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の達成

◯内部 英治　　浅田 稔　　野田 彰一　　細田 耕
大阪大学工学部

The Coordination Of Multiple Behaviors For A Mobile Robot Acquired by Vision-Based Reinforcement Learning

◯Eiji UCHIBE　　Minoru ASADA　　Shoichi NODA　　Koh HOSODA
Osaka University

1　はじめに

環境の明確なモデルを作ることなく，タスクを達成する手法として，強化学習法[1]があるが，学習時間が状態数に対して指数関数的に増大する問題がある．学習時間を短縮させるため，タスクを時系列のサブタスクに分解し[2]，個々のサブタスクを独立に学習させて，その結果を統合する[3]研究がある．これらの研究では，想定しているタスクは，状態空間が干渉しないサブタスクに分解可能であり，また同時に達成する必要がない．しかし，全てのタスクが，独立で非干渉のサブタスクに分解できるとは限らない．

本報告では，強化学習の一つであるQ学習を用いて，このようなタスクを一旦独立なサブタスクに分解し，それぞれ学習させた後に，統合する法を提案する．それらは個々の学習結果の単純和(シンプルサム)，学習結果の状況による使い分け(スイッチング)，先験的知識としての学習の結果の使用(再学習)の三種類である．再学習ではサブタスク間の状態空間の干渉によって生じる新たな状態を追加し，その状態を重点的に学習させる．サッカーロボットがキーパーを回避しながらシュートするタスクのシミュレーションを行ない，この学習法の有効性を検証する．

2　Q学習による多重行動の統合

2.1　Q学習のアルゴリズム

以下に1ステップQ学習のアルゴリズムを示す．$\boldsymbol{S}$を環境の有限な離散状態集合，$\boldsymbol{A}$をロボットが実現することのできる有限な離散行動集合とする．

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. $$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \quad (1)$$
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$(0 < \alpha < 1)$，γは減衰率$(0 < \gamma < 1)$である．

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，負の報酬とする．学習中の行動価値関数の更新式として
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \min_{a' \in \boldsymbol{A}} Q(s',a')) \quad (3)$$
を使用する．これにより学習中は，衝突することを学習し，学習後は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${}^gQ$, ${}^rQ$を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，
$${}^cQ_{ss}({}^cs,a) = {}^gQ(({}^gs,*),a) + {}^rQ((*,{}^rs),a) \quad (4)$$
ここで$({}^gs,*)$，$(*,{}^rs)$は統合後の状態を前の状態${}^gs$，${}^rs$で表現するのに用いており，$*$は任意の状態を表す．つまり，${}^gQ$の計算には${}^gs$，${}^rQ$の計算には${}^rs$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
$${}^cQ_{sw}({}^cs,a) = \begin{cases} {}^rQ({}^rs,a) & (\text{ ある条件 }) \\ {}^gQ({}^gs,a) & (\text{ それ以外 }) \end{cases} \quad (5)$$
で与えられる．

2.3.3　再学習による統合

再学習による統合では，サブタスク間の状態空間の干渉を考慮して，新たに状態を追加し，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．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,

$$\begin{aligned} {}^cQ_{rl}({}^cs,a) &= {}^cQ_{ss}({}^cs,a) \\ {}^cQ_{rl}({}^cs_{sub},a) &= \text{分割前の対応する状態の} \\ & \quad {}^cQ_{ss}({}^cs,a) \end{aligned} \tag{6}$$

によって行なう．また ${}^cQ_{rl}$ の更新は (1) 式で行なう．

3 仮定とタスクの分解

3.1 タスク

タスクは，キーパー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 1 対 1 の簡単なサッカーの試合を想定する．

ロボットは，左右に動輪をもち，一つのカメラを中心部に搭載している．また，ロボットは自分自身やキーパーの運動学や動力学を知らないと仮定する．このタスクを以下の二つのサブタスクに分解する．

- ${}^sT$: キーパーのいない環境でボールをゴールにシュートする (shooting)[1)]
- ${}^aT$: キーパーとの衝突を回避する (avoidance)

3.2 サブタスクの学習結果の統合

シンプルサムでは二つの行動価値関数 ${}^gQ$ と ${}^aQ$ の単純和を使用する．スイッチング時の切替え条件は,「キーパーのみが見えた」場合とする．

再学習の場合は,「ボールがキーパーに隠された」という状態を新たに追加する．再学習の初期段階ではキーパーを静止させた環境で学習させ，その後，キーパーが移動する環境に変更して学習させる．

4 シミュレーション

4.1 シミュレーション環境

環境内に存在するのはボール，学習ロボット，キーパー，及びゴールだけと仮定する． また α は

Table 1 Parameters

	# of states	γ
${}^sT$	319	0.9
${}^aT$	11	0.1
learning	3770	0.9

全て 0.25 で一定，行動集合の大きさ $|\boldsymbol{A}|$ は左右輪の状態 (前進，停止，後退) の組合せから両輪停止を除いた 8 とする．

4.2 シミュレーション結果

キーパーがシュートコースを遮る行動をしたときのシミュレーション結果を **Table 2** に示す. shooting はシュート率 (ゴール数/試行回数) であり，steps/collisions はキーパーが見えていて衝突するまでの平均のステップ数である．

シンプルサムでは，学習ロボットとキーパーとボールが一直線上に並んだ時，停留点に陥るという問題があり，シュート率が一番低い．また，スイッチングではボール，ゴールとキーパーが同時に見えたとき，キーパーを無視してしまうため，衝突数が多くなっている．再学習による方法は，

Table 2 Simulation results

	shooting(%)	steps/collisions
simple sum	36.1	154.3
switching	52.6	50.7
learning	63.5	7215.2

シュート率，衝突数ともに最善であり，試行の途中で学習ロボットが停留する (局所解に陥る) ことも，他の方法と比較して少なかった． **Fig.1** に再学習した後の行動の様子を示す.

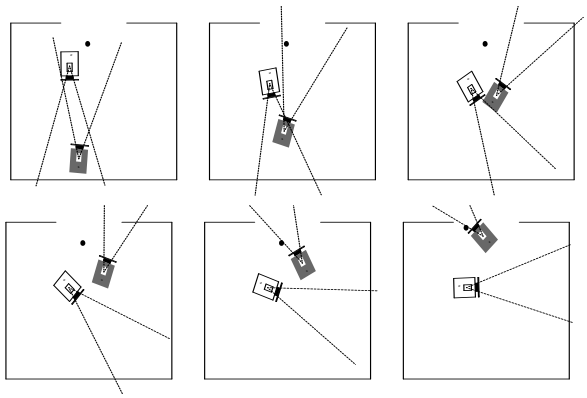

Fig.1 A shooting behavior of the *learning*

黒いロボットが学習したロボットである．ロボットからでている 2 本の直線は，ロボットの視野を表している．

5 おわりに

本報告では移動ロボットの行動を強化学習を使用して獲得するために，サブタスクに分解して，それぞれを独立に学習させた後，統合することによって達成した．サブタスク間の状態集合の干渉という問題に対して，新たに状態を追加し再学習を行なう方法を提案し，その有効性をシミュレーションで確認した．

今後の方針としては，キーパーの防御の学習，味方を加えた場合の協調の学習などが考えられる．

参考文献

[1] 野田，浅田，俵積田，細田. “強化学習によるロボットの行動獲得の効率化に関する考察-簡単なタスクからの学習 LEM-”. 第 4 回ロボットシンポジウム予稿集, pp. 67–72, 1994.

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

[3] S. Whitehead, J. Karlsson, J. Tenenberg. “Learning Multiple Goal Behavior Via Task Decomposition And Dynamic Policy Merging”. *Robot Learning*, pp. 45–78, 1993.