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遂行のための協調行動の獲得

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

◯内部 英治　浅田 稔　野田 彰一　細田 耕

◯Eiji Uchibe　Minoru Asada　Shoichi Noda　Koh Hosoda

大阪大学工学部電子制御機械工学科

Dept. of Mechanical Engineering for Computer-Controlled Machinery, Osaka University

Abstract: Reinforcement learning has been recently used to build systems that learn to accomplish non-trivial sequential decision tasks. However, there are few attempts at real robot applications and at achievement of multiple tasks of which state space are partly inconsistent with each other. This paper proposes a method which acquire a behavior that achieve such multiple goals based on Q-learning, one of the most widely used reinforcement learning methods. In the first part, three kinds of coordinations are considered: simple sum of different action value functions, switching them according to some situations, and Q-learning with previously learned behaviors. Simulation results show that the simple sum and switching methods could not cope with situations that are inconsistent with the state space of sub-tasks because they did not consider such situations. Only the learning method could cope with them. However, the specification of such situations is done by hand, and the robot so often loses the ball due to its narrow visual field. In the second part, we propose a method to cope with these problems. Virtual visual field is proposed to expand the visual field of the robot based on the estimation of the state transition probabilities, which is further used to detect the situations inconsistent with the state space of sub-tasks. Experimental results are shown and a discussion is given.

1 はじめに

複雑で動的な環境下でも自律的に行動できるロボットには，例えば工場内での荷物の運搬や，ゴミの回収など様々な利用価値があると考えられる．古典的な移動ロボットに関する研究は，まずセンサ等によって環境の状態を観測し，環境をモデリングし，得られたモデルをもとにして計画を立て，実行する．しかしこの様な手法では，環境のモデリングやプランニングに多大な時間を必要とするため，静的な環境下においてもロボットの動作は緩慢になってしまうという問題があった．そこで動的な環境に対処するためにリアクティブプランニングの研究が始まった [16]．Arkin らは目的の行動を生成するために環境の先験的な知識をモータースキーマの中に組み込むことにより，動的な環境下でうまく振る舞うことのできる移動ロボットのためのアーキテクチャAuRA を提案している [1,2]．これらのリアクティブプランニングによる手法では，環境の先験的な知識をいかに有効に組み込むことができるかが問題となってくる．

一方，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，強化学習法が最近注目されている．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 [10,14] や，隠れ状態が存在するような環境下での学習 [5,13,15] などの理論的な考察をしているものが多い．しかし，これらの研究では，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．

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．Gachet らは，タスクを基本となる行動として参照される要素タスクの集合として表し，それを制御する AFREB アーキテクチャを実現した [6]．しかし，これらの研究ではバンパーセンサ，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，野田らは，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．その中で，ロボットはカメラからの画像だけから情報を得て，ボールをゴールにシュートする行動を獲得している [4,8]．しかしながら，そこでは単一タスクのみが扱われ，状態が干渉するタスクの実現には向かない．そこで本稿では実ロボットを対象としてタスク間で干渉が生じる場合の多重タスクの統合法を提案する．

まず，前半では，キーパーロボットとの衝突を回避しながらボールをゴールにシュートするという多重タスクを達成するための行動を獲得するための手法を提案する．それぞれ独立に学習したサブタスクの学習結果を，単純に統合する手法，ある条件のもとで切り替える手法などか考えられる．しかし，サブタスク間で状態空間が干渉する場合には有効でない．そこで干渉部分を事前に与えて再学習する方法を提案する [3,11]．

後半では提案した手法のいくつかの問題点について指摘し，その解決のための一手法を提案する．最も大きな問題は，ボールをゴールにシュートすることはできない場合があることである．その理由として，ボールやゴールが見えない状態に含まれる状況が大きすぎて，一度ボールを見失ってしまうと非常に限られたケースでしか目的の行動は学習されないということがある．また，キーパーロボットの手前にボールがある場合にも，環境に関

する指標が少なすぎるため，学習ロボットが停留してしまうことがある． 以上のことから環境の状態を識別するために選んだ指標がボール，ゴール，ロボットでは不十分であることがわかる．現実の世界ではボール，ゴールの他に多くの指標が存在しており，そのうち適当な指標と対象との関係に着目して，人間は行動をしていると考えられる．

理想的には，ロボットもタスク達成のために必要な指標に着目し，それを用いて学習することが望まれる．そこで，その最初のステップとして，コーナーポスト，フィールドラインといった指標を増やし，できるだけ任意の場所からシュートできるような行動を獲得するための手法を考える．まず，具体的には，シュートするタスクに関して，ボール，ゴール以外に，ロボットのセンサ情報の欠如を補うために，対象物 (ボール) に関する状態遷移確率行列を用いた，仮想視野を提案する．これによって，見えない領域においても対象物の状態を推定できる．また，状態遷移確率を用いて，状態空間の干渉の発見することができる．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 学習は最近注目されている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 [9,12]．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
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'\in \boldsymbol{A}} Q(s',a')) \tag{1}$$
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，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

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

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，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

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

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 (Perceptual Aliasing Problem) [13] 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，

$$ \begin{array}{lcl} {}^cQ_{rl}({}^cs, a) & = & {}^cQ_{ss}({}^cs, a) \\ {}^cQ_{rl}({}^cs_{sub}, a) & = & {}^cs_{sub}\text{に最も近い状態の } {}^cQ_{ss}({}^cs, a) \end{array} \tag{7} $$

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ことである．ロボットに関しては，搭載されたカメラ画像だけから情報を獲得し，自身の幾何学的パラメータや動的特性などは知らされていない．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) システムを持っている．まず，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．それぞれのパラメータを次のように定義する．

1. 状態空間 $\boldsymbol{S}$:
   ボールにの状態は，画像上での位置 (重心の水平軸上の位置:左，中央，右)3 通りと大きさ (ボール半径:大，中，小) の3通り，および，観測されない場合の「右に消えた」,「左に消えた」の2状態の合計 $3 \times 3 + 2 = 11$ 通りとする [8]．　ゴールについては，ボールと同様の位置 (水平軸上の座標重心)，大きさ (垂直軸方向の長さ) に加えて向き (ゴールの傾き:右向き，正面，左向き) の状態 および観測されない場合の「右に消えた」,「左に消えた」の 2 状態の合計 $3 \times 3 \times 3 + 2 = 29$ を設定する．

2. 行動空間 $\boldsymbol{A}$:
   ロボットの移動する速度を分割し，行動空間を構成する．ロボットとは PWS 車であるため，左右輪がそれぞれ独立に前進，停止，後退ができるので，合計 9 通りの行動をとることができる．この行動は状態と行動のずれ問題のため，状態が変化するまで同一の行動をとり続けるとする [4]．従って，両輪停止の行動は状態の変化を起こさないため，これを除外する．したがって行動は 8 通りとする．

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)
- ${}^aT$: キーパーとの衝突を回避する (avoidance)

サブタスク ${}^sT$ の学習は [8] を参照されたい．また，サブタスク ${}^aT$ の学習は2.2節で述べた通り，学習中は衝突することを学習する．

二つのサブタスクを統合した場合に,「ボールがキーパーに隠された」という状態はもとのサブタスク間の状態空間の直積には含まれていない．これがサブタスク間で干渉する状態に相当する．そのため，再学習による手法では，この状態を新たに追加して学習を行う．追加した状態を重点的に学習するために，追加した状態ではボルツマン分布の温度パラメータ T を高く設定し，それ以外では低く設定する．さらに，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，次第にキーパーが移動する環境に変更して学習させる．また,「ボールが隠された」という判断は，一つ前にボールが見えていた所に，現在キーパーロボットが見えている場合とする．

シミュレーションに用いたパラメータを **Table 1** に示す．また学習率 α は全て 0.25 で一定である．

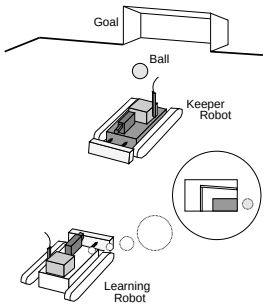


Fig.1 Learning robot regards this situation as ball-lost

Table 1 Parameters

task	# of states	γ
${}^{s}T$	319	0.9
${}^{a}T$	11	0.1
learning	3770	0.9

Table 2 Simulation results

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

3.2 シミュレーション結果

スイッチング時の政策の切替え条件は,「キーパーのみが見えた」場合であり，その時 ${}^{a}Q$ を用いる．キーパーがシュートコースを遮るような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．

Fig.2 において，黒，白の長方形がそれぞれ学習ロボット，キーパーロボットに対応している．各ロボットから出ている 2 本の線は，ロボットの視野を表している．再学習による手法がシュート率，衝突までのステップ数が最大になっている．獲得された行動の特徴を述べると，シンプルサムによる手法ではポテンシャル法と同様に停留点問題が生じてしまうが，再学習ではそのような状況が発生しにくい結果となった．また，スイッチングによる手法では，政策の切り替えを人間が与えており (キーパーロボットだけが見えた場合)，それが適切でないため，再学習による手法ほどの結果は得られていない．このことから，どの状況で政策を切り替えるかを決定することは容易ではない．

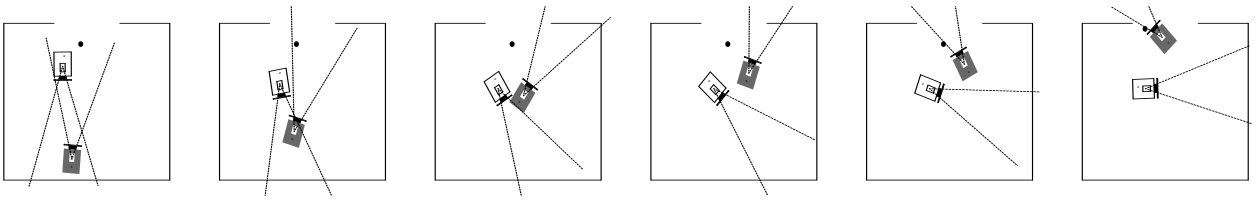


Fig.2 A shooting behavior avoiding collisions with a keeper robot (learning method)

4 指標の追加と状態遷移確率の導入による視野の拡大および干渉の検出

4.1 上で述べたシステムの問題点

今回提案した多重タスク達成のための手法の欠点として，以下のようなことが考えられる．

1. 指標の欠如
   環境内に存在するのがボールとゴール，そしてキーパーロボットだけであり，指標が少なすぎる．例えば，ボールが見えていて，ゴールが左に消えていた状態のとき，どの程度右回転すればゴールが見えるようになるかということが，これだけでは区別がつかない．

2. 知覚的見せかけ問題 (Perceptual Aliasing Problem) [13, 15]
   自律的な移動ロボットを想定しているため，センサは全てロボットに搭載されることになる．視覚を用いて環境の情報を獲得する場合には，カメラの画角が問題になる．実験で用いたカメラの画角は約 36 度しかなく，ロボットが回転することで対象物を容易に見失ってしまう．結果としてロボットは環境の状態を部分的にし

か観測できない．具体的にはボールやゴールなどが見えなくなる状態の発生確率が高く,「右に消えた」「左に消えた」といった状態が有効である場合が非常に限られている．

3. 状態空間の干渉部分の検出
サブタスク間の状態空間が干渉することを人間が，あらかじめ与えてから学習を行っている．今回のタスクは，サブタスク間での状態空間が干渉する部分は「ボールがキーパーロボットに隠された」という状態であることは容易に分かる．しかし，より複雑なタスクに適用する場合，干渉する状態をあらかじめ想定することは困難である

4.2 状態遷移確率を用いた仮想視野と干渉部分の発見

環境内の指標の欠如の問題は，実際に指標を増やすことで容易に達成できる．しかし，知覚的見せかけ問題や，状態空間の干渉部分の検出といった問題は従来のQ学習だけでは対処できない．知覚的見せかけ問題に対処するために，カメラを動かすなどのアプローチも考えられるが，ここでは状態遷移確率を用いた仮想視野に基づく手法を提案する．

仮想視野を用いた手法では，ロボットのセンサ情報の欠如を補うために，本来は対象物が見えていない場合でも，見えている範囲で学習した結果を用いて，見えない領域に入った場合の位置を推定する．これは，見えない領域での対象物に関する状態遷移は，見えている範囲と同様の状態遷移を行うという仮定に基づいている．Q学習では，環境の状態遷移確率を推定することなしに目的行動を獲得する手法であるため，Q値を用いて，次の状態を推定するのは困難である．そこで，対象物の状態遷移確率$P_{ij}(a)$の推定は最尤推定

$$P_{ij}^{k}(a) = \frac{n_{ij}^{a}(t)}{\sum_{j \in \boldsymbol{S}} n_{ij}^{a}(t)} \tag{8}$$

を用いる．ここで$n_{ij}^{u}(t)$は時刻tまでに，行動aによってiからjに状態遷移が行われた回数を表す．ロボットは環境との相互作用を繰り返してQ値を推定しながら同時に対象物の状態遷移確率を推定する．

Fig.3 に仮想視野の概念図を示す．学習ロボットは，環境との相互作用により，真の画面で対象物が見えている場合に限り，状態遷移確率$P_{ij}(a)$を推定する．もし，真の画面上から対象物が消えた場合，推定した$P_{ij}(a)$を用いて，ボールの状態を推定する．状態遷移のタイミングは，真の画面上で観測した状態が変化した時に遷移する．仮想視野から対象物が消えた場合に，その対象物は，真に消えたと判断する．

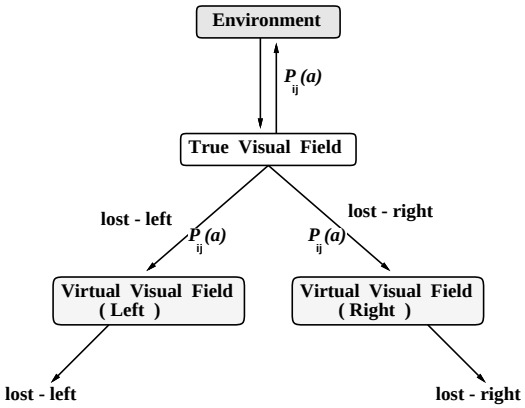


Fig.3 Relation true visual field and virtual visual field

また，状態遷移確率を用いて，サブタスク間の状態空間の干渉を統合時に発見することが可能である．遷移確率の検定についてであるが，Chrisman らと同様にχ^2検定を用いる [5].

5 仮定と拡張したタスク

5.1 各空間の構成

環境内でにロボットが認識できるのは，ボール，ゴール，コーナーポスト0〜3，フィールドライン0〜4とする．それぞれのパラメータを次のように定義する．

1. 状態空間$\boldsymbol{S}$:
仮想視野を使用する場合，ボールの状態は，これまでの画像上での位置を3通りと大きさの3通りの組合せが3枚分必要となる．更に，観測されない場合の「右に消えた」,「左に消えた」の2状態の合計し，$3\times3\times3+2=29$通りとする．

   ゴールについては，これまでの画像上での位置，大きさ，向きの状態を考える．追加するコーナーのポストについては画像上の位置(水平軸上の座標重心)，大きさ(垂直軸方向の長さ)を，ラインについては位置(垂直

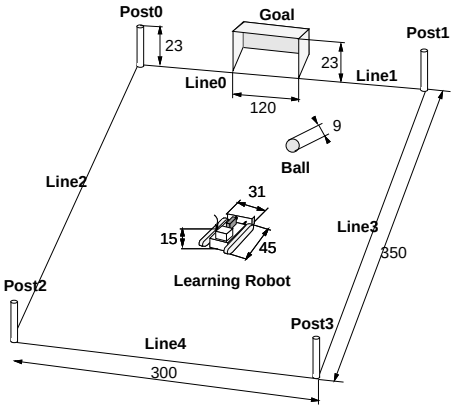


Fig.4-a The task of shooting a ball into the goal

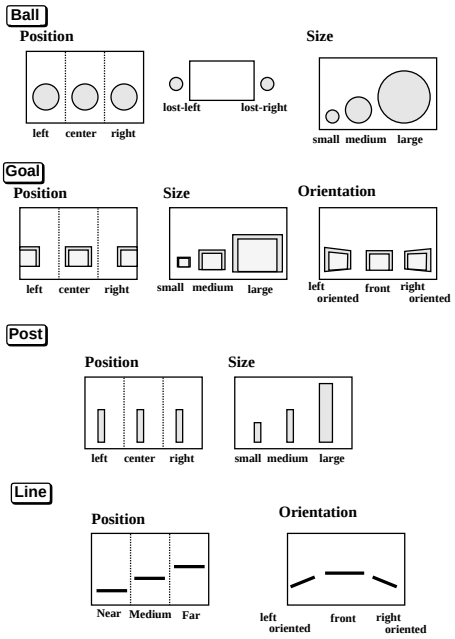


Fig.4-b The substates of the ball, goal, post, and line

軸上の座標重心，距離の代わり) と向き (傾き:負，0，正) の状態を設定した，つまりゴールは $3 \times 3 \times 3 = 27$, ポストは $3 \times 3 = 9$，ラインは $3 \times 3 = 9$ 通りになる．ただし，ゴール，ポスト，ラインを単純に組み合わせると状態数が増えすぎ，しかも物理的には発生し得ない状態を構成する可能性がある．今回は優先順位をつけることにより対処した．優先順位はゴール，ポスト，ラインの順で，これらが同時に見えた場合には，優先順位の高いものしか見えていないとして，状態を判断する．ゴール，コーナーポスト，フィールドラインに関しては，観測されない場合の「右に消えた」,「左に消えた」という状態は考えず，ゴール，ポスト，ラインのいずれも見えない状態の一つにまとめた．これらをまとめると 109 通りになる．

2. 報酬 r:
ロボットがボールをゴールにシュートし，かつ，ボールを蹴った時にボールとゴールが見えている場合に，ボールを蹴った状態と行動の行動価値関数に対して 3 の報酬を与え，それ以外は 0 とした．

3. 学習率 α および減衰係数 γ:
学習率 α の値は 0.25，また減衰係数 γ の値は 0.9 とそれぞれ固定した．

5.2 シミュレーション結果

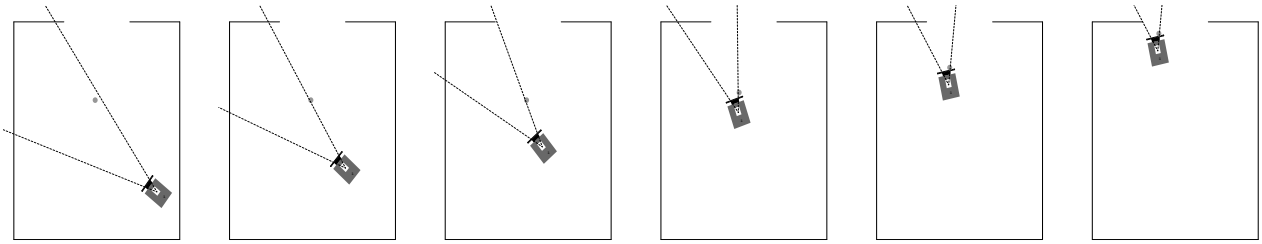


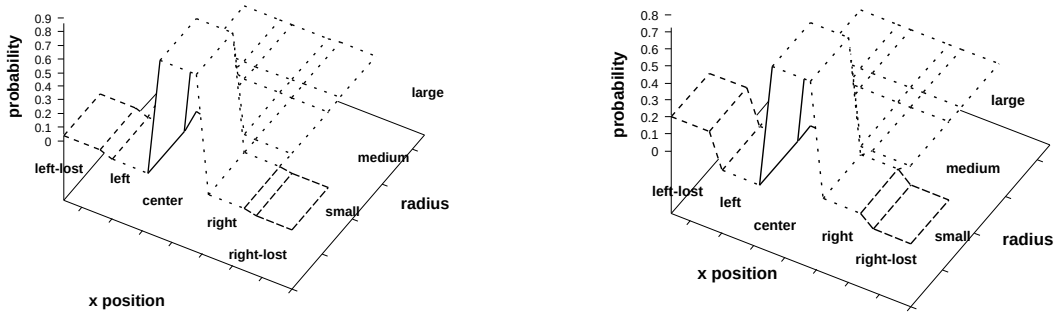
Fig.4 A shooting behavior with virtual visual fields

提案した手法の有効性を検証するために，シミュレーションを行った．シミュレーションの環境は，**Fig.4-a** に示す通りである．また，カメラはロボットの中心部に搭載されており，画角は約 36 度である．シミュレーションの 1step は 1/30[sec] と対応している．ロボットの最大速度は 3.8/step，最大回転角速度は 9.1 度/step，ロボットの質量とボールの質量の比は 100:1，跳ね返り係数は 0.5 とした．ボールの転がりに関して，床との摩擦係数は 0.2 とした．

まず，仮想視野を用いた場合のシミュレーション結果を **Fig.4** に示す．指標を設定し，仮想視野を用いた場合にはゴールが見えていない場合でも，ボールをゴールをシュートすることができた．

次に状態遷移確率を用いて，サブタスクの干渉部分の検出を行なった結果を示す． **Table 3** は状態空間が干渉されたと判断された状態と行動を表している．ここで，ロボットの行動は，forwardが前進，backward が後退，right(left) spin turn が右 (左) まわりに旋回，right(left) pivot turn が右 (左) まわり信地旋回，right(left) pivot reverse が右 (左) まわりに逆信地旋回をあらわしている． **Table 3** において，例えば (left small), (right small), (forward) の組は，ボールが左に小さく，キーパーロボットが右に小さい時に，前進の行動をとったあと遷移した状態が干渉する状態であると判断されたことを意味している．

Fig.5 は， **Table 3** のボールが左に小さく見えている時 (∗ の所) に後退の行動をとった場合のボールに関する状態遷移確率である (統合の場合の状態遷移確率はキーパーロボットが左に小さく見えている時)．シューティングの時と比較して，統合した時には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．



(a) shooting

(b) integration

Fig.5 Simulation results

5.3 考察

始めから，ゴールが見えていない場合でも，ボールをゴールにシュートできる状況は増えた．しかし，より複雑な状況 (学習ロボットの前方にゴール，後方にボール，もしくはその逆) については，行動が獲得されなかった．このような状況でもシュートするためには，一度ボールを完全に見失う必要がある．ボールを見失った時，次にどちらに行動すべきということが，不確定になりすぎるためであると考えられる．

次にサブタスク間の状態集合の干渉する部分の発見であるが，学習中の経験に偏りがあるため，発見された「干渉する状態」も偏りが発生した．また，はじめ想定していた「干渉する状態」は「キーパーロボットにボールが隠された」という状態 (キーパーの状態が lost から遷移した時に発見された時に相当) であったが,「キーパーロボットがボールを移動させた」という状態 (ボールが左に見えて，キーパーロボットが右に見えている状態から遷移した時に相当する) も干渉状態として発見された．

キーパーロボットによってボールが動かされたという状態は，3.1節で想定されていない干渉状態である．この状態は自分以外に積極的に行動するロボットのために発生する状態である．このようなマルチエージェントの環境では，強化学習の枠組はそのまま適用できない [7]．発見された状態に対して，どのような学習アルゴリズムを適用するかということが問題となってくる．

今回は状態空間の増加のため，仮想視野と状態空間の干渉部分の検出を同時に行なうことは行なわなかった．状態数を少なくするために，環境の指標に優先順位を設定したが，結局それは人間が与えているため，最適な状態空間の構成になっているとは限らない．

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」といった二つのサブタスクを視覚を用いた強化学習を適用して，それぞれ独立に学習し，統合した．また，サブタスク統合の際，サブタスク間で状態空間が干渉する場合には，単純に政策の切替えを行うだけでは不十分で，干渉する状態を追加して再び学習を行う必要があることを指摘し，シミュレーションによってその有効性を示した．次に，環境の指標を増やした環境下で，センサ情報の欠如を補うために，対象物の状態遷移確率を用いた仮想視野による手法を提案した．同時に，状態遷移確率を用いて，タスク統合時のサブタスク間の状態空間の干渉する部分を発見できた．

今後の課題としては，まず，発見した干渉する状態を分割して再学習を行なう枠組を構築することが挙げられる．状態数の増加に対処する点からは，各タスクで獲得された行動の政策をいかにして切り替えるか，といったことが考えられる．4.1節でも述べた通り，提案した手法では，サブタスク間での状態集合の直積プラス干渉部分と拡張されるため，統合されるタスクが増加すると，それだけ状態数が組合せ的に増加し，結果的に学習時間は増加してしまう．しかし，実際の多重タスクは常に同時に達成しなければならない状況にあるのではなく，限られた状況だけが多重タスクの達成に必要であると考えられる．つまり，通常は政策の切り替えによって多重タスクは達成され，統合時に発生する状況だけを追加して学習する事が望ましい．そういった意味で，状況に応じて行動の政策を切り替えるスイッチングによる手法は重要であり，学習時間の短縮化の点から考えても有効である．

Table 3 Interferent states which learning robot detects

state(ball)	state(keeper)	action	state(ball)	state(keeper)	action
left small	*right small*	*forward*	*left small*	*lost-left*	*backward*
right small	*left small*	*right spin turn*	*left small*	*lost-left*	*left pivot reverse*
right small	*left small*	*forward*	*left small*	*lost-left*	*left spin turn*
$*$ *left small*	*left small*	*backward*	*left small*	*lost-right*	*right pivot turn*
right small	*left small*	*right pivot turn*	*center small*	*lost-left*	*left pivot turn*
center small	*right small*	*right pivot turn*	*center small*	*lost-left*	*forward*
right small	*center small*	*backward*	*center small*	*lost-right*	*left spin turn*
right small	*center small*	*right spin turn*	*center small*	*lost-right*	*forward*
left small	*left small*	*left spin turn*	*right small*	*lost-right*	*right pivot reverse*
right small	*right small*	*right spin turn*	*right small*	*lost-right*	*right spin turn*
left small	*right small*	*right pivot turn*	*right small*	*lost-right*	*right pivot turn*

参考文献

[1] Ronald C. Arkin. Motor Schema-Based Mobile Robot Navigation. *The International Journal of Robotics Research*, Vol. 8, No. 4, pp. 92–112, August 1989.

[2] Ronald C. Arkin. Integrating Behavioral, Perceptual, and World Knowledge in Reactive Navigation. In Patie Maes, editor, *Desigining Autonomous Agents*, pp. 105–122. MIT/Elsevier, 1990.

[3] Minoru Asada, Eiji Uchibe, Shoichi Noda, Sukoya Tawaratsumida, and Koh 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.

[4] 浅田稔, 野田彰一, 俵積田健, 細田耕. 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, Vol. 13, No. 1, pp. 68–74, 1995.

[5] Lonnie 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.

[6] D. Gachet, M. A. Salichs, L. Moreno, and J. R. Pimentel. Learning Emergent Tasks for an Autonomous Mobile Robot. In *Proc. of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 290–297, 1993.

[7] Michael 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.

[8] 野田彰一, 浅田稔, 俵積田健, 細田耕. 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.

[9] Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming.* Academic Press, NY, 1983.

[10] Satinder Pal Singh. Transfer of Learning by Composing Solution of Elemental Sequential Tasks. *Machine Learning*, Vol. 8, pp. 99–115, 1992.

[11] 内部英治, 浅田稔, 野田彰一, 細田耕. 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. 第12回日本ロボット学会学術講演会予稿集, pp. 609–610, 1994.

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

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

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

[15] Steven D. Whitehead and Long-Ji Lin. Reinforcement Learning of Non-Markov Decision Process. *Artificial Intelligence*, Vol. 73, pp. 271–306, 1995.

[16] 山田誠二. リアクティブプランニングにおける学習. 日本ロボット学会誌, Vol. 13, No. 1, pp. 38–43, 1995.