視覚を有する移動ロボットの強化学習による複数タスクの達成

Achievement of Multiple Tasks for a Mobile Robot with a Visual Sensor Using Delayed Reinforcement Learning

○ 准　内部　英治　(阪大)　　正　浅田　稔　(阪大)
准　野田　彰一　(阪大)　　准　細田　耕　(阪大)

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

Abstract : Reinforcement learning(RL) has been recently used to build an autonomous robot that learn to accomplish non-trivial tasks. We have applied the vision-based RL method for the integration of a set of tasks of which states are not completely independent of each other, therefore the robot sometimes misunderstands world states. This is called perceptual aliasing problem. This paper presents the method for the robot with a limited sensor to achieve multiple tasks. The virtual sensor is implemented to cope with perceptual aliasing, and the state transtion probability distribution allows the robot to detect the states which are inconsistent between subtasks. Simulation results are shown and the discussion is given.
Key Words : reinforcement learning, perceptual aliasing, multiple task,soccer robot

1　はじめに

近年，環境の明確なモデルを作ることなく，反射的かつ適応的に行動を獲得し，タスクを達成する手法として，強化学習法が注目されている．強化学習では環境の状態数の増加に応じて学習時間が指数関数的に増大するといった問題があるが，ロボットがある行動を起こした時に報酬を与えるだけで目的の行動が得られるという長所をもつ．

強化学習に関する従来の研究では，複数タスクへの拡張[5,10]や，知覚的見せかけ問題[3]と呼ばれる，隠れ状態が存在するような環境下での学習[2,9]などの理論的な考察をしているものが多い．前者の例として，Singhはサブタスク間で状態空間と行動空間を共有することのできる場合について，逐次的な多重タスクを学習する手法について提案している[5]．しかし，サブタスク間で状態空間が干渉しない理想的な環境を対象としており，簡単なシミュレーションによる結果しか示しておらず，実ロボットへの適用可能性について論じているものは少ない．また，知覚的見せかけ問題に対する研究としては，隠れマルコフモデルを用いる手法[2]，リカレントネットを用いる手法[4]などがあげられる．これらもまた，問題を簡単化するために，格子上の世界で学習を行っている．

一方，実ロボットに強化学習を適用した研究では．Connel and Mahadevan が，箱押し作業を事前に「箱の発見」,「箱押し」,「スタック状態からの回避」の3つに分解しておくことによって実現している．この研究に代表される従来の研究はバンパーセンサ，超音波センサなど近接センサのみを使用しているので，局所的なタスクの達成のためには有効であるが，大局的な目的行動を獲得することには向いていない．

これに対し浅田らは，視覚に基づく強化学習をサッカーロボットに適用している[1]．彼らは，搭載されたカメラからの画像だけから，ボールをゴールにシュートするタスクを達成している．

そこで本稿では，[1]で想定しているタスクをより複雑にした，サブタスク間で状態空間が干渉する場合の多重タスクを達成するための手法について提案する．また，その干渉によって発生する知覚的見せかけ問題に対して，その状態を自律的に検出する手法について提案する．簡単な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学習による多重行動の統合

Q学習は強化学習法の一つであり，Watkins によって提案された，確率的動的計画法に基づく学習アルゴリズムである．ロボッ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]．ここで，γは減衰係数である．

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. ロボットが実行する行動 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
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha(r+\gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \quad (1)$$
で行う．

6. 行動の方策 f の更新は
$$f(s) \leftarrow a \text{ such that } Q(s,a) = \max_{a' \in \boldsymbol{A}} Q(s',a') \quad (2)$$

7. 2 に戻る

ここで，α は学習率 $(0 < \alpha < 1)$，減衰係数 γ 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が，学習中は未探索領域の探索と探索した領域の利用という2つの矛盾する要求があるが，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
$$P(a|x) = \frac{\exp(Q(s,a)/T)}{\sum_{b \in \boldsymbol{A}} \exp(Q(s,b)/T)} \quad (3)$$
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，γ を低くし，衝突したときに負の報酬を与えるとする．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
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha(r+\gamma \min_{a' \in \boldsymbol{A}} Q(s',a')) \quad (4)$$
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，
$${}^cQ_{ss}({}^cs,a) = {}^sQ(({}^ss,*),a) + {}^aQ((*,{}^as),a) \quad (5)$$
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
$${}^cQ_{sw}({}^cs,a) = \begin{cases} {}^sQ({}^ss,a) & (\text{ ある条件 }) \\ {}^aQ({}^as,a) & (\text{ それ以外 }) \end{cases} \quad (6)$$
で与えられる．

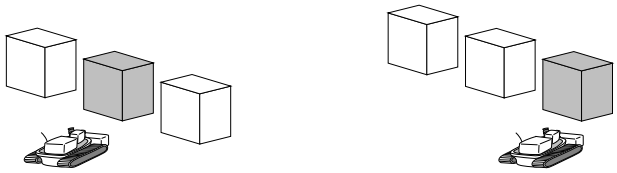
2.3.3 再学習による統合

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 (Perceptual Aliasing Problem)[9] が発生する．そこで，再学習による統合では，サブタスク間で，干渉する部分の状態 ${}^cs_{sub}$ を人間が追加し，その部分を重点的に再学習させ最終的なタスクを達成させる．また，サブタスクの学習結果を先験的知識として与える (Q の初期値を与える) ことで，学習時間を短縮できる．

行動価値関数 ${}^cQ_{rl}$ の初期化は，
$$\begin{aligned} {}^cQ_{rl}({}^cs,a) &= {}^cQ_{ss}({}^cs,a) \\ {}^cQ_{rl}({}^cs_{sub},a) &= {}^cs_{sub}\text{に最も近い状態の } {}^cQ_{ss}({}^cs,a) \end{aligned} \quad (7)$$
によって行なう．また ${}^cQ_{rl}$ の更新は (1) 式で行なう．

3 センサの拡張

自律ロボットは，それ自身に搭載されたセンサだけから環境の状態を認識しなければならない．しかし，センサの能力には限界があり，現在の環境の状態を完全に認識することはできず，前節で示したような知覚的見せかけ問題が発生する．たとえば，ロボットは自



(a) (b)

Fig.1 Perceptual aliasing

身に搭載されたカメラだけから環境の状態を認識しなくてはならない場合を考える．このとき，**Fig.1(a)** と **Fig.1(b)** では異なる状況にロボットは置かれているのに，その違いをロボットは認識することはできない．強化学習を適用する場合，知覚的見せかけ問題が発生していると，学習が困難になってしまう．

そこで，ここでは環境に対して相対的に運動する物体それぞれに対して，状態遷移確率を導入し，それを用いて仮想的にセンサ能力を拡張する．つまり，実際のセンサを用いて対象物に関するモデルを獲得しながら，そのモデルを用いて学習をおこなう．強化学習にモデルを用いる点からすると，このアプローチは Sutton の Dyna アーキテクチャと類似しているが[6]，Sutton の Dyna アーキテクチャは環境のモデルを獲得し，実世界で学習を行うコストの削減を狙ったものである．それに対し，提案する手法では，空間の連続性を仮定することにより，実センサの両側に仮想センサを設けて，センサの能力を拡張することが目的である．

状態遷移確率 $P_{ij}(a)$ の推定は，
$$P^k_{ij}(a) = \frac{n^a_{ij}(t)}{\sum_{j \in \boldsymbol{S}} n^a_{ij}(t)}, \quad (8)$$
を用いる．ここで，$n^a_{ij}(t)$ は時刻 t において，状態 i で行動 a をとったときに状態 j に遷移した回数を示す．実

センサを通して，状態遷移確率 $P_{ij}(a)$ を推定し，それを用いて仮想センサのモデルとする．また，状態遷移確率を用いて 2.3 で想定した「統合時の干渉する状態」を発見できる．これは再学習時に，統合前の状態遷移確率と統合後の状態遷移確率を χ^2 検定を用いて検定することにより，可能となる．

4 タスク

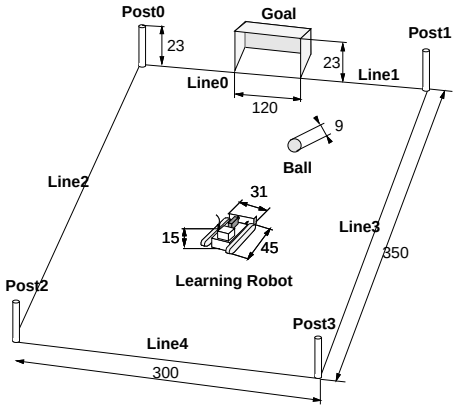


Fig.2 The task of shooting a ball into the goal with a keeper robot

タスクとして，ロボットがキーパーロボッ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．ロボッ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，ロボッ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，行動する．また，環境内には以下のものが存在し，これらの組合せで状態空間が構成される．

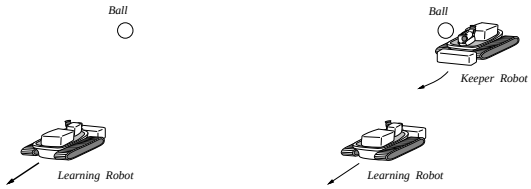
- ボール → 位置と直径をそれぞれ3通り，
- キーパーロボット → 位置と高さをそれぞれ3通り，
- ゴール → 位置，高さ，傾きをそれぞれ3通り，
- コーナーポスト → 位置と高さをそれぞれ3通り，
- ライン → 位置と傾きそれぞれ3通り．

ゴール，コーナーポスト，フィールドラインに関しては，簡単のため優先順位を設定し，構成される状態数を削減した．詳細については [7] を参照されたい．

5 シミュレーション結果および考察

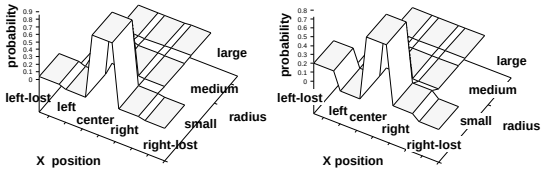
Fig.4 は，ボールが左に小さく見えている場合に学習ロボットが後退の行動をとったときの，状態遷移確率である．キーパーロボットは左に小さく見えているとする (**Fig.3** 参照)．ここで X position はボールの画像上での中心の x 座標，radius は半径である．シューティング時と比較して，統合時はボールが左に消えた確率が高くなっている．これはキーパーロボットにボールが隠されたためである．

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



(a) shooting (b) integration

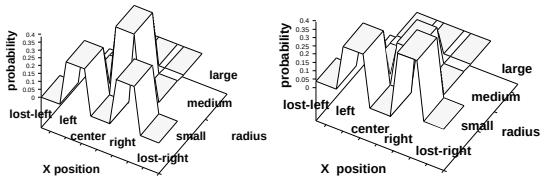
Fig.3 The example of inconsistent states



(a) shooting (b) integration

Fig.4 Probability distribution (case:occlusion)

のものだけではなく，キーパーロボッ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.5** 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，shooting 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，integration 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，前進中にキーパーロボットと衝突したり，キーパーロボットがボールを運んだりしたことが原因である．



(a) shooting (b) integration

Fig.5 Probability distribution (case:pushing,etc.

ロボット自身が干渉状態の意味を知る必要はないが，干渉状態として検出される状態遷移確率にバラツキがある場合には問題となる．たとえば，現在キーパーロ

ボットの速度に関する状態は組み込んでいないため，キーパーロボットが複数の戦略を持っていた場合，得られる状態遷移確率はばらついてしまい，知覚的見せかけ問題が発生してしまう．しかし、このタイプの知覚的見せかけ問題に対しては，提案した手法では対処することはできない．

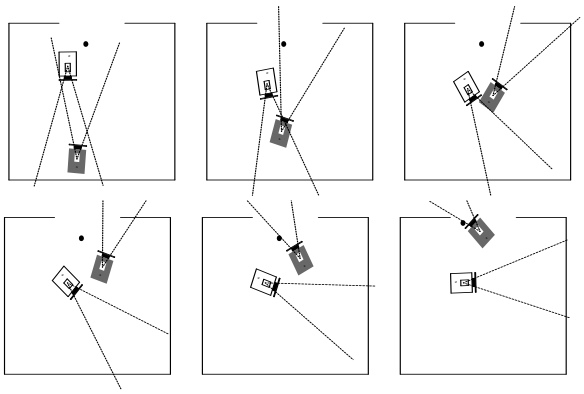


Fig.6 A shooting behavior

次に，獲得されたシューティングの様子を **Fig.6** に示す．黒いロボットが学習ロボットで，ロボットから出ている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．**Table 1** は統合法の違いによる結果の比較である．再学習法がシュート率，衝突までの平均ステップ数ともに優れていることがわかる．

Table 1 Simulation results (integration methods)

	shooting(%)	steps/collisions
simple sum	36.1	154.3
switching	52.6	50.7
learning	64.2	7215.2

Table 2 shooting rate (with/without virtual sensor)

	visible	unvisible
without virtual	63.0 %	10.4 %
with virtual	64.2 %	37.4 %

また，仮想視野の有無によるシュート率の違いを **Table 2** に示す．初期位置として，ゴールまでの距離が遠くキーパーロボットがゴール前にいるような位置から始めている．ボールが常に見えていた場合(visible)は，シュート率にそれほど違いはない．しかし，ボールを一度見失わなければならない場合(unvisible)，仮想視野を用いた場合の方がシュート率は良くなっている．仮想視野を用いない場合でもシュートができるのは，最適な行動が画像上でのボールの半径と関係ない(左に消えたら，回転してすぐに見えるように行動すれば良い)状況があるからである．状態数が増加しているため，単純に比較はできないが，仮想視野を用いない場合には，知覚的見せかけ問題に対処できない．仮想視野はある程度，この問題を解決していると考えられる．

6 おわりに

「ボールをゴールにシュートする」サブタスクと，「キーパーロボッ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た．また，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．

今後の課題としては，獲得した Q 値や，状態遷移確率をより有効に用いて抽象化，学習中の行動選択に関する研究が考えられる．また，キーパーロボットの速度も考慮して，運動を予測しながらの衝突回避といったことがあげられる．

参考文献

[1] 浅田, 野田, 俵積田, 細田. 視覚に基づく強化学習によるロボットの行動獲得. 日本ロボット学会誌, 13(1):68–74, 1995.

[2] 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.

[3] 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.

[4] L.-J. Lin and T. M. Mitchell. Reinforcement Learning With Hidden States. In *Proc. of the 2nd International Conference on Simulation of Adaptive Behavior: From Animals to Animats 2.*, pp. 271–280, 1992.

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

[6] R. S. Sutton. Integrated Architecture for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Proc. of the 7th International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, 1990.

[7] 内部, 浅田, 野田, 細田. 視覚に基づく強化学習による移動ロボットの多重タスク遂行のための協調行動の獲得. 第21回 人工知能基礎論研究会(SIG-FAI-9403), pp. 25–32. 人工知能学会, 1995.

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

[9] S. D. Whitehead and D. H. Ballard. Active Perception and Reinforcement Learning. In *Proc. of the 7th International Conference on Machine Learning*, pp. 179–188. Morgan Kaufmann, 1990.

[10] 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.