郎, 浅田稔. エピポーラ幾何を利用した呈示者視野復元に基づく模倣の実現. 日本ロボット学会誌, Vol. 22, No. 1, pp. 68–74, 2004.

[3] Aude Billard and Maja J. Mataric. Learning human arm movements by imitation: evaluation of a biologically inspired connectionist architecture. In *Robotics and Autonomous Systems*, Vol. 941, pp. 1–16, 2001.

[4] R. Sutton and A. Barto. *Reinforcement Learning : An Introduction*. MIT Press, 1998.

[5] Y.Takahashi and M.Asada. Multi-controller fusion in multi-layerd reinforcement learning. In *IEEE/RSJ International Conferrence on Multisensor Fusion and Integration for Intelligent Systes(MFI2001)*, pp. 7–12, 2001.

[6] Jun Morimoto and Kenji Doya. Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. In *Proceedings of International Conference on Machine Learning*, pp. 623–630, 2000.

[7] 谷口忠大, 椹木哲夫. 報酬設計による社会的相互作用に基づいた多様行為概念の自律的獲得. 第 15 回インテリジェント・システム・シンポジウム (FAN2005), pp. 365–370, 2005.

[8] Dare A. Baldwin, Jodie A. Baird, Megan M. Saylor, and M. Angela Clark. Infants parse dynamic action. *Child Development*, Vol. 72, No. 3, pp. 709–717, May/June 2001.

[9] Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, June 2005.