1A3-4

マルチエージェントタスクを考慮した二階層型強化学習

A Two Layered Reinforcement Learning for a Multi-Agent Task

*1金　天海　　*1辻野　広司　　*2中原　裕之
Chyon Hae Kim　　Hiroshi Tsujino　　Hiroyuki Nakahara

*1株式会社　ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co.,Ltd.

*2理化学研究所 脳科学総合研究センター
RIKEN Brain Science Institute

We propose a two layered reinforcement learning system that learns switching of attentive levels using the two layers in multi agent environments. This study investigates the needed learning steps while a robot learns strategies to approach an agent. We conducted two experiments using the proposed system. As results of a capture experiment, The proposed learning system, which learns switching of attentive levels, got higher success rate than conventional systems, which can not learn the switching. As results of a guidance experiment, we could confirm the same effect in a more realistic environment. We can expect the use of the proposed system for a robot to learn the way to approach a human.

1. はじめに

近年のロボット技術の進歩により，人間と近い環境において活躍するロボットに期待が寄せられている．特に移動式のロボットは自律的，または半自律的に目標のタスクを遂行し，必要に応じて人間とのインタラクションを行い，人間の行う活動を補助することを目的として研究・開発されている場合が多い．

そのようなロボットの人間との関わり方を考えるうえで特に欠かせない機能が，タスク目標や人間を含んだ周囲の状況を考えて人間との位置関係を適切に調整する機能である．本稿では人間へ接近するための方策を取り上げ，その方策を適切に選択・学習するためのシステムについて考察する．

人間が他者に接近する場合，接近の目的や他者の個性，周囲の障害物などを考慮して接近する方策を選択する．例えば人間は他者と握手をする場合には，握手に備えた距離まで接近し，握手の準備をする．通路で人の脇を通る際には他者と十分な距離をとる．後方から接近されることを嫌がる他者には迂回して正面から近づくし，広いパーソナルスペースを持つと思われる他者に対しては十分な距離をとる．人間はこの多様な接近方策をタスク遂行と他者とのかかわり合いの中で学習している．

一方で，従来のロボットは接近の目的や人間の個性に関わらず画一的な方策によって人間との位置関係を調整する場合が多い [Sisbot07]．ロボットが接近の目的や人間の個性に応じて接近方法を選択できるようにするには，その方策を学習させる必要がある．目的毎の方策獲得の学習方法としては強化学習 [Sutton00] がある．強化学習を行うロボットは報酬関数によって定義された目的に従って行動を学習するため，ロボットが人間に接近する際にも目的毎に適した接近方策が獲得できる．また，一部の強化学習システムは他者の個性にも対応している．例えば Tesauro が提案した Hyper-Q はマルチエージェント強化学習 (MARL) の枠組みに基づいてジャンケンゲームを行う [Tesauro03]．ジャンケンゲームは，個々の相手の癖を見抜いて自らの行動を選択することで平均勝率上げることができる．

連絡先: 金天海，ホンダ・リサーチ・インスティチュート・ジャパン，埼玉県和光市本町 8-1，TEL:048-462-5219，FAX:048-462-5221，E-mail:tenkai@jp.honda-ri.com

このように，強化学習システムは他者の個性にも対応できる．ただし，強化学習システムを他者への接近のようなマルチエージェント環境に用いる場合には学習速度の問題がある．通常，強化学習は目的や個性に対応した適切な動作を獲得するまでに膨大な訓練が必要となるが，この訓練には他者の参加が不可欠であり人の負担が大きい．

このようなマルチエージェント環境下において学習速度を改善するための方法の一つとして，強化学習が使用する入力情報から行動選択の判断に必要の無い情報を取り除くという手段がある．これにより強化学習器の状態空間が縮小し，学習速度は改善する．ただし，行動選択の判断に必要または必要の無い情報は容易には特定できず，ロボットの置かれた状況やタスク，遭遇する他者によっても異なる．人間はこのような情報の取捨選択をアテンションの切り替えによって実現すると考えられる．従来，ロボットのアテンションを強化学習的に切り替えるシステムが提案されている [Paletta05, Yoshikai04]．しかしながら，これらのシステムはアテンションを切り替えるために強化学習を用いてはいるが，強化学習の学習効率を向上させることが目的ではない．本稿では学習効率の向上を目的として，ロボットに他者へと接近する際の方策を学習させると同時に他者の姿勢や速度の情報に対するアテンションの深さの切り替えを学習させることができる二階層型強化学習を提案する．

2. 提案法

2.1 着眼点

ロボットが人へと接近する際の最も単純な方法は人間の位置 y に対するフィードバック制御によって動作することである．本稿ではこの y へのアテンションを第一レベルアテンションと定義する．

ロボットが人へと接近する際の次のレベルの制御としては，人間の位置と速度 $(y, \dot{y})$ から他者の動きを予測することで自らの動きを決めるということが考えられる．この場合，人間の位置 y に加え，速度 $\dot{y}$ を使用することで，他者の短期的な動きを推定できるため，より適切な接近を行うことが期待される．この短期的な情報には人のジェスチャーや習慣的な動きが含まれるため，人間の個性が表現されている可能性もある．本稿で

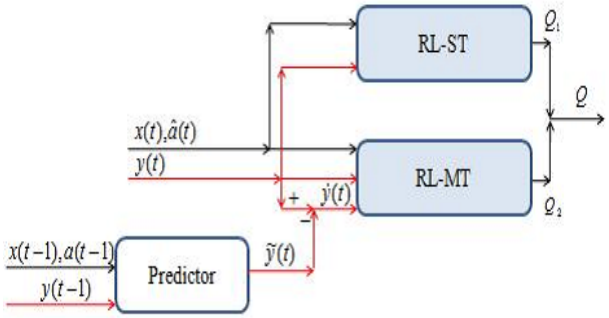

図 1: Proposed reinforcement learning system

x is the position and posture of a robot. y is the position and posture of an agent that the robot approaches. z is the position and posture of objects in the environment. This information is obtained from the sensory inputs of the robot. Here, a and $\hat{a}$ are actions that have been performed in the previous step and actions that have not yet been performed.

はこの $(y, \dot{y})$ へのアテンションを第二レベルアテンションと定義する．

第二レベルアテンションを持つ学習器の学習効率を考える場合，$\dot{y}$ の扱いが重要となる．人間の位置 y のみを参照すれば良い状況においては $\dot{y}$ の情報は参照せずとも良いノイズ情報となり，学習効率を低下させうるからである．よって，ロボットが効率良く学習するためにはこれら両方のアテンションレベルを切り替えながら学習することが望ましい．

2.2 学習システム

図 1 に提案する強化学習システムを示す．提案システムでは静的対象強化学習器 (RL-ST) と動的対象強化学習器 (RL-MT) の二つの層を用いて学習を行う．RL-ST は他者情報として第一レベルアテンション y により学習を行い，RL-MT は第二レベルアテンション $(y, \dot{y})$ により学習を行う．他者の速度 $\dot{y}$ が明示的に得られない状況では，予測器を用いることで $\dot{y}$ を抽出する．

提案手法の動作について説明する．まず，予測器がロボットの移動座標系上から観測した他者状態 $\tilde{y}$ を予測する．この他者状態 $\tilde{y}$ は他者が動かなかったとすればどのように観測されるべきかという情報である．我々は他者の絶対速度 $\dot{y}$ を実際に観測される他者の状態 y と $\tilde{y}$ との差分として定義する $(\dot{y} = y - \tilde{y})$．

RL-ST は $\dot{y}$ の情報を用いずに学習を行い，出力値 Q_1 を生成する．RL-MT は $\dot{y}$ の情報を用いて学習を行い，出力値 Q_2 を生成する．最終出力 Q はこれらの Q 値の重み付き和によって得ることとする $(Q = \lambda_1 Q_1 + \lambda_2 Q_2)$．この Q によってロボットの接近手段を選択することになる．よって，RL-ST と RL-MT はロボットの接近手段を調和的に決定しているといえる．後に 3 章においてこの調和がアテンションレベル間の遷移を引き起こすことを示す．

提案手法は有限状態によって実装した場合，次のように定式化できる．RL-ST は RL-MT と比較して次元の少ない入力空間で学習を行うため，RL-MT がマルコフ決定過程において学習している場合にも RL-ST は部分観測マルコフ決定過程において学習することになる．RL-ST への入力状態を s_i と定めると，RL-MT への入力状態は s_{ij} として定めることができる．また，RL-ST の出力を $Q_1 := O_i(s_i)$，RL-MT の出力を $Q_2 := O_{ij}(s_{ij})$ として定めることができる．この定義に従った場合のベルマン誤差は以下のようになる．

$$E = \frac{1}{2}\sum_{i,j}\sum_{i',j'} p_{ij} P_{s_{ij},s_{i'j'}} (r_{s_{ij},s_{i'j'}} + \gamma Q'(s_{i'j'},\pi) - \lambda_1 O_i - \lambda_2 O_{ij})^2 \quad (1)$$

ただし，p_{ij} は RL-MT が状態 s_{ij} を取る確率，$P_{s_{ij},s_{i'j'}}$ はロボットが選択した行動により状態 s_{ij} から状態 $s_{i'j'}$ へと遷移する確率，$r_{s_{ij},s_{i'j'}}$ はその遷移の間で与えられた報酬値，Q' はロボットの次の状態 $s_{i'j'}$ において方策 π によって定まる Q 値である．

$r_{s_{ij},s_{i'j'}} + \gamma Q'(s_{i'j'},\pi)$ を O_i と O_{ij} とは独立に求まる出力目標値として捉え勾配法を適用することで，ベルマン誤差 E より各 O の更新関数を得ることができる．

$$\Delta O_m = -\alpha \frac{\partial E}{\partial O_m} \approx \alpha\lambda_1 \sum_j \sum_{i',j'} p_{mj} P_{s_{mj},s_{i'j'}} (r_{s_{mj},s_{i'j'}} + \gamma Q'(s_{i'j'},\pi) - \lambda_1 O_m - \lambda_2 O_{mj}) \quad (2)$$

$$\Delta O_{mn} = -\alpha \frac{\partial E}{\partial O_{mn}} \approx \alpha\lambda_2 \sum_{i',j'} p_{mn} P_{s_{mn},s_{i'j'}} (r_{s_{mn},s_{i'j'}} + \gamma Q'(s_{i'j'},\pi) - \lambda_1 O_m - \lambda_2 O_{mn}) \quad (3)$$

これらをオンラインの更新関数へと置き直すことで次式が得られる．

$$\Delta Q_1 = \alpha_1(r + \gamma Q' - \lambda_1 Q_1 - \lambda_2 Q_2) \quad (4)$$

$$\Delta Q_2 = \alpha_2(r + \gamma Q' - \lambda_1 Q_1 - \lambda_2 Q_2) \quad (5)$$

実験では，メッシュ型の関数近似器と ϵ-Greedy 法を用いてシステムを構成し，$\lambda_1 = \lambda_2 = 1$ の条件のもと解析を行う．

3. 接近接触タスク

我々は提案システムの基本特性を調べるため，他者への接近接触するタスクについての実験を行い，アテンションレベルの切り替えと学習速度を検証した．

3.1 実験設定

このタスクは他者 (円) の重心をロボット (半径 R の半円) の中に捉えるタスクである (Fig.2)．

3.1.1 他者の設定

我々は他者の動きについて二種類の要素を導入した．一つ目はランダム要素であり，他者は正規乱数 $\phi(u, \sigma^2)$ により左右に移動する．二つ目は非明示的なルールの要素であり，他者は正規乱数の平均 u の符号を周期的に変更する．他者の移動は次の手順による．

1. 正規乱数のパラメータ u と σ の値を定める．
2. 正規乱数 $\phi(u, \sigma^2)$ を他者位置 y に加える．
3. u を反転させる．
4. 手順 2 と 3 を繰り返す．

u は他者の内部変数であり，ロボットからは直接観測できない．もしも u が 0 で一定であればロボットはが他者へ接触する際の最適制御方策を他者位置 y の情報だけをもとに作成できる．この場合の最適制御は他者位置 y へのフィードバック制御によって得られる．しかしながら，u が周期的に変化する場合には，単純な他者位置 y へのフィードバック制御では他者にうまく接触できない．この場合，ロボットは周期的に変化する u を推測しながら接触するための方策を見つける必要がある．

3.1.2 ロボットの設定

垂直方向へはロボットは一定速度 v で他者へと接近する．水平方向へはロボットは提案システムの出力する Q 値に従った ϵ-Greedy 法により 3 通り（右へ Δ 移動，左へ Δ 移動，移動しない）の行動をとる．ロボットの行動は u の反転と同じタイミングで切り替える．他者への接触に成功した場合，ロボットは報酬 1 を，失敗した場合-1 を得る．

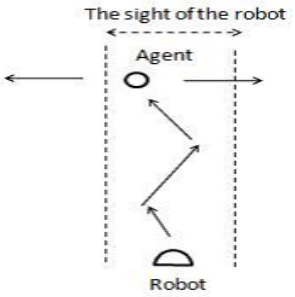

図 2: Simulated approach and contact experiment

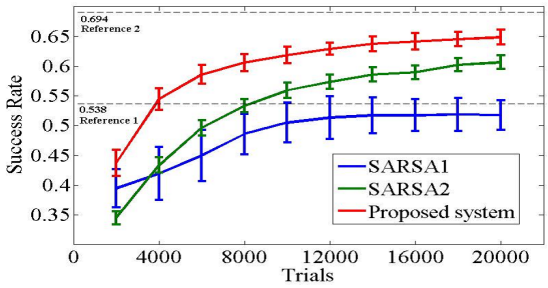

図 3: Success rate (2,000 trial moving average)

3.1.3 その他の設定

他者の初期位置はロボットの正面とした．1 セット 20000 トライアルとして $\Delta = 0.2$, $u = 0.2$, $\sigma^2 = 0.15$ 及び $R = 0.1$ のパラメータを用いて学習を行った．学習初期には RL-ST 及び RL-MT の全ての O を 0 と設定した．初期の Q 値として楽観的初期値 (optimistic initial value [Sutton00]) を設定するため，$Q = Q_1 + Q_2 + 0.007$ とした．

3.2 学習システムへの入力

入力 y としては，ロボットと他者との相対位置ベクトル $(\Delta p, \Delta q)$ を用いた．この実験では $\dot{y}$ を計算するために予測器は用いず，他者の水平絶対速度を用いた ($\dot{y}$=$\dot{p}$)．入力空間を 100×4 (垂直方向 $\times$ 水平方向) に分割した．

3.3 実験

比較のため，RL-ST と RL-MT の入出力に対応した二つの SARSA を用いた．提案システム及び二つの SARSA の学習率を 0.08 に設定した．

3.3.1 学習速度

各学習システムに対して 100 セットの学習を行い，成功率を算出した．成功率は 100 セット毎の平均をとり，さらに 2,000 トライアル移動平均として標準偏差とともにプロットした 3. グラフより，提案手法を用いると SARSA1，SARSA2 よりも高い成功率が得られることがわかる．図 3 の点線はリファレンスのための 2 種類の最適行動による成功率である．Reference 1 は現在の u の符号を知らないロボットが $y(t)$ へのフィードバックを行った場合の成功率である．Reference 2　は現在の u の符号を知っているロボットが $y(t) + u$ へ向けてフィードバックを行った場合の成功率である．

3.3.2 アテンションレベルの解析

接近接触タスクを実行中の提案システムの持つアテンションレベルを調べるため，我々は RL-ST 及び RL-MT によるロボットの行動支配に着目した．RL-ST がロボットの行動を支配している場合，ロボットは $\dot{y}$ を参照せずに行動していることになる．RL-MT がロボットの行動を支配している場合，ロボットは $(y, \dot{y})$ を参照して行動していることになる．このように，RL-ST と RL-MT の行動支配を分析することにより，ロボットのアテンションレベルが両アテンションレベルのうちのどちらにあるかを分析できる．

我々は RL-ST のアクション $a_1(\Delta p, \Delta q, \dot{p})$ を，システムが Q_2 を無視した場合のロボットの行動出力として定義する ($Q_2 = 0, Q = Q_1$)．a_1 が a と等しい場合には RL-ST がロボットの行動を支配しているといえる．この間ロボットは第一レベルアテンションに従って行動を行っている．一方で，$a \neq a_1$ の場合，RL-ST の決定したアクション a_1 は RL-MT の影響を受けて改変されていることになる．この場合 RL-ST は行動支配を行えておらず，ロボットは第二レベルアテンションに従って行動している．RL-ST の支配 (dominance) を D_1 として定義する．

$$D_1(\Delta p, \Delta q) = \int \delta_{a(\Delta p, \Delta q, \dot{p}), a_1(\Delta p, \Delta q, \dot{p})} d\dot{p} \qquad (6)$$

ただし δ はクロネッカーのデルタである．定義した支配の強さに従って接近接触タスクを学習中のロボットのアテンションレベルを描いた図が Fig.4 である．Fig.4 の両図における各ピクセルの色は，ロボットが相対位置 $(\Delta p, \Delta q)$ の入力を得た際の RL-ST の支配 D_1 を描いたものである．上方のカラーバーの青に相当するピクセルは RL-ST の支配が強い (D_1 が大きい) 領域を示しており，赤に相当するピクセルは RL-MT の支配が強い (D_1 が小さい) 領域を示している．学習初期 (Fig.4 left) には RL-ST は空間上の広い領域に渡って支配的であった．この場合ロボットは $\dot{y}$ を無視した第一レベルアテンションを用いる傾向が強い．学習後期 (Fig.4 right) には，ロボットの視界の両端と中央において RL-MT の支配が強くなった．他者がロボットより遠い場合には第一レベルアテンションを用い，他者がロボットより近い場合，または視界の端にいる場合は第二レベルアテンションを用いていることがわかる．

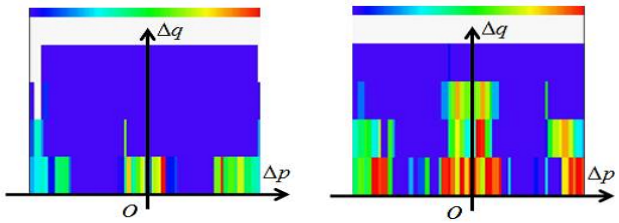

図 4: Attentive leveles during the early stage (left) and attentive levels during the last stage (right)

3.4 実験の結論

提案手法がアテンションレベルの切り替えを学習できることがわかった．また，アテンションレベルの切り替えを行わない他の手法との比較により，提案手法のほうが学習効率が良いことが分かった．

4. 接近誘導タスク

提案手法を接近誘導タスクへと適用した．接近誘導タスクは誘導ロボットが非誘導ロボットへと接近し，指示を与えることにより非誘導ロボットを所定の位置まで誘導するタスクである．ロボットのハードウェアモデルをもとにシミュレーション環境を作成した Fig.5.

4.1 誘導ロボット

誘導ロボットは頭部カメラ画像と自己姿勢の情報をもとに学習を行い，非誘導ロボットを所定の場所へと誘導する．学習シ

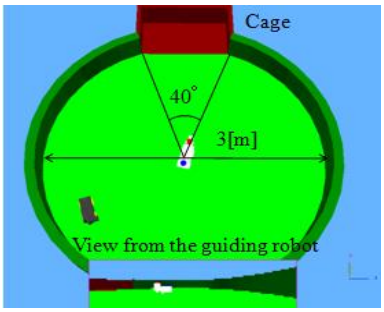

図 5: Experimental environment

表 1: State space of the learning system

Target	Information (dimension)	Range
Robot (x)	Neck yaw (1)	$[-1, 1]$
Agent (y)	Horizontal weight center (1) Rotation ($cos\theta$, $sin\theta$) (2)	$[0, 1]$ (detected) -1 (not detected)
Cage (z)	Horizontal weight center (1) Horizontal corner position (2)	$[0, 1]$ (detected) -1 (not detected)

ステムの状態空間を表 1 に示す．このタスクでは他者の動き $\dot{y}$ を捉えるための予測器を必要とする．そこで予め，誘導ロボットをランダムに動かし，動かない非誘導ロボットの見え方を予測器に学習させた．誘導ロボットは非誘導ロボットへ「向かう」,「遠ざかる」等 8 通りの行動プリミティブと，指示 (距離 0.15[m] 以内で行う) により非誘導ロボットを誘導する．行動プリミティブの選択を強化学習器が行う．誘導ロボット，非誘導ロボット，ゴールの三つとが並んだことを頭部カメラ画像上で確認できた時点 (a) で報酬 0.1 を，(a) の状態のまま直進した際に報酬 1 を，非誘導ロボットが定められたゴールへ入ったことを頭部カメラ上で確認できた時点 (c) で報酬 10 を与えた．

4.2 非誘導ロボット

非誘導ロボットは力場に従ってフィールドの中央へ戻る動作と，衝突回避の緊急動作を行うように設定した．非誘導ロボットは衝突回避用の赤外線センサが反応しない場合には，図 6(left) のようにフィールドの中央へ戻る動作をする．赤外線センサの反応がなく，誘導ロボットが半径 0.15[m] 以内に近づいた場合には，誘導ロボットからの指示を受ける．力場はその指示に従い図 6(right) のように変形する．誘導ロボットからの指示により，非誘導ロボットの移動する方向には誘導ロボットと逆方向へ移動するバイアスがかかる．このバイアスの強さによりタスクの難易度を調整できる．また，非誘導ロボットは赤外線センサに反応がある際には表 2 のように回避行動をとる．

表 2: Collision Avoidance

Activated sensors	Command
Two front sensors	Turn left or right at random
Two rear sensors	Move forward
Right sensor only	Turn left
Left sensor only	Turn right

4.3 Results

図 7 に SARSA1，SARSA2 と提案システムの成功率を示す．提案システムを用いた場合，成功率が早期に上昇した．

5. まとめ

我々はマルチエージェント環境下における静的・動的対象の区別に着目したアテンションレベルの切り替えを学習できる二層型強化学習システムを提案した．接近接触タスクの結果より，アテンションレベルの切り替えを用いた提案システムのを用いたほうが，アテンションレベルの切り替えを用いないシステムに比べ高い成功率が得られることがわかった．接近誘導タスクの結果からも，提案システムを用いたほうが高い成功率が得られることがわかった．

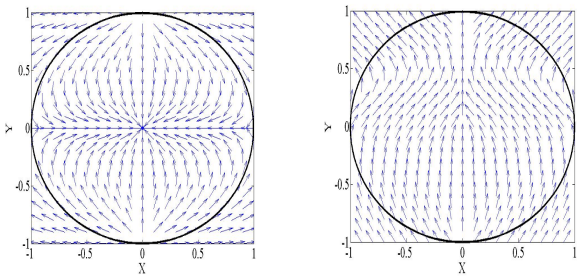

図 6: Force field (left) and force field while taking the guideance (right)

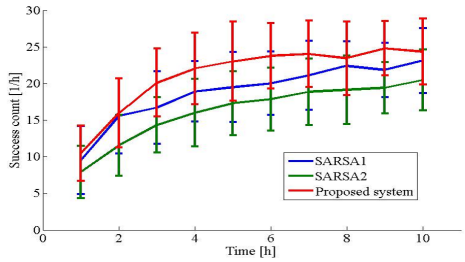

図 7: Success rate of the guiding robot simulation

参考文献

[Sisbot07] Emrah Akin Sisbot, Luis F. Marin-Urias, Rachid Alami, and Thierry Siméon: A Human Aware Mobile Robot Motion Planner, IEEE Transactions on Robotics, Vol.23, No.5, (2007).

[Sutton00] Richard S. Sutton and Andrew G. Barto: Reinforcement Learning, MIT Press, 55 Hayward Street Cambridge, MA 02142-1493 USA, (2000).

[Tesauro03] Gerald Tesauro: Extending Q-Learning to General Adaptive Multi-Agent Systems, Advances in Neural Information Processing Systems, (2003).

[Paletta05] Lucas Paletta, Gerald Fritz, and Christin Seifert: Reinforcement Learning of Informative Attention Patterns for Object Recognition, Proceedings of IEEE International Conference on Development and Learning, (2005).

[Yoshikai04] Tomoaki Yoshikai, Noritaka Otake, Ikuo Miznuchi, Masayuki Inaba, and Hirochika Inoue: Development of an Imitation Behavior in Humanoid Kenta with Reinforcement Learning Algorithm Based on the Attention during Imitation, Proceedings of IEEE/RSJ International Conference on intelligent Robots and Systems, (2004).