# 競合エージェントの存在する環境での視覚に基づく強化学習によるロボットの行動獲得

大阪大学　工学部　○　内部　英治
浅田　稔
細田　耕

## Behavior Coordination for a Mobile Robot Based on Reinforcement Learning in an Environment Including a Competitive Agent

○ Eiji UCHIBE Osaka University, 2-1, Yamadaoka, Suita, Osaka
Minoru ASADA
Koh HOSODA

**Abstract** : Reinforcement learning has been recently used to build an autonomous agent that learns to accomplish non-trivial tasks. We have applied the method for the integration of multiple behaviors of which set of states are partly inconsistent of each other, therefore the agent sometimes misunderstands world states. This paper presents the method to achieve multiple tasks in an environment including a competitive agent, and shows how the learning agent improves its performance according to the behavior of opponent agent. Simulation results are shown and the discussion is given.

***Key Words*** : reinforcement learning, multiple task, behavior coordination, competitive agent, soccer game

## 1　はじめに

環境とのインタラクションを繰り返し，目的のタスクを達成するための行動を獲得するための手法として，強化学習が注目されている．強化学習では，エージェントがある行動を起こした時に，報酬を与えることで目的行動を獲得することが可能である．

これまで，複数のタスク達成のための行動を強化学習を用いて獲得した研究にSingh[6]やWhitehead[9]がある．例えばSingh はサブタスク間で状態空間と行動空間を共有できる場合に，逐次的な複数タスクを達成するための行動獲得の方法を提案している．しかし，サブタスク間で必ずしも状態空間が共有できる訳ではない．また，エージェントは鳥瞰図的なセンサを想定しており，自律エージェントへの適用に関して考慮されていない．

また，競合エージェントが存在する環境で強化学習を取り扱った研究として，Littman の研究[5]がある．Littman は1対1のサッカーゲームについてシミュレーションを行なっているが，これも鳥瞰図的なセンサを用いて問題を簡単化している．

一方，Asada らはエージェントに搭載されたカメラからの情報をもとに強化学習を適用し，ボールをゴールにシュートするという，簡単なサッカーのゲームを実現している[1]．本稿では，[1]で取り扱ったタスクをより複雑にしたタスクを想定する．具体的には，キーパーエージェントを導入し，シューティングの行動と，キーパーエージェントとの衝突の回避行動の二つの行動を協調させる手法について提案する．

今回想定する1対1の簡単なサッカーゲームでは，Littman や Singh らのタスクにはない，状態空間が干渉するといった知覚的見せかけ問題[3,4]が発生する．この問題に対処するため，ここでは統計的な解析を用いて，その干渉状態を自律的に発見する．また，学習エージェントの学習に，キーパーエージェントのビヘービアが及ぼす影響について述べる．最後に，シミュレーションによる結果を示し，今後の方針について述べる．

## 2　$Q$ 学習による複数行動の統合

$Q$学習はWatkins によって提案された確率的動的計画法を利用した，強化学習の一つである．エージェントを含む環境全体がマルコフ性を満足する場合には，$Q$学習は状態$s_0 = i$から始まる場合の減衰した積算報酬の条件つき期待値

$$\lim_{N\to\infty} \mathrm{E}\left[\sum_{t=0}^{N-1} \gamma^t r_{s_t} \middle| s_0 = i\right]$$

を最大とするような政策を獲得できることが証明されている[8]．

### 2.1　$Q$学習のアルゴリズム

最も基本的な1ステップ$Q$学習のアルゴリズムを以下に示す．

1. 状態$s(\in \boldsymbol{S})$の時，行動$a(\in \boldsymbol{A})$をとる時の行動価値関数$Q(s,a)$をある値(通常は0)で初期化する．
2. 現在の状態$s$を観測する．
3. エージェントが実行する行動$a$を選択する．

4. 行動 $a$ を実行し，環境から報酬 $r$ を受けとる．環境は $s'$ に遷移する．

5. $Q(s,a)$ の更新は

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha(r+\gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \tag{1}$$

で行う．

6. 行動の方策 $f$ の更新は

$$f(s) \leftarrow a \text{ such that } Q(s,a) = \max_{a' \in \boldsymbol{A}} Q(s',a') \tag{2}$$

7. 2 に戻る

ここで，$\alpha$ は学習率 $(0 < \alpha < 1)$，減衰係数 $\gamma$ は $0 < \gamma < 1$ である．学習後は状態が $s$ のとき $a = \arg\max_b Q(s,b)$ である行動を選択することで最適行動を実現できるが，学習中は未探索領域の探索と探索によって獲得した知識の利用という2つの矛盾する要求を満足しなければならないが，学習中の行動戦略としては，しばしばボルツマン分布に基づく確率的手法が用いられる．つまり，状態 $s$ において行動 $a$ を選択する確率 $P(a|x)$ は

$$P(a|x) = \frac{\exp(Q(s,a)/T)}{\sum_{b \in \boldsymbol{A}} \exp(Q(s,b)/T)} \tag{3}$$

によって与えられる．ここで $T$ は温度パラメータであり，$T$ が大きいほど行動戦略はランダムになり，$T$ を 0 に近づけると保守的になる．

### 2.2 $Q$ 学習の反射的なタスクへの適用

学習パラメータや更新式の変更によって，反射的な行動を必要とするタスクを学習する場合にも，$Q$ 学習は適用できる．例えば衝突回避の場合，$\gamma$ を低くし，衝突したときに負の報酬を与える．学習中の行動価値関数の更新式は max のかわりに min を用いた

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha(r+\gamma \min_{a' \in \boldsymbol{A}} Q(s',a')) \tag{4}$$

を使用する．これにより学習中は，衝突することを学習する．学習後は通常の $Q$ 学習の政策 ($Q$ 値を最大にする行動の選択) をとることにより，目標状態の近傍で衝突行動をとらないようになる．

### 2.3 サブタスクの学習結果の統合

目標指向的タスク $(\gamma \approx 1)$ と反射的タスク $(0 < \gamma \ll 1)$ の学習結果を統合することにより，複雑なタスクに対処できる．ここでは学習結果の統合法として，以下の三つの方法を提案する．

#### 2.3.1 シンプルサムによる統合

二つの行動価値関数 ${}^sQ$，${}^aQ$ を単純に加えることによって，新しい行動価値関数 ${}^cQ_{ss}$ を構成する．すなわち，

$${}^cQ_{ss}({}^cs,a) = {}^sQ(({}^ss,*),a) + {}^aQ((*,{}^as),a) \tag{5}$$

ここで $({}^ss,*)$，$(*,{}^as)$ は統合後の状態を前の状態 ${}^ss$，${}^as$ で表現するのに用いており，$*$ は任意の状態を表す．つまり，${}^sQ$ の計算には ${}^ss$，${}^aQ$ の計算には ${}^as$ だけを使用する．

#### 2.3.2 スイッチングによる統合

状況に応じて行動の政策つまり行動価値関数を使い分ける．新しい行動価値関数 ${}^cQ_{sw}$ は

$${}^cQ_{sw}({}^cs,a) = \begin{cases} {}^sQ({}^ss,a) & (\text{ ある条件 }) \\ {}^aQ({}^as,a) & (\text{ それ以外 }) \end{cases} \tag{6}$$

で与えられる．

#### 2.3.3 再学習による統合

上記2つの統合法では，サブタスク間の状態空間が非干渉である事を暗黙のうちに仮定している．そのため，サブタスク間で状態空間が干渉する場合には，異なる状態を同一の状態とみなしてしまう知覚的見せかけ問題が発生し，学習に悪影響を及ぼす．

そこで，再学習による統合では，まず，サブタスクの学習によって得られた行動価値関数を先験的知識として与える ($Q$ の初期値を与える) ことで，学習時間の短縮を実現し，また統合前後の状態遷移確率の分布を比較することで，タスク達成時に問題となる干渉状態を自律的に検出する．具体的には，$\chi^2$ 検定を用いる．

行動価値関数 ${}^cQ_{rl}$ の初期化は，シンプルサムによる方法で与える．

## 3 タスク

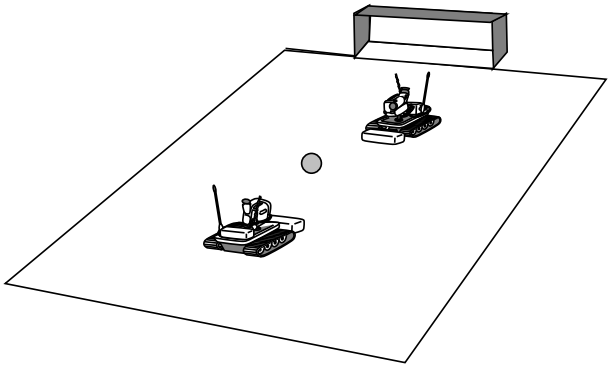


**Fig.1** The task of shooting a ball into the goal with a keeper robot

タスクとして，学習エージェントがキーパーエージェントとの衝突をできるだけ回避しながら，ボールをゴールにシュートするタスクについて考える．エージェントに搭載されているのはカメラだけであり，自身の幾何学的パラメータや動的特性などに関する知識はもっていない．

また，エージェントは左右の車輪を独立に動かすことのできる PWS(Power Wheeled Steering) システムにより，駆動される．また，状態空間の構成は一般に困難な問題であるが，ここでは環境内にはボール，キーパーエージェント，ゴール，ラインだけが存在すると仮定し，それぞれ，画像上で位置や大きさなどについて **Fig.2**の様に量子化し，その組合せで構成する．

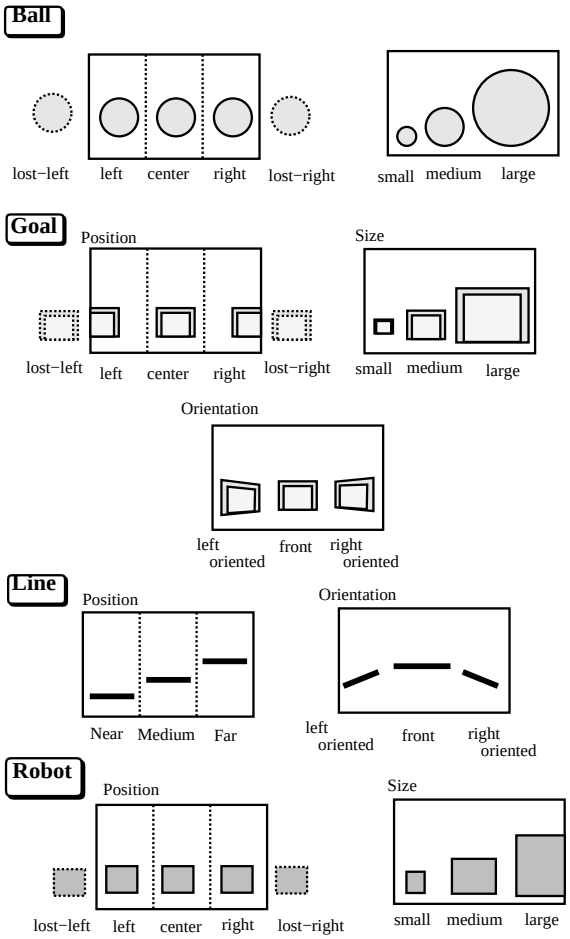



**Fig.2** The substates of state space

## 4 シミュレーション

### 4.1 設定

以上のことをシミュレーションで検証した．ここで，スイッチングにおける行動価値観数の切替えの条件については，

- switching1： ボールとゴールだけが見えている場合には，シューティング，それ以外は衝突回避行動をとる．
- switching2： キーパーだけが見えている場合に，衝突回避，それ以外はシューティング行動をとる．

という二つの条件を準備した．また，シミュレーションにおけるキーパーエージェントの行動戦略は，まず低速で防御行動をとらせた．防御行動は，学習エージェントを観測し，その進行方向を妨害することで実現した．

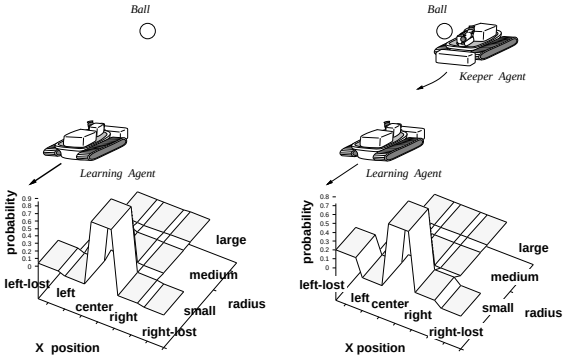
### 4.2 結果および考察

まず，複数行動の統合法の違いによる結果を **Table 1** に示す．シンプルサムによる方法では，しばしば二つの行動価値がつりあい，停留点に陥ってしまった．またスイッチングによる方法では，シュート率か衝突までの平均ステップ数どちらかを優先させることしかできなかった．一方，再学習法がシュート率，衝突までの平均ステップ数の両方に対して良好な結果を獲得された．

**Table 1** Simulation results

| | shooting(%) | steps/collisions |
|---|---|---|
| simple sum | 36.1 | 172.3 |
| switching1 | 39.5 | 6207.4 |
| switching2 | 49.6 | 95.2 |
| learning | 60.8 | 5048.5 |

次に，再学習時における，$\chi^2$ 検定によって推定された干渉状態について述べる．まず，**Fig.3**は，ボールが左に小さく見えている場合に学習ロボットが後退の行動をとったときの，状態遷移確率を示している．キーパーエージェントは左に小さく見えているとする (Fig.3 参照)．ここで X position は観測された画像上におけるボール中心の $x$ 座標，radius は半径である．シューティングのみの場合と比較して，統合時ではボールが左に消えた確率が高くなっている．これはキーパーエージェントにボールが隠されたためである．

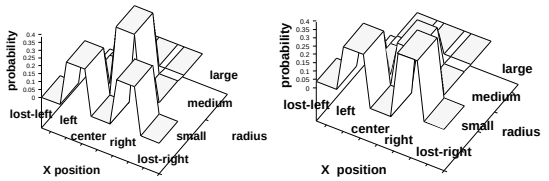



(a) shooting (b) integration

**Fig.3** The example of inconsistent states and probability distribution(case:occlusion)

最終的に矛盾が生ずると判断された状態は，全部で 22 状態となった．この中には，オクルージョンが原因のものだけではなく，キーパーエージェントがボールを動かしてしまうことが原因であると考えられる状態も検出された．たとえば，**Fig.4** はボールが小さく見えていて，前進の行動をとった場合の状態遷移確率を示しているが，シューティング (Fig.4(a)) の場合には中央に中ぐらいに見える確率が高くなっているのに対し，統合 (Fig.4(b)) の場合には右に小さく見えた確率が最も高くなっている．左に消えた確率が 0 であることから，オクルージョンは発生していない．これは，前進中にキーパーエージェントと衝突したり，キーパーエージェントがボールを運んだりしたことが原因である．

次に，獲得されたシューティングの様子を**Fig.5**に示す．黒いエージェントが学習エージェントで，エー




(a) shooting (b) integration

**Fig.4** Probability distribution (case:pushing,etc.)

ジェントから出ている2本の線は視野を表している．一度ボールを見失っているが，結果的にシュートすることができた．

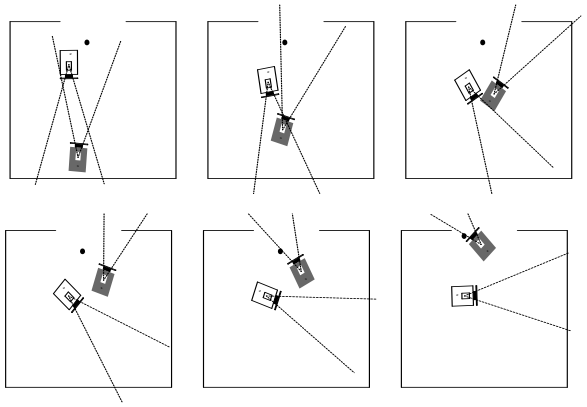


**Fig.5** Shooting a ball into the goal without collisions

最後に，キーパーエージェントの戦略が学習ロボットの戦略に与える影響について述べる．**Fig.6**は，再学習時における LEM(Learning from Easy Mission)[1)] を用いた場合と，そうでない場合のシュート率の推移の相違を示している．最初の矢印までは，キーパーエージェントは静止していて，そこから2番目の矢印までは低速で防御行動をとっている．それ以降は，高速で防御行動をとっている．

キーパーエージェントの戦略をスケジューリングしない場合は，学習により多くの時間を費し，また，シュート率もそれほど改善されにくいことが判る．これは，はじめからキーパーエージェントが最適な行動を実現していた場合，学習エージェントがタスクを達成できず，強化信号を受け取ること頻度が低くなるためである．

## 5 おわりに

「ボールをゴールにシュートする」サブタスクと，「キーパーエージェントとの衝突を回避する」という状態空間が干渉するような，2つのサブタスクを統合する手法を提案し，状態空間の干渉する状態を自律的に検出する手法について提案した．また，学習エージェント以外のエージェントが学習に及ぼす影響について考察した．

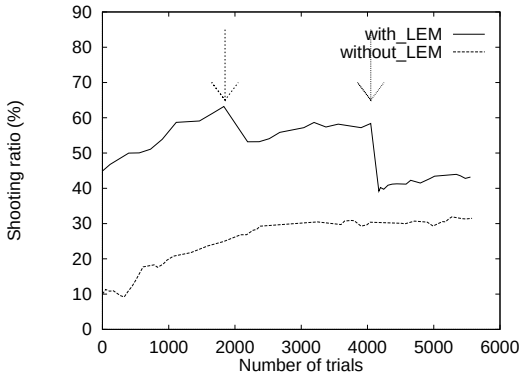



**Fig.6** Curves of shooting rate while *re-learning* (with/without) LEM according to the behavior of keeper agent

今後の方針として，現在のところ，キーパーエージェントを実機で動かせていないため[2)]，それを含めた実験を行なう必要がある．また，タスクのバリエーションとして，キーパーエージェントを含めた同時学習や，エージェントの台数を更に増やし，エージェント間の協調といった問題を取り扱った時の，状態空間の構成や戦略の検討，行動理解[7)] について考える必要がある．

## 参考文献

[1] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-Based Reinforcement Learning for Purposive Behavior Acquisition. In *Proc. of IEEE International Conference on Robotics and Automation*, pp. 146–153, 1995.

[2] M. Asada, E. Uchibe, S. Noda, S. Tawaratsumida, and K. Hosoda. Coordination Of Multiple Behaviors Acquired By A Vision-Based Reinforcement Learning. In *Proc. of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, pp. 917–924, 1994.

[3] L. Chrisman. Reinforcement Learning with Perceptual Aliasing: The Predictive Distinctions Approach. In *Proc. of the 10th International Conference on Artificial Intelligence*, pp. 183–188, San Jose, CA, 1992. AAAI Press.

[4] 開, 松原. 機械学習から見たロボット学習 — 能動的学習機構に向 けて—. 日本ロボット学会誌, 13(1):5–10, 1995.

[5] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. of the 11th International Conference on Machine Learning*, pp. 157–163, 1994.

[6] S. P. Singh. Transfer of Learning by Composing Solution of Elemental Sequential Tasks. In *Machine Learning*, Vol. 8, pp. 99–115, 1992.

[7] 内部, 浅田, 細田. 他のエージェントの行動理解 —サッカーロボットにおける強化学習のマルチエージェント環境への適用に向けて. 第12回日本ロボット学会学術講演会予稿集, pp. 241–242, 1995.

[8] C. J. C. H. Watkins and P. Dayan. Technical note: *Q*-learning. *Machine Learning*, pp. 279–292, 1992.

[9] S. D. Whitehead, J. Karlsson, and J. Tenenberg. Learning Multiple Goal Behavior Via Task Decomposition And Dynamic Policy Merging. In J. H. Connel and S. Mahadevan eds., *Robot Learning*, chapter 3. Kluwer Academic Publishers, 1993.