複数の報酬による強化学習を用いたサッカーロボットのゴール守備行動の獲得

Acquisition of a goal keeping behavior of a soccer robot by reinforcement learning

○加藤　龍憲　(阪大)　　鈴木　昭二　(阪大)　　浅田　稔　(阪大)

Tatsunori KATO, Osaka University, 2-1, Yamadaoka, Suita, Osaka

Sho'ji SUZUKI, Osaka University

Minoru ASADA, Osaka University

Abstract: We apply Q-learning, one of major method of reinforcement learning, for behavior acquision by a mobile robot. We choosed a shooting behavor of soccer game as an example and a robot with a single camera could learn it. However, the robot was difficult to acquire more complex behavior such as a goal keeping behavior, because its view was narrow and a reword for learning was too simple. In this paper, we introcude an omnidirectional vision to a mobile robot and apply Q-learning. We propose a new reward function, so that the robot can learn a goal keeping behavior. We perform simulation to verify our proposed method.

Keywords: behavior acquisition, reinforcement learning, omnidirectional vision, keeper, soccer robot

1　はじめに

ロボットに先験的な知識をほとんど与えることなく自律的に合目的な行動を獲得させる手法として強化学習が注目されている[2]．強化学習の枠組みでは，ロボットは観測した環境情報に基づいて自身のとるべき行為を選択し，行動目的の達成度合に応じた報酬を受け取る．観測，行為の選択，報酬受取のサイクルを繰り返すことにより，ロボットは徐々に高い報酬を受取る行為を選択するようになり合目的な行動を獲得する．

これまでに著者らは，視覚を持った移動ロボットに強化学習の代表的な手法であるQ学習を適用し，サッカーにおけるシュート行動を獲得させた[3]．ボールとゴールが存在する環境下で，ロボットは視覚情報に基づきボールをゴールに押し込む行動を学習した．しかし，ロボットの視野が狭いためシュート以外の行動，例えば，ゴールとボールを常に見続ける必要のあるゴール守備行動を学習させることはできなかった．

そこで，著者らは視野を広げるために全方位視覚を導入し全方位視覚の画像情報を用いたQ学習によりゴール守備行動の獲得を試みた[4]．しかし，報酬の与え方に対する考察が不十分であったために獲得された行動は性能が高くなかった．

本報告では，ロボットにより高度なゴール守備行動を学習させるための報酬の与え方について述べる．シュート行動の場合，ロボットはゴール近くでボールをゴールに向かって押し出せばよく，比較的単純な報酬でシュート行動を学習できた．これに対し，ゴール守備行動の場合は，ロボットはゴール又はボールの位置に応じてゴール守備に適切な位置に移動する必要があるため単純な報酬では学習できない．そこで著者らは，複数の報酬を組合せた報酬関数をロボットに与えることによりゴール守備行動を学習させる手法を提案し，シミュレーションにより有効性を検証する．

2　環境及びタスク設定

近年，動的な環境下で作業するロボットチームの実現に必要な技術をサッカーを通じて追求することを目的とするロボカップが注目されている[1]．本研究は，ロボットの自律的な行動の獲得をサッカーを通じて試みている．環境は，ロボカップ中型部門の競技場を想定し，幅4575[mm]長さ8220[mm]の水平な平面を仮定する．ゴールの大きさは幅1500[mm] 高さ600[mm]とする．環境にはゴールとボールのみが存在する．ゴールとボールは画像上で色により識別できるものとし，それぞれ青と赤に色分けされている．また，ボールの大きさは直径200[mm]で床面上を転がる．競技場の周囲は壁で囲われており，ロボットとボールは競技場外に出ないものとする(図1)．

図2に著者らの用いる移動ロボットを示す．ロボットはPWS(Power Wheeled Steering)方式の移動機構と全方位視覚センサで構成されている．ロボットの大きさは，幅300[mm]長さ450[mm]高さ350[mm]であり，最高速度は4.8[m/s]である．ロボットの行為として前後移動・左右回転・停止が予め実装されておりホストコンピュータから無線を通じて実行される．

全方位視覚は，双曲面鏡とCCDカメラにより構成され，双曲面鏡はその中心軸がCCDカメラの光軸と一致するように取り付けられている．ロボットの周囲にある物体は図4(a)に示すように双曲面鏡を通じてカメラの撮像面上に投影される．図4(b)に全方位視覚により得た画像の例を示す．全方位視覚は，ロボットが回転しても視野の中心が変化しないように，視野の中心がロボットの回転中心に一致するように取り付けられている．

ゴール守備行動においては，ロボットはボールを押し出すことによりゴールから遠ざけなければならない．また，同時にロボットはゴール近くから遠ざかりすぎないことも求められる．したがって，ロボットの学習

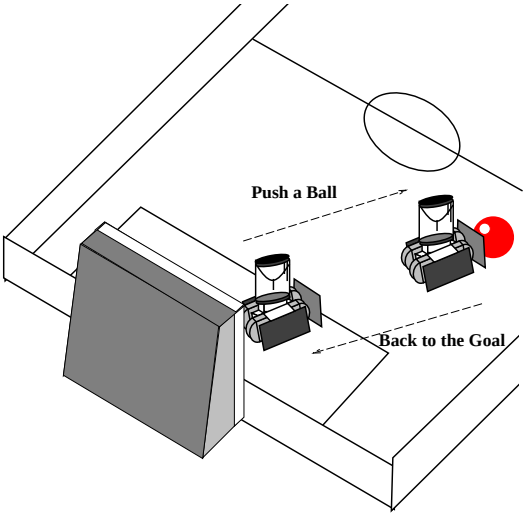


Fig.1 The environment and the task

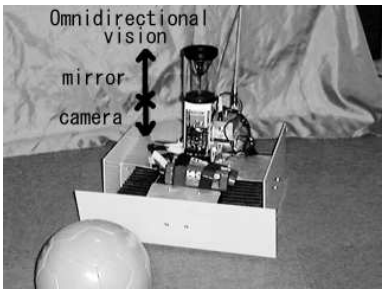


Fig.2 The robot

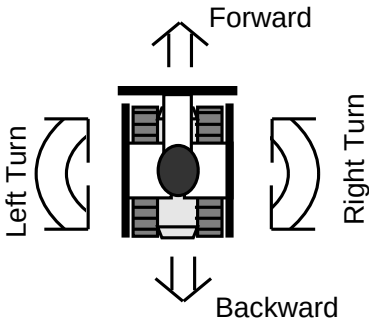
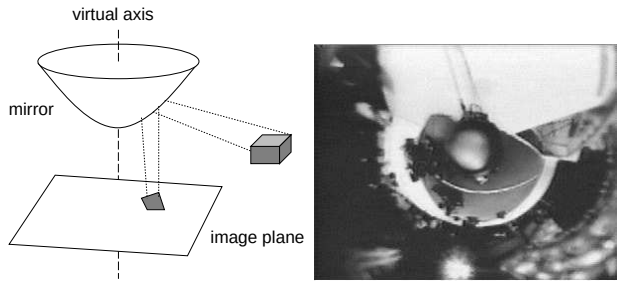


Fig.3 Actions of the robot

すべき行動には，ボールをゴールから遠ざけることと，ゴールから離れすぎないことの複数の目的がある．本研究では，このような行動を学習するための報酬の与え方を提案し，Q学習によりロボットにゴール守備行動を獲得させる．



(a) projection (b) image

Fig.4 The omnidirectional vision on the robot

3 複数報酬によるゴール守備行動の学習

3.1 Q学習の概要

ここでは，Q学習[6]の基本的な枠組を述べる．Q学習においては，ロボットは現在の環境の状態sを観測し，実行すべき行為aを選択する．行為aの実行により，環境から報酬rを受け取り，環境の状態はs'へと遷移する．この時，行動価値関数$Q(s,a)$の更新を(1)式によって行う．

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

ここで，αは学習率であり報酬の影響の受けやすさを決める定数である．γは減衰係数であり，将来得る報酬の影響の受けやすさを決める．ロボットが観測と行為選択を繰り返すことにより，行動目的に有効な行為に対するQ値が更新されていく．

Q学習をロボットの行動獲得に適用するためには，ロボットに与える報酬，ロボットの観測できる状態およびロボットの選択できる行為を定義しなければならない[3]．以下それぞれについて詳しく述べる．

3.2 報酬関数の設計

複数目的を同時に達成するためにそれぞれの目的に応じた報酬を与える．具体的には，ロボットがゴールの近くに留まる，及びボールをゴールから遠ざけるための報酬を考える．図5にロボットに与える報酬を図示する．最終的にロボットに与える報酬はこれらの重み付き和を用いる．

ロボットがゴール近くに留まるために，ゴールからロボットまでの距離x_aを変数とする報酬r_aを，

$$r_a = \begin{cases} -2x_a/\Delta + 1 & (0 \leq x_a \leq \Delta) \\ -1 & (\Delta < x_a) \end{cases} \quad (2)$$

と定める．ここで，Δはゴール付近を設定するパラメータである．ここでは，ゴール前のゴール幅を直径とする半円内をゴール付近とし$\Delta = 750mm$とする．

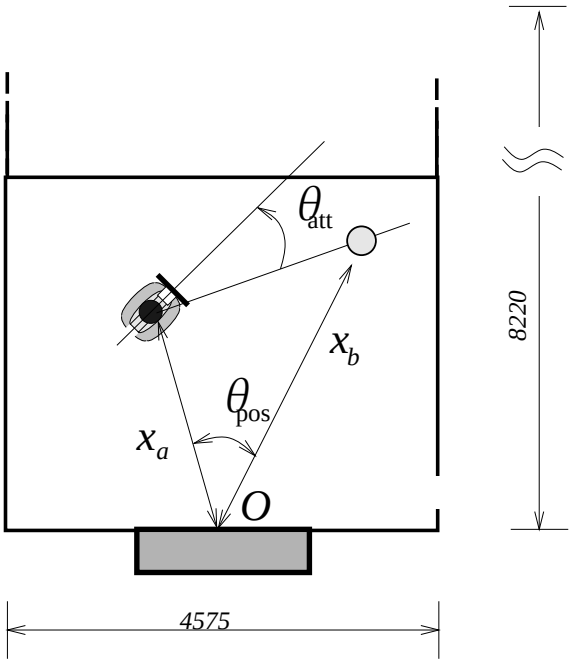
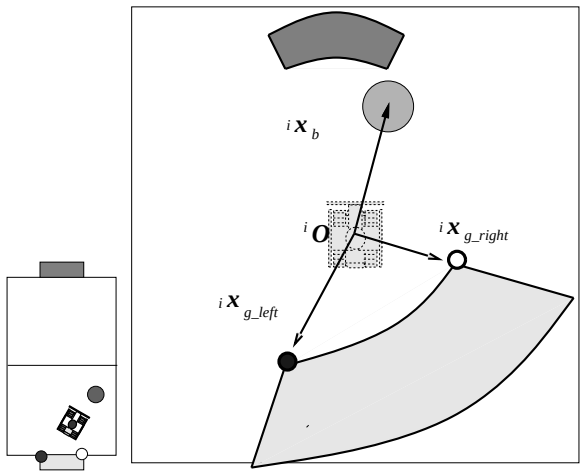


Fig.5 報酬のための変数設定



(a) the Real world

(b) an image of the Robot

Fig.6 definition of the state set

次に，ボールをゴールから遠ざけるために，ゴールからボールまでの距離x_bを変数とする報酬r_bを，

$$r_b = \begin{cases} 2x_b/L - 1 & (0 \leq x_b \leq L) \\ 1 & (L < x_b) \end{cases} \quad (3)$$

と定める．Lはゴールを守備する範囲を定め，ここではLをフィールド全長(=8220mm)の3/8の長さにとった．また，ロボットは前後移動の行為で主にボールを操るので，ボールをロボットの正面に捉えるための報酬，

$$r_{\theta_{pos}} = \cos\theta_{pos} \quad (4)$$

および，

$$r_{\theta_{att}} = \cos\theta_{att} \quad (5)$$

を定義する．θ_{pos}はゴール中心に対するロボットとボールの重心がなす角度であり，θ_{att}はロボット正面を基準としてボールの方向を表す変数である．ボールがロボットの正面に近いほど大きな報酬となる．

以上 4種類の報酬は環境内のロボットとボールとゴールの位置から計算される．ロボットにはこれらを足し合わせたものを報酬関数として与える．

$$\begin{aligned} r &= w_a r_a + w_b r_b + w_{\theta_{pos}} r_{\theta_{pos}} + w_{\theta_{att}} r_{\theta_{att}} \\ w_a &+ w_b + w_{\theta_{pos}} + w_{\theta_{att}} = 1 \end{aligned} \quad (6)$$

$w_a, w_b, w_{\theta_{pos}}, w_{\theta att}$は重みであり，重み付けによりロボットの行動の特徴が変化する．例えば，w_bが大きい程ボールへ積極的に向かう動きとなる．

3.3 状態の定義

ゴール守備行動獲得のための状態を画像上のゴールとボールに関する情報をもとに次のように定義する．

ゴールについては，ゴールポスト下端 2点の観測から得られる画像上での 2つのベクトル$\boldsymbol{x}_g$を，それぞれ画像中心からの距離r_gと画像上での方位θ_gに分解し状態変数とする(図6)．r_gは図7のように同心円状に4段階に分割する．θ_gは図8(a)の様に8等分に分割する．

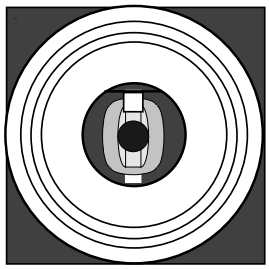


Fig.7 距離成分の分割

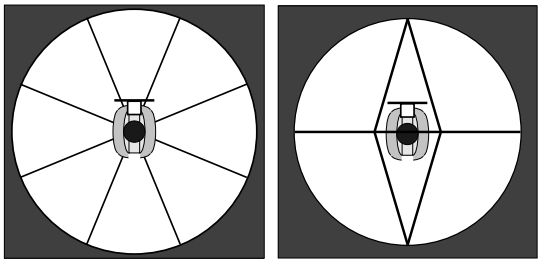
ボールについては，ボール重心を観測して得られる画像上でのベクトル$\boldsymbol{x}_b$を，画像中心からの距離r_bと画像上での方位θ_bに分解し状態変数とする(図6)．r_bに関してはr_gと同様に4 段階に分割し，θ_bの分割に関してはロボットの前後方向に対して敏感になるように図8(b)の様に行なう．

このように状態を定義するとロボットの識別する状態は24576種類となる．

3.4 行為の選択とQ値の更新

ロボットが可能な行為は前後移動・左右回転・停止とする．学習中の行為選択は，90%の確率でそれまでに獲得した最適な行為を選択し，10%の確率でランダムとした．Q値の更新は(1)式に従う．この時，学習率αは0.80とし報酬の影響を受けやすいようにした．減衰係数γは$\gamma = 0.1$とし，遷移後の状態において受け取る報酬の伝播を小さくして学習させた．

学習中はある行為を選択した結果，状態が変わればQ値を更新する．状態が変わらない場合は，変わるま



(a) ゴール　　(b) ボール

Fig.8 方位成分の分割

でQ値を更新せずに選択した行為を実行し続ける．しかし，停止が最善の行為である場合は，状態遷移が起こらないため最善の行動を実行しているにも関わらずQ値が更新されない．これを防ぐために，停止を選択した場合は状態が遷移しなくとも一定時間毎にQ値を更新する (図9).

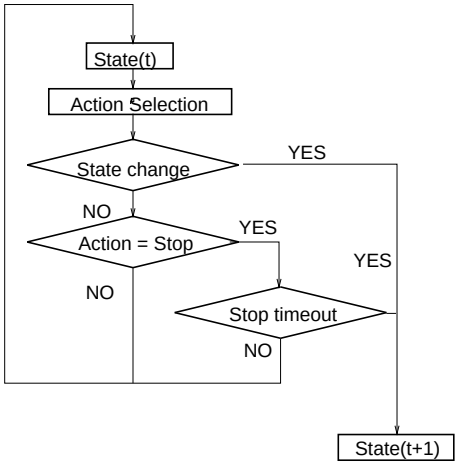


Fig.9 Action selection

4　実験結果

ゴール守備行動を獲得するシミュレーションを行った．ロボットとボールの初期位置を図10に示す．ロボットはゴール直前に配置し，ボールはロボットが観測しやすいようにロボットに接近した位置に配置する．ロボットによる行動学習を効率よく行うためには，段階的な学習が有効である[5]．そこで，学習を3段階的に分けて行った．まずは，図10に示される“center”の位置に配置する．そこで，ある程度の学習を行った後，“left”“center”“right”の 3ヶ所のゴール・ボール・ロボットの初期配置について学習を行う．最後に，ロボットの姿勢を正面から正負90度以内の角度でランダムに配置し学習を行った．この時も，ボールはロボットに接近した位置に配置する．

一試行は，1000step経過した時に終了する．1stepはカメラが画像を取り込むのに要する時間で 33[ms]である．シミュレーションに要する時間は，1000試行当たり15分程度である．

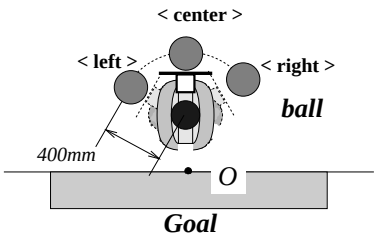


Fig.10 initial position

獲得された行動の例を図13，14に示す．図13ではロボットは試行開始と同時にボールを押しながら前進し(図13(a)(b)(c))，ある程度ゴールから離れると後退してゴール前に戻っていく様子が示されている (図13(d)(e)(f))．異なる初期姿勢から試行を開始した時も図14に示す通りボールを押し出してゴール前へ戻って来ることが分かる．

図11から図12に学習経過のグラフを示す．

図11は各状態のQ値の最大値の和を示している．グラフ上では8000試行までが第一段階，14000試行までが第二段階，残りの試行が第三段階である．それぞれの段階である程度のQ値の収束が示されるまで学習が行われていることが分かる．

図12は，学習途中における各試行において，その時点で獲得された行動を評価した結果である．各時点において，学習時とは別に 100試行を行い得られた報酬の和を示している．学習過程の第一段階では得られる報酬の値は大きいが，ばらつきが大きい．第二段階ではばらつきが小さくなり，得られる報酬の和の最小値が，第一段階より大きくなっている．報酬の和の最大値が第一段階より小さく成ったのは，ロボットの経験する状況が増えたために各状況全てにおいて最大の報酬を得るのが困難になったためと考えられる．第三段階においても，第一段階から第二段階への変化と同様な傾向が見られる．学習過程を通して得られる報酬の和の最小値が次第に大きくなっていることから，徐々に行動が改善されていると考えられるが，最大値が増大していないことから行動が無難な程度に留まっていると考えられる．これに関しては，学習過程において第一段階，第二段階とすすむにしたがってゴールからの離れる距離が短くなっていることが例に挙げられる．

ロボットは試行開始時の初期位置に停止していても報酬を得ることができる．しかし，ボールを押し出した後ゴール前に戻って来る事により，ボールを押し出している間は報酬が減少するものの，最終的には初期状態より大きな報酬が得られる．このことは，Q学習のパラメータである減衰係数γの値に大きく影響を受ける．γが1.0に近い場合，行動価値関数全体に報酬が伝播してしまうため，どの行動が最良なのか明確で無くなってしまう．また，報酬が次々に伝播して行くためQ値の収束は確認できなかった．一方，$\gamma = 0.1$のように極端に減衰係数を小さくした時，Q値は収束し，学習完了の目安となる．γを小さくすることは，ある瞬間での状態に対する最適な行動がその時の報酬に強く影響を受けることを意味している．逆に，報酬関数の設計が獲得される行動に反映される．

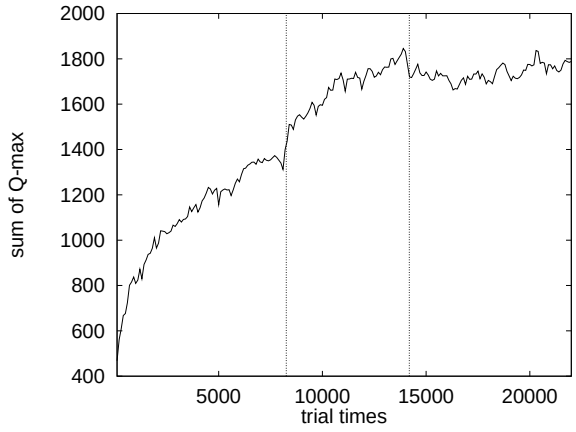


Fig.11 Q 値の最大値の総和

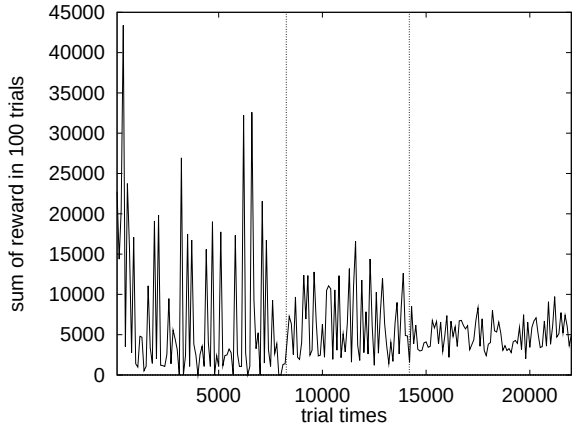
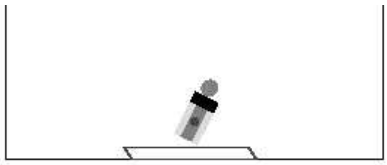
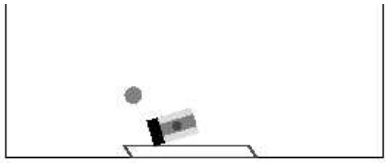


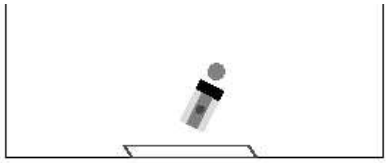
Fig.12 100 試行当たりの報酬の和



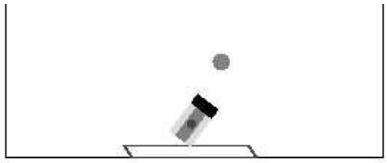
(a) Trial start



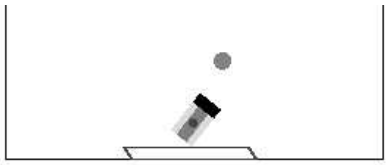
(b) Push a ball



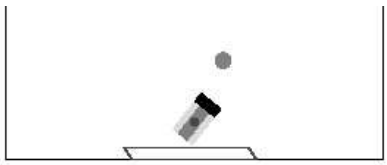
(c)



(d) Back to the goal

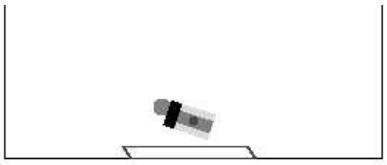


(e)

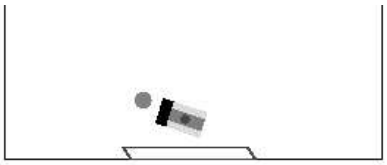


(f) Stay front of the goal

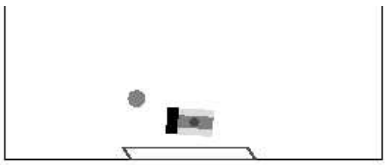
Fig.13 An acquired behavior (1)



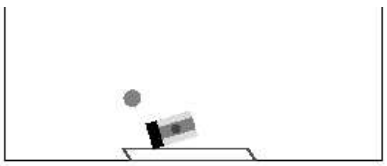
(a) Trial start



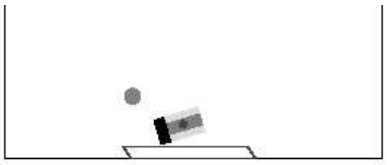
(b) Push a ball



(c) Back to the goal



(d) Stay front of the goal



(e)

Fig.14 An acquired behavior (2)

5 まとめ

本報告では，高度な行動を獲得するために複数の報酬を組み合わせて強化学習の報酬関数としてを用いる手法を提案しシミュレーションにより検証した．

今回行ったシミュレーションで得た学習結果は大部分において良好な結果を示した．しかし，タスクに失敗することも幾つかの初期姿勢において見られる．その原因としては，

1. 衝突の仕方によりボールの転がり方が変わるので状態遷移が一意でない．
2. 移動機構のノンホロノミック拘束のため，即座に動きたい方向へ移動できない．

などの要因が考えられる．

一つ目の問題点に対しては，状態量として速度を含めることで改善が見られると考える．ただし，その場合，現在より識別すべき状態量が増大してしまうため学習に要する時間や記憶容量の増大に対処する方法が必要となる．

二つ目の問題点に対しては，駆動機構そのものを変更するのもよいが，現在の機構のままモーターへの入力を工夫することで改善が可能だと考えている．

今後の課題は，これらの問題への対処および実ロボットによる実験である．

参考文献

[1] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara. “RoboCup: A challenge problem of ai”. *AI Magazine*, 18:73–85, 1997.

[2] Connel, J. H., Mahadevan, S.: Robot Learning. Kluwer Academic Publishers (1993)

[3] 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.

[4] 加藤龍憲・鈴木昭二・浅田稔: “強化学習によるゴール守備行動の獲得”, 第3回 JSME ロボメカ・シンポジア講演論文集, pp.37-40, 1998.

[5] 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.

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