視覚に基づく強化学習によるサッカーロボットのシューティング行動の実現

◯俵積田 健　　浅田 稔　　野田 彰一　　細田 耕
大阪大学 工学部

Realization of A Shooting Behavior of Soccer Robot By A Vision-Based Reinforcement Learning

◯Sukoya TAWARATSUMIDA　　Minoru ASADA　　Shoichi NODA
Koh HOSODA
Osaka University

1　はじめに

自律移動ロボットに，環境に対する適応的，反射的行動を獲得させる一手法として，環境やロボットの明確なモデルを必要としない学習法である強化学習 (Reinforcement Learning)[1] が注目されている．しかし，強化学習に関する従来の研究は，簡単なシミュレーション上における学習の解析がほとんどであり，実際のロボットシステムに応用した例は少ない [2].

我々は，強化学習の一種である Q 学習を適用した自律ロボットシステムの実現を目的としている．強化学習を実ロボットへ適用する際には，複雑な実世界に対して状態空間をどう設定するか，状態数の指数関数のオーダを必要とする学習時間 [3] をどのように低減するかという問題がある．状態空間は，実センサーの使用に基づいて設定されるべきであるが，これによりロボットの行動と一対一に対応する状態空間を構成できるとは限らない．そこで，この “状態と行動のずれ” を解消する学習法をとり入れる．学習時間に対しては，我々の提案する学習の高速化の手法を用いること，実ロボットでの学習を行なわず，コンピュータシミュレーションによる学習結果を利用することで低減化する．

実ロボットを想定したシミュレーションおよび学習の高速化については [4] を参照されたい．本稿では，強化学習を実システムに適用する際の問題点を明らかにし，製作した実システムによる実験を通して，本手法の有効性を検証する．

2　強化学習 [4]

ロボットの識別できるロボットと環境の状態を表す集合を状態空間 $\mathbf{S}$，環境に対してロボットのとることのできる行動の集合を行動空間 $\mathbf{A}$ とする．現在の状態 $s \in \mathbf{S}$ において，ロボットのとった行動 $a \in \mathbf{A}$ により，ある確率で，次の状態 s' に遷移する．その時の状態 s と行動 a に対し，その評価として報酬 $r(s,a)$ が環境からロボットに与えられる．一般の強化学習では，この報酬の積算を最大にするように学習する．

強化学習の一種である Q 学習は，状態 s において行動 a を選択した時，行動価値関数 $Q(s,a)$ を，

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

によって更新する．ここで，α は学習の速さを決める定数 $(0 \leq \alpha \leq 1)$ であり，γ は時間の重み付けをする減衰係数 $(0 \leq \gamma \leq 1)$ である．また，学習後の $Q(s,a)$ に対して，とるべき行動の政策 f は，

$$f(s) \leftarrow a \ \text{ such that } \ Q(s,a) = \max_{b \in \mathbf{A}} Q(s,b) \quad (2)$$

と表され，状態 s に対して, 行動のセット A の中から行動価値 Q が最大となる行動 a を選択する．

3　実ロボットへの強化学習の適用

強化学習を実ロボットの行動獲得へ適用する際に生じる問題について，以下に 3 つの問題点とその対処法を示す．

3.1 実センサーの利用の問題

状態空間は，実ロボットの持つセンサーから得られる情報に基づいて分割する必要がある．ここでは，センサーとしてロボットに搭載された単一テレビカメラのみを用いることを考える．二次元画像情報から，対象物の画像上の位置と大きさが得られるので，これらによって状態を分割するのが自然である．ここで，対象物の見かけの大きさは，ロボットと対象物との距離を暗に示すと考えられる．

3.2 状態と行動のずれの問題

実際のカメラを用いた場合は，視覚情報の特性によって，対象物が近い時は，小さな変化でも画像上では大きな変化になり，その逆もあるため，元の3次元空間での状態を等質に分割することは難しい．そのため，行動と状態遷移が一対一に対応せず，一回の行動によって状態遷移が起こるとは限らないという “状態と行動のずれ” が生じる．そこで，状態が遷移するまでは同じ行動を続け，状態が遷移した時に行動価値関数の値を更新するようにする．つまり，状態が遷移するまでの一連の行動を一つの行動とみなす．

3.3 実ロボットでの学習の問題

状態数が増加すれば，学習の収束までにかかる時間も状態数の指数関数的に増大する [3]．学習のさせ方によって，ある程度学習を高速化できる [4] が，多大な試行回数を要する学習を，実ロボットで行なうことは，コストがかかり過ぎる．そこで，実ロボットによる学習は行なわず，ワークステーション上のシミュレーションによって学習された政策を適用することで，実ロボットの行動政策を効率良く得ることができる．

4 実システム

Q学習の実システムへの適用として，視覚を持った移動ロボットにボールをゴールに入れるというタスクを与える．

4.1 システム構成

システムにおける情報の流れを Fig.1 に示す．このシステムは，Inaba らの提案するリモートブレインシステム [5] を採用しており，ボディ(移動ロボット本体) とブレイン (計算処理装置) は無線で情報のやりとりをする．計算処理装置をロボット本体上に搭載しないので，ロボット本体の小型，

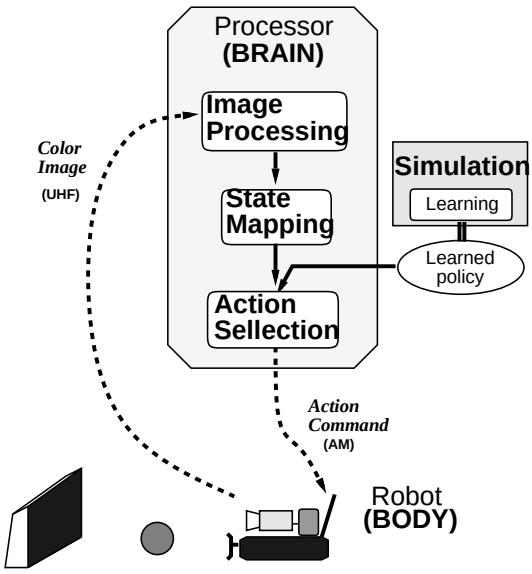

Fig.1 システムにおける情報の流れ

軽量化が可能であり，処理装置の規模も制限されない．

ロボットに搭載されたテレビカメラによって得られる画像は，UHF 電波でブレイン側に送られ，そこで画像処理，ボールとゴールの状態の識別，行動選択を行なう．行動選択には，あらかじめコンピュータシミュレーション [4] によって学習された結果が用いられる．そして，選択された行動をロボットにとらせる制御指令を，AM 電波でロボットに送る．

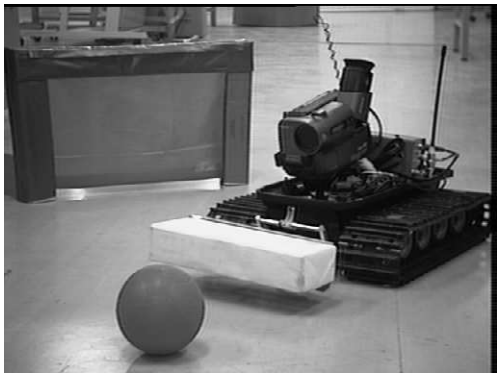

Fig.2 サッカーロボットとボール，およびゴール

Fig.2 に示す移動ロボットの台車は，市販のラジコン車を一部改造したものである．この台車は，2

個のモータコントロールアンプを使用し，左右のモータの回転を独立に制御できる．ロボットの操蛇方式としては，ロボットがその場で回頭できるPWS(Power WheeledSteering) 方式を採用している．ロボットには画像を得るためのカラーテレビカメラ (Sony handy-cam TR-3) と，その画像を画像処理装置に送るためのUHF 送信器がとり付けられている．

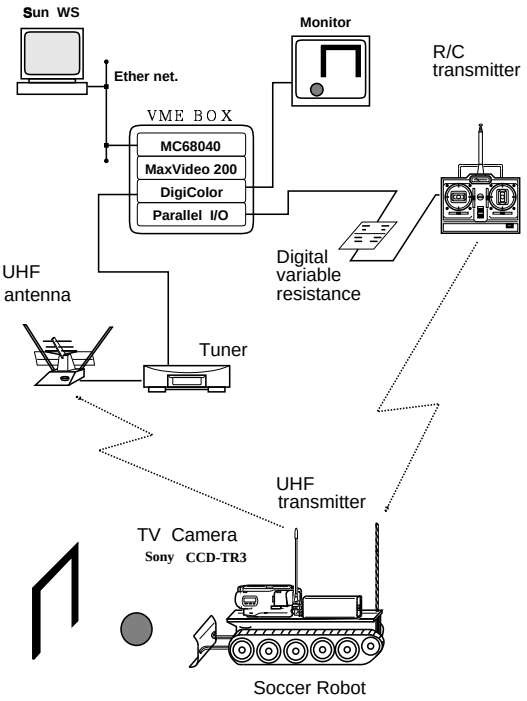

Fig.3 実システムの構成

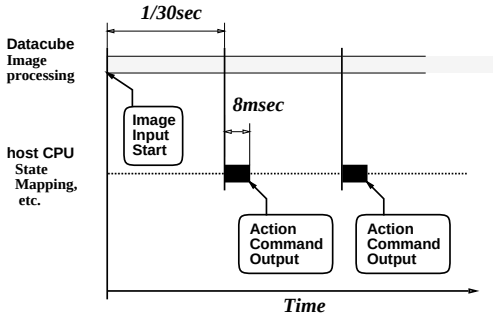

Fig.4 タイミングチャート

実システムの構成を Fig.3 に示す．画像処理には，リアルタイムパイプライン画像処理装置である Datacube Max Video 200 を用いる．DigiColor は Datacube のカラー画像入出力ボードである．また，ホスト CPU には MC68040 を用い，リアルタイム OS として VxWorks を使用する．このホスト CPU は，状態を識別し，行動を選択する．行動の制御指令を，パラレル I/O ボードを介して，ラジコン送信器によって，ロボットに送る．

本システムにおける処理のタイミングチャートを Fig.4 に示す．画像処理による時間遅れはほとんどなく，画像処理にかかる時間が 1/30sec(33.3msec)，状態識別等の処理時間が約 8msec であり，1/30sec 毎にロボットに制御指令を送る．

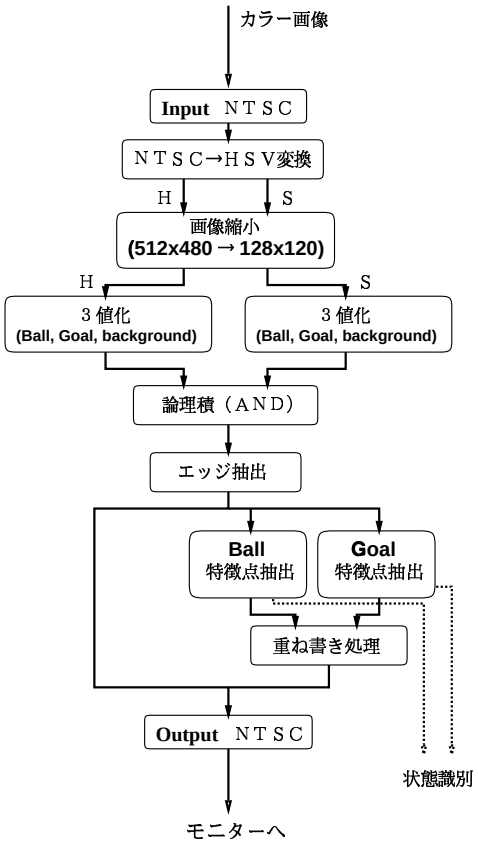

Fig.5 画像処理の流れ

4.2 画像処理

画像処理では，テレビカメラから得られた画像から，ボールとゴールを識別し，ボールとゴールの特徴点を抽出する．ボールとゴールの識別にカラー画像の色情報を用いることによって，画像処

理を簡略化し，その速度を上げることができる．Fig.5 に画像処理の流れを示す．

カメラから得られるカラー画像は，複合映像信号 (NTSC) として入力される．この複合映像信号を HSV(H:色相，S:彩度，V:明るさ) 信号に変換する．彩度の高い赤色のボールと，彩度の低い青色のゴールを使用し，色相と彩度によってボールとゴールを識別する．具体的には，H 画像と S 画像をそれぞれ適当なしきい値によってボール・ゴール・背景に分類 (3 値化) し，色相による識別と，彩度による識別の論理積をとる．その画像から，ボールとゴールのエッジを抽出し，ボール，ゴールの特徴点を抽出する．抽出した特徴点をホスト CPU に読みとり，ボールとゴールの状態を識別する．

Fig.6 に入力画像を， Fig.7 に処理画像を示す．Fig.7 では，エッジ抽出した画像上に，ボールとゴールのそれぞれを囲む枠を重ね書きしている．

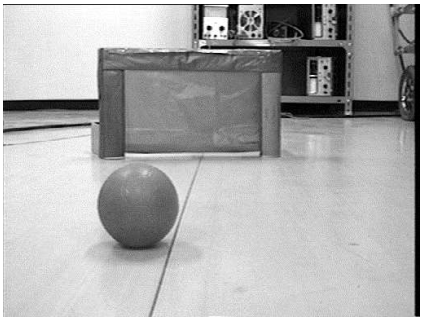

Fig.6 入力画像

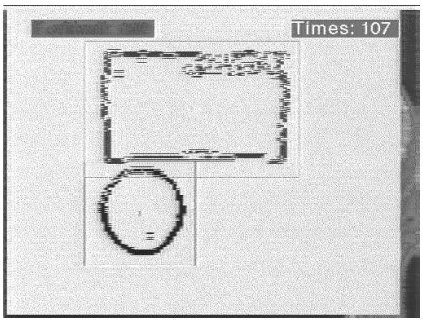

Fig.7 処理画像

4.3 状態識別

ロボットが知覚できるのは，カメラから得られるボールとゴールの二次元画像情報である．状態空間を

- ボールの状態については，

  大きさ (ボール半径): 大・中・小

  位置 (中心位置 X 座標): 左・中央・右

  見えていない: 左に消えた・右に消えた

- ゴールの状態については，

  大きさ (ゴールの高さ): 大・中・小

  位置 (中心位置 X 座標): 左・中央・右

  ゴールの見え方 (ゴールバーの傾き): 左上がり・水平・右上がり

  見えていない: 左に消えた・右に消えた

と設定した．これらの組み合わせによる 319 の状態で状態空間を構成する．このように状態を比較的粗く分けることによって，学習速度を上げることができる上に，画像処理の誤差を吸収できるという利点がある．

画像処理の結果として得られるボールとゴールの特徴点の画像上の座標から，ボールとゴールの画像上での位置，大きさおよびゴールバーの傾きを求め，状態を識別する．この時，画像上でボールは円であり，ゴールは二辺が画像の水平線に垂直な四辺形であると仮定して計算する．

4.4 行動選択

行動空間は，ロボットの左右輪それぞれの前進，停止，後進の組合せによる 9 通りの行動に設定した．ただし，両輪が停止する行動は，その行動によって状態変化が起きないので，選択しないように設定している．

学習によって，状態空間の中の 1 つの状態に対して，ロボットがとるべき最適な 1 つの行動が獲得されている．状態識別の結果から，あらかじめシミュレーション上で学習された学習結果を用いて，ロボットの行動を決定し，実行する．

5 実験結果

ロボットがシューティングに成功した回数を測定した．80 回の試行の内シュートに成功したのは

23 回であり，そのシュート率は約 30% であった．シュート率は初期位置に依存するが，本実験では，ロボットとゴールが約 1.5m から 2.5m 離れた位置から，ボールの位置とゴールの向きを適当に変えて行なった．

また，ロボットがシューティングに成功した時の状態識別と行動選択の結果の例を Table.1 に示す．この時のロボットのシューティング行動の様子を Fig.8 に示す．表中の time step は画像処理のステップ (1/30sec) を示し，state step は状態変化のステップを示す．状態は，ボールとゴールの画像上での位置を **L**(左)・**C**(中央)・**R**(右) で表し，大きさ (距離) を **F**(遠い)・**M**(中間)・**N** (近い) で表し，ゴールの向きについては **Lo**(左向き)・**Fo**(正面)・**Ro**(右向き) で表す．ボールまたはゴールが見えない状態は **D**isappeard と表す．各状態でロボットのとった行動を action 欄に左右輪のそれぞれの回転で **F**(前進)，**S**(停止)，**B**(後進) と表す．また，入力画像と処理画像を比較し，画像処理の失敗による状態識別のエラーには"*****" を付け，time step 毎のエラーの総数を右の列に示した．

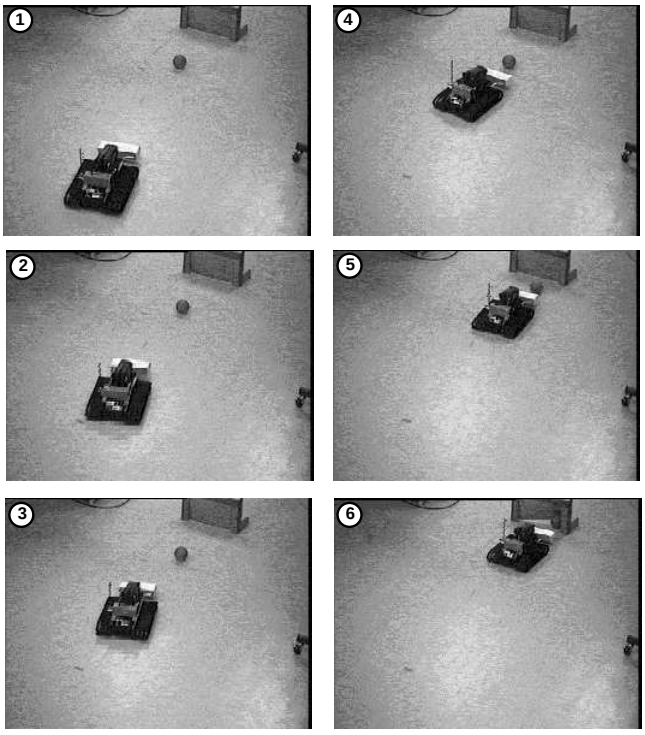

Fig.8 シューティング行動の様子

Table 1 状態識別と行動選択の結果

time step	state step	state: ball	state: goal	action: L	action: R	# err
1	1	(C,F)	(C,F,Fo)	F	F	
2	2	(R*,F)	(C,F,Fo)	F	F	1
3	3	(D*,D*)	(C,F,Ro*)	B	B	3
4	4	(C,F)	(C,F,Lo*)	B	S	1
5	5	(C,F)	(C,F,Fo)	F	F	
6		(C,F)	(C,F,Fo)	F	F	
7		(C,F)	(C,F,Fo)	F	F	
8		(C,F)	(C,F,Fo)	F	F	
9	6	(C,F)	(C,F,Ro*)	B	S	1
10	7	(C,F)	(C,F,Fo)	F	F	
11	8	(C,F)	(R,M,Fo)	F	F	
12	9	(R,F)	(R,M,Fo)	F	F	
13	10	(R,M*)	(R,F*,Lo*)	F	B	3
14	11	(L*,F)	(R,M,Ro*)	F	S	2
15	12	(L*,F)	(R,M,Fo)	F	S	1
16	13	(R,M)	(R,M,Fo)	S	B	
17	14	(C,M)	(C,M,Fo)	F	F	
18	15	(L,M)	(L,M,Fo)	S	F	
19	16	(L,N)	(L,M,Fo)	B	S	
20		(L,N)	(L,M,Fo)	B	S	
21	17	(L,M*)	(L,M,Fo)	S	F	1
22	18	(L,N)	(L,M,Fo)	B	S	
23		(L,N)	(L,M,Fo)	B	S	
24	19	(C,N)	(C,M,Fo)	F	B	
25	20	(C,M)	(C,M,Fo)	F	F	
26		(C,M)	(C,M,Fo)	F	F	
27	21	(C,M)	(C,N,Fo)	F	S	
28	22	(C,M)	(C,M*,Lo*)	F	S	2
29	23	(C,M)	(C,M*,Ro*)	S	B	2
30	24	(C,F)	(D,D,D)	F	S	

Ball: (position {**L**eft,**C**enter,**R**ight,**D**isappeard},
size {**N**ear,**M**iddle,**F**ar,**D**isappeard})
Goal: (position {**L**eft,**C**enter,**R**ight,**D**isappeard},
size {**N**ear,**M**iddle,**F**ar,**D**isappeard},
orientation {**L**eft-**o**riented,**F**ront-**o**riented,
Right-**o**riented,**D**isappeard})
error *****

total step	errer	errer step	errer /step×5	errer step /step
30	17	10	11.3%	33.3%

6 考察

6.1 ロボットのシュート率

実験結果に示したように，ロボットがシューティングに成功した割合は，約30%であった．同じような条件で行なったシミュレーションにおけるシュート率は，約70%である．この差の要因は，

(1) 画像上のノイズによる状態識別の間違い．

(2) シミュレーションと実機との不整合．

にある．

6.2 画像上のノイズによる状態識別の間違い

画像処理において，画像をUHF電波によって受信しているために生じるノイズにより，良好な画像処理結果が一定して得られない．それに伴いボールとゴールの状態識別が失敗する． Table.1の例では，総状態数(ステップ数×5)に対する総エラー数は11.3%で，総ステップ数に対するエラーステップ数は33.3%であった．実験時の電波状態によるが，平均的なエラーの割合は，それぞれ15%,40%程である．

Table.1の例には見られないが，ノイズがひどくボールもゴールも間違って状態を識別してしまう場合がある．このような状態が何ステップも続くとロボットがタスクを達成することは不可能だが，ほとんどの場合この状態は連続して起こることはなく，次の瞬間にほぼ正確な処理ができることによって，ロボットの動作を修正できる．

状態識別の間違いはほとんどボールまたはゴールの大きさ(距離)とゴールの傾きの間違いである．ボールやゴールの画像上での位置はそれらがある程度見えていればほぼ間違うことはないが，それらの大きさとゴールの傾きはその輪郭が明確でなければ間違えてしまうことが多い．従って，ボールとゴールの距離とゴールの向きをある程度正しく推定しなければシュートが困難な場合，例えば，ロボットに対してボールとゴールが別々の方向にある場合は，ノイズによって状態識別を間違い，シュートに失敗することがある．

これに対して，ロボットに対してボールとゴールがほぼ一直線に位置する状態や，ボールとゴールが同じ方向にありロボットが向きを変えるだけで上記の状態になる場合は，ボールとゴールの大きさやゴールの向きに関係なく，ロボットはボールとゴールが画像の中央に位置するように向きを変えつつ，直進することによって，シュートに成功する．Fig.8の例は，後者の場合であるといえる．

6.3 シミュレーションと実ロボットとの整合性

実験では，ワークステーション上でのシミュレーションにおける学習結果を用いている．従って，シミュレーションにおける仮定と実際のロボットのおかれた環境(物理世界)との整合性が問題になる．例えば，実際のボールの重心が偏心しているために，ボールの転がり方が不規則であり，ボールをロボットが意図した方向に運ぶことができない．また，ロボットの急な動作変化時に，クローラと床の間には大きな滑べりが生じるため，ロボットの行動が失敗する可能性がある．後者の場合，ロボットは状態が遷移するまでは同じ行動をとるので，状態が正確に識別されていれば，ほとんど問題はない．

7 おわりに

本稿では，強化学習の一種であるQ学習を，視覚を持った実ロボットへ応用することによって，強化学習を実システムに適用する際に生じる問題点を明らかにした．そして，製作した実システムによる実験を通して，本手法の有効性を検証した．

今後は，タスク達成のための最適な状態空間を得るために，ロボットが状態空間を自律的に分割し，再構成するような学習について検討したい．

参考文献

[1] J. H. Connel, S. Mahadevan, editors, “Robot Learning,” Kluwer Academic Publishers, 1993.

[2] J. H. Connel, S. Mahadevan, “Rapid task learning for real robot,” In *Robot Learning*, chapter 5, Kluwer Academic Publishers, 1993.

[3] S. D. Whitehead, “A complexity analysis of cooperative machanisms in reinforcement learning,” Proc. of the AAAI-91, pp.607-613, 1991.

[4] 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”，第4回ロボットシンポジウム予稿集, 1994.

[5] M. Inaba, “Remote-Brained Robotics: Interfacing AI with Real World Behaviors,” Proc. of the 6th Int. Symp on Robotics Research, Oct.1993.