方策勾配法を用いたサッカーエージェントの学習：パス選択

五十嵐 治一[+]，福岡 仁志[+]，佐野 直人[+]，石原 聖司[++]

[+]芝浦工業大学工学部情報工学科，[++]近畿大学工学部電子情報工学科

人工知能の分野ではマルチエージェントシステムにおける協調行動の研究のために，ロボットサッカーの競技会である RoboCup が提唱されている．本研究ではヒューリスティクスを組み込んだ方策関数を考え，その中の重み係数を強化学習法の一種である方策勾配法を用いて学習することで，精度の高いパスの実現を目指す．この学習機能をプログラムを公開している２つの代表的なチーム（UvA Trilearn2003 と HELIOS)へ移植して学習実験を行い，チーム力が向上することを確かめた．

Learning of Soccer Player Agents Using a Policy Gradient Method: Pass Selection

Harukazu IGARASHI[+], Hitoshi FUKUOKA[+], Naoto SANO[+], Seji ISHIHARA[++]

[+]Shibaura Institute of Technology, [++]Kinki University

This research develops a learning method for the pass selection problem of midfielders in RoboCup Soccer Simulation games. A policy gradient method is applied as a learning method to solve this problem because it can easily represent the various heuristics of pass selection in a policy function. We implement the learning function in the midfielders' programs of two well-known teams, UvA Trilearn 2003 and HELIOS. Experimental results show that our method effectively achieves clever pass selection by midfielders in full games. Moreover, in this method's framework, dribbling is learned as a pass technique, in essence to and from the passer itself. It is also shown that the improvement in pass selection by our learning helps to make a team much stronger.

1. はじめに

近年，人工知能の分野ではマルチエージェント環境下における協調行動の学習が研究されている[1][2][3]．この研究の題材としてロボットサッカーの競技会であるRoboCupが提唱されている[4][5]．このRoboCupの中の一部門であるサッカーシミュレーション・リーグ[6][7]では，実環境でロボットを動かす難しさから解放されるため，複数エージェントによる協調的な行動に研究の焦点を当てることができる[8]．

一般に，マルチエージェント学習には，状態空間の爆発問題，同時学習問題[9]，不完全知覚問題[10]，報酬割り当て問題[2]等のマルチエージェント系特有の問題があるとされている[3]．本研究で取り扱うRoboCupサッカーシミュレーション2Dリーグでは，不完全な知覚情報とある程度の通信が許されたマルチエージェント系をシミュレートしていると考えられるが，本研究でもこれら４つの問題に対応する．

その中でも，「状態空間の爆発問題」が第一の大きな問題と考えられる．2Dリーグでは，各プレーヤーの視野内の物体への相対角度と相対距離とが視覚情報としてサーバから定期的に与えられ，各プレーヤーはこれを基に状況判断と行動決定を行い，決定した行動コマンド（kick, dash, turn など）をサーバへ送信すると言うサイクルの繰り返しでゲームが進められる．しかし，各エージェントに入力される情報の状態空間はそのままでは大きすぎて，かつ，不完全・不確実である．本研究では，エージェントの行動決定（方策）の強化学習をテーマとしているが，毎時刻のすべての行動決定を学習するのではなく，局面とエージェント（複数）とを限定し，そこでのマクロ的行動の決定方法を学習することにより状態空間の爆発問題を回避する．具体的には，ゲーム中のミッド・フィルダー（MF）3～4人の味方プレーヤーへのパス選択問題を取り扱った．

本論文の構成は以下のとおりである．第2章では，RoboCupの公式サッカーシミュレータと，本研究で用いたチーム，本リーグでこれまでに行われてきた協調行動の研究例について簡単にまとめた．第3章では，本研究の対象であるパス選択問題について述べ，その問題における学習の目的と，その目的の実現に向けた報酬の設計法について説明する．第4章では，本研究で用いた方策勾配法を説明し，パス選択に用いたヒューリスティクス（先験的知識）の詳細を記した．続く2つの章では，2つのチームに本研究のパス選択法を移植し，自己対戦を通して学習させた実験について述べた．第7章には本研究のまとめと今後の展開について記した．

2. サッカーシミュレータ

2.1 サーカーフィールド

本研究ではシミュレーション環境としてRoboCupの公式シミュレータであるSoccer Server ver. 12.1.X [4]を用いた(ただし，5章ではX=3，6章ではX=4)．フィールドはセンターサークルの中央を原点として，105[m]×68[m]の大きさである(Fig.1)．x座標は自陣から敵陣への方向を正に，y座標は自陣から敵陣を見て右方向を正に取る．Fig.1には左が自陣である例を示してある．

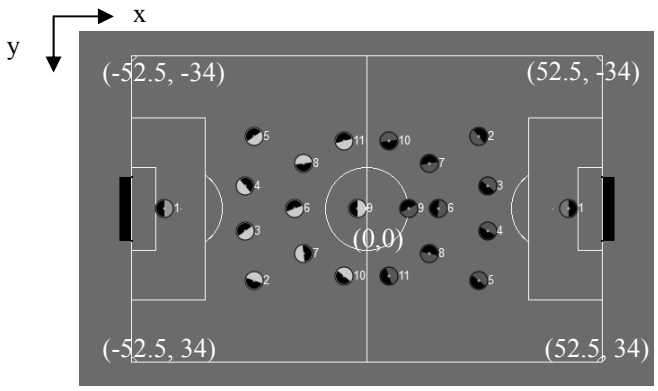

Fig.1 仮想サッカーフィールド

2.2 サッカーサーバーとサッカーエージェント

Soccer Serverはクライアント・サーバ方式をとっており，ユーザは11人分のプレーヤープログラムを作成し，クライアントとしてサーバにUDP/IPで接続する(Fig.2)．サーバは物体の動きをシミュレートして，クライアントに視覚・聴覚情報等を送信し，クライアントはサーバへkick/dash/turnなどの基本コマンドを送信してプレーヤーを動かす．しかし，クライアント同士はsay/hearコマンドなどを用いたサーバを介した通信はできるが，直接通信はできない．つまり，チームは自律分散制御であり，その中で協調行動が必要とされる．

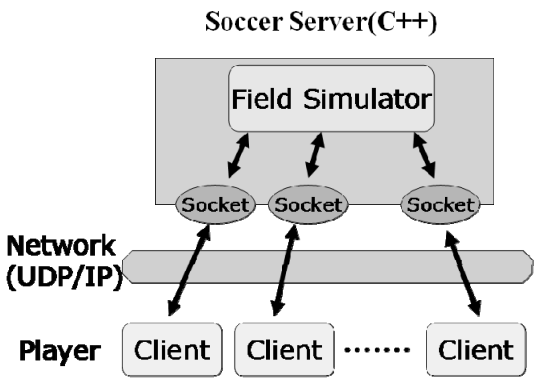

Fig.2 サッカーサーバーとクライアントプログラム

2.3 本研究で用いたチーム

本研究ではアムステルダム大学が公開しているUvA Trilearn2003チーム(以下，UvAと表記) [11]と，個人が公開しているHELIOS ver.1.0 [12]というチームをそれぞれベースとして我々が開発した２つのチームを使用する．

UvAは2003年の世界チャンピオンチームであり，ボールやプレーヤーの情報の取得，パスする相手の探索など，多くの基本的な機能がライブラリとして提供されている．しかし，公開されているプログラムには戦術的な行動は何も書かれていない．各プレーヤーは，ボールを追って行き，ゴールに向かってボールを蹴るだけである．そのため，開発者は状況に応じたドリブルやパス，インターセプトなど，実戦に必要なプレーを具体的に実装する必要がある．

一方，HELIOSは日本のチームの中では最強（2008年世界大会3位）であり，国内では多くのチームがベースとして採用している．UvAと同様あるいはそれ以上に，幾何計算，観測データや内部状態の管理，プレイヤー動作に関する多彩なライブラリが用意されている．しかし，パスやドリブル，タックルなどプレーヤーの行動プログラムも実戦用のものがそのまま公開されているため，プログラムに手を加えるとかえって弱くなってしまうという傾向があった．したがって，プレーヤーに学習機能を組み込んでチームを強化することは難しいと考えられてきた．本研究では，上記の2種類のチームのMFに対してパス選択時の学習機能を組み込んだ．

2.4 シミュレーションリーグにおける協調行動の研究例

マルチエージェント系のエージェントの行動学習には強化学習[3][13][14]がよく用いられる．強化学習を用いたシミュレーションリーグにおける研究としては次のような例がある．例えば，動的なホームポジションを決定するための方策をニューラルネットワークモデルで表現し，その重みパラメータをKimuraらの確率的傾斜法（方策勾配法の一種）で求めたAndou[15]の研究や，3v2（攻撃側3名と守備側2名の対戦の意．以下同様）における少人数戦術タスクをProfit Sharing法的な強化学習により学習させた熊田らの研究[16]，2v1から7v8までの対戦において個人スキルの行動列（例，パス・インターセプト，ドリブル，nキック）の生成と選択の各問題をTD法により学習するRiedmillerらの研究[17]，3v2におけるKeepaway（ボール保持）[18]と4v5でのhalf field offence（ハーフサイズの競技フィールドを用いた攻撃タスク）における攻撃側エージェントのマクロ行動（例，ボール保持，パス，ドリブル，シュート）の選択をSarsa(λ)法により学習するStoneらの研究[19]などがある．

1章で述べた「状態空間の爆発問題」は，状態空間を適切に設計すれば，ある程度，状態空間の爆発を抑えることができる[3]．上記のシミュレーションリーグのエージェント学習問題においては，状態空間の削減のためには，競技フィールドを粗いセルに分割する方法[15][16]，プレーヤーとボールの位置座標を入力とするニューラルネットワークモデルによる状態価値関数の近似[17]，タイルコーディングなどによる行動価値関数（Q関数）の近似[18]，状態の特徴量空間を用いる[15][19]などの方法が考案されている．また，行動空間の削減のためにも，dashコマンドやkickコマンドなどの基本コマンドそのものを選択するのではなく，個人スキル[17]やマクロ行動[18]などの少し抽象化されたレベルの行動を定義して用いるという工夫がなされてきた．

さらに，11v11のフルゲームを対象タスクとするのではなく，ボールを持っていないときのホームポジションの決定問題[15]や，2v2のパス＆シュート問題[20][21]，4×3セルから構成される矩形の小領域内においてその矩形領域の決められた一辺にボールを通過させる問題[16]，3v2のボール保持問題(Keepaway)[18]，2v1～7v8のように人数を制限した対戦問題[17][19]など，問題自体を小さなものに限定することにより「状態空間の爆発問題」を回避している．

本研究ではパス選択というマクロ的行動の学習に問題を限定することにより，11v11のフルゲームでの学習を可能にするという方針を取った．

3. パス選択問題

3.1 学習の目標

本研究では，サッカーにおけるチーム力の強化のためには，MFが局面に応じて味方プレーヤーへ適切なパスを出すことにより，自チームのボール保持時間が長くすることが極めて有効であると考えた．また，適切なパス相手がいない状況では，自らがドリブルを行って保持時間を伸ばすことが望ましいであろう．そこで，次節に述べる報酬を強化学習において用いることにした．

3.2 報酬の設計

本研究の学習ではエピソードごとに行う．そこは次の2種類のエピソードを定義する．第１の定義は，自チームがボールを保持してから敵チームにボールを奪われるまでの状態・行動列を１エピソードとすることである．第２の定義は，各MFがボールを保持してから敵チームにボールを保持されるまでの状態・行動列を1エピソードとすることである．ただし，「ボールを保持する」とは，シミュレーションの5サイクルステップ以上の時間，ボールに最近接である状況とここでは定義している．次に，報酬rはエピソード長Lの逆数の逆符号，r≡-1/L (<0)とする．したがって，自チームまたは各MFがボールを得てからの保持時間が短いほど大きな罰が与えられる．

なお，上記のエピソードの定義から，第１のエピソードの定義の場合は，報酬は全学習エージェントに共通な値が与えられ，第２の定義の場合には，各学習エージェントごとに異なる値が与えられる．学習（パラメータの更新）は各エピソード終了時に行われるが，エピソード中にボールに触れなかった学習エージェントのパラメータは更新されない．どの定義が適切かどうかは一概に言えないが，UvAベースのチームを用いた学習実験では第１の定義を，HELIOSベースのチームを用いた学習実験では第２の定義を用いた．

4. 学習方式

4.1 方策勾配法

本研究では強化学習の一種である方策勾配法を用いて学習を行う．方策勾配法とは，報酬の期待値が最大（極大）になるように方策中のパラメータを更新する学習法である．このときの最大化の手段として確率的勾配法を用いる．方策勾配法は数学的な基礎がはっきりしており，理論的に取り扱いやすい．また，方策として if-then 型のルールや，ポテンシャルなどの様々な関数が利用できるので，方策への知識表現が容易であるという長所がある[22]．元々は，Williams により提案された手法[23]であるが，Kimura らにより確率的傾斜法[24]として部分観測マルコフ決定過程（POMDP: Partially observable Markov Decision Process）へ適用されている．さらに，一般的な非マルコフ決定過程においても方策勾配法が適用可能であることが示されており[25]，実際に追跡問題やカーリングゲームに適用されている[22][26]．近年，マルコフ決定過程(MDP: Markov Decision Process)においても方策勾配法は見直されており，理論面と応用面の両方において研究が進められている[27][28][29][30]．

4.2 目的関数と方策

プレーヤーλが状態(＝λの周囲の局面) s_λ においてボールをキックできる時に，どの味方プレーヤーへパスを出すかという行動決定を考える．このときのパスを出す相手を「パス先」と呼ぶ．プレーヤーλがパス先を選択する行動 a_λ を評価するのに有用なヒューリスティクスを $U_i(a;s)$ (i=1, 2, …, N_U)を用意し，これらの線形和で表現される次の目的関数 $E(a_\lambda; s_\lambda, \{\omega^\lambda_i\})$ を考える．

$$E_\lambda\left(a_\lambda;s_\lambda,\left\{\omega_j^\lambda\right\}\right)=-\sum_j \omega_j^\lambda \cdot U_j\left(a_\lambda;s_\lambda\right) \quad (1)$$

具体的なヒューリスティクスの形は 4.5 で述べるが，$0 \leqq U_i \leqq U_{max}$ というようにある程度正規化された関数として設計する．重み係数については $\omega_i \geqq 0$ と仮定する．したがって，学習時にはどの重みも $\omega_i \geqq 0$ という範囲で勾配法を用いる．これは上記のヒューリスティクスが正しいという前提の下で，その重要度を学習しようと考えているからである．もし，重みが負であれば，それは逆の内容の知識を意味することになる．しかし，本論文ではヒューリスティクスの真偽を学習することまでは意図していないので，重みは非負であるという制限を設けた．

上で定義した目的関数を用いて，プレーヤーλの方策を次の Boltzmann 分布関数で与える．

$$\pi_\lambda\left(a_\lambda;s_\lambda\right) \equiv \frac{e^{-E_\lambda\left(a_\lambda;s_\lambda\right)/T_\lambda}}{\sum_a e^{-E_\lambda\left(a;s_\lambda\right)/T_\lambda}} \qquad (2)$$

Boltzmann 分布は目的関数の期待値$\langle E_\lambda \rangle$を一定に固定した時にエントロピー（行動選択の不確定性）を最大にする分布である．したがって，この分布に従って行動を選択すれば，$\langle E_\lambda \rangle$の大きさ，すなわち，行動の質が期待値の意味で保証される．物理系を考えるとわかるように，$\langle E_\lambda \rangle$は温度パラメータ T を大きくとるほど大きくなる．また，T を大きくするほどランダムに行動を選択するようになり，小さくするほど最も大きい価値の行動を選択しやすくなる．特に，T→0 の極限では決定論的な行動決定となる．

4.3 自律分散的な行動決定

本研究では，協調行動(チームプレイ)の実現を目的としている．したがって，報酬 r を特定のエージェントの特定の状態や行動に与えるのではなく，チーム全体として良い状態や行動が生じたときに与えたい．そこで，あるエピソードσにおいて生起した一連の自チーム全体の状態・行動に対して報酬 $r=r(\sigma)$を与えるものとする．また，報酬を与えるタイミングは，各エピソードの終了時とし，各エージェントに同一の報酬を与える．

しかし，チーム全体の状態 s や行動 a を基にした方策$\pi(a;s)$を取り扱うのは，「状態空間の爆発」を招くので以下の近似を行う[22][31]．

$$\pi(a;s) \approx \prod_\lambda \pi_\lambda\left(a_\lambda;s_\lambda,\left\{\omega_j^\lambda\right\}\right) \qquad (3)$$

右辺のπ_λはプレーヤーλの方策関数であるが，本研究では(1)と(2)で表される関数を用いる．なお，(3)の近似は，エージェント間の行動選択の相関を無視することを表している．しかし，各エージェントは行動選択時に味方エージェントの存在（位置情報）までを無視しているわけではなく，状態 s_λに含まれている．したがって，このような近似を行っても協調行動を学習することはある程度までは可能と考えられる．

4.4 学習則

(3)の近似を用いると，報酬の期待値の勾配は次のように表すことができる[22][31]．

$$\frac{\partial E\left[r(\sigma)\right]}{\partial \omega_j^\lambda} = E\left[r(\sigma)\cdot\sum_{t=0}^{L-1}\frac{\partial}{\partial \omega_j^\lambda}\ln\pi\left(a(t),s(t)\right)\right] \qquad (4)$$

$$\approx E\left[r(\sigma)\cdot\sum_{t=0}^{L-1}\frac{\partial}{\partial \omega_j^\lambda}\ln\pi_\lambda\left(a_\lambda(t);s_\lambda(t)\right)\right] \qquad (5)$$

(5)に対して確率的勾配法の考えを用いると以下の学習則を得ることができる[22][31]．

$$\Delta\omega_j^\lambda = \varepsilon\cdot r(\sigma)\sum_{t=0}^{L(\sigma)-1} e_{\omega_j^\lambda}(t) \qquad (6)$$

ただし，

$$e_{\omega_j^\lambda}(t) \equiv \frac{\partial}{\partial \omega_j^\lambda}\ln\pi_\lambda\left(a_\lambda(t);s_\lambda(t)\right) \qquad (7)$$

は，離散時刻 t におけるパラメータ$\omega^\lambda{}_j$の特徴的適正度[23]であり，$a_\lambda(t)$と $s_\lambda(t)$は離散時刻 t におけるエージェントλの行動と状態である．(1)，(2)を(7)の右辺へ代入すると，特徴的適正度は，

$$e_{\omega_j^\lambda}(t) = \frac{1}{T}\left[U_j\left(a_\lambda(t)\right) - \sum_{a_\lambda} U_j\left(a_\lambda\right)\pi_\lambda\left(a_\lambda;s_\lambda(t),\left\{\omega_j^\lambda\right\}\right)\right] \qquad (8)$$

と表わされる．したがって，報酬はチーム全体のエピソードを評価して各プレーヤー共通に与えられるが，それ以外はプレーヤーごとの実際に生起した状態と選択した行動から更新量を計算しており，自律分散的な学習則であると言える．

4.5 パス選択に用いたヒューリスティクス

本研究ではパス相手（以下，パス先）の決定において，ヒューリスティクスとして次の U_1~U_5 $(0 \leqq U_i \leqq 10)$を用いる：(i)U_1=パスコースにおける敵の有無(Fig. 3a)，(ii) U_2=パス先と最も近い敵との距離(Fig.3b)，(iii) U_3=パス先周辺の敵の数(Fig. 3c)，(iv)U_4=パス先とゴールとの距離(Fig.3d)，(v)U_5=パス先の信頼度(U_5)．各ヒューリスティクスの定義の詳細を以下に記す．

(i)U_1=パスコースにおける敵の有無：パスコースが空いているほど価値が高くなる．

$$U_1 = \begin{cases} 10.0\ (if\ \ diff > 30\ or\ OppDist > AgentDist \\ \qquad or\ \ パス先が自分自身) \\ 9.0\ (if\ \ 20 < diff \le 30 \,\&\, OppDist < AgentDist) \\ 7.0\ (if\ \ 10 < diff \le 20 \,\&\, OppDist < AgentDist) \\ 4.0\ (if\ \ diff \le 10 \,\&\, OppDist < AgentDist) \\ 2.0 \quad (else) \end{cases} \tag{9}$$

ただし，*diff*=パス先と敵との角度差，*OppDist*=ボールと敵との距離，*AgentDist*=ボールとパス先との距離である．３つのどれかの値が取得不可であるときには 2.0 を返す.

(ii) U_2=パス先と最も近い敵との距離 : パス先と敵との距離が離れているほど価値が高い.

$$U_2 = \begin{cases} 10-(-\min\_dist+30.0)/5.0\ (if\ \min\_dist < 30) \\ 10 \qquad\qquad\qquad\qquad (if\ \min\_dist \ge 30) \\ 2.0 \qquad\qquad\qquad\quad (else) \end{cases} \tag{10}$$

ただし，*min_dist*=パス先と最も近い敵との距離であり，敵の位置がわからない場合は 2.0 を返す.

(iii) U_3=パス先周辺の敵の数：パス先周辺（半径 10m）に敵が少ない方が価値の高い.

$$U_3 = \begin{cases} EnemyNum\ (if\ \ EnemyNum \le 10.0) \\ 0.0\ (else) \end{cases} \tag{11}$$

ただし，*EnemyNum*=パス先周辺の敵の数である.

(iv) U_4=パス先とゴールとの距離 : パス先がゴールに近いほど価値が高い.

$$U_4 = \begin{cases} 10-goal\_dist/10.0\ (if\ \ dist \le 40.0) \\ 2.0 \qquad\quad (else) \end{cases} \tag{12}$$

ただし，goal_dist=パス先と敵ゴールの距離である.

(v) U_5=パス先の信頼度 : パス先の信頼性が高いほど価値が高い.

$$U_5 = \begin{cases} 9.0 \quad (if\ パス先が自分自身) \\ 10-(-50.0*confidence+50.0) \quad (else) \end{cases} \tag{13}$$

ただし，*confidence*=パス先の信頼度であり，最後にパス先を見た時間 *LastSeeCycle* と現在の時間 *CurrentCycle* との差から次のように算出する.

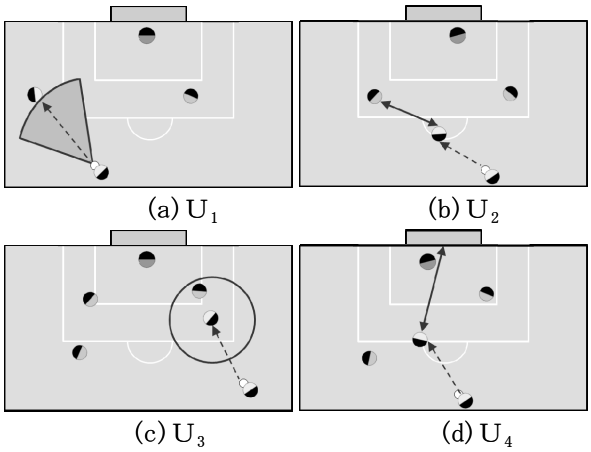

Fig.3 パス選択問題に用いたヒューリスティクス

$$confidence \equiv 1 - \frac{CurrentCycle - LastSeeCycle}{100} \tag{14}$$

ヒューリスティクスのU_1～U_3は，敵に奪われない安全なパス先を選ぶための知識であり，U_4は攻撃に結びつく前方へのパスを好むという知識であり，U_5は不確実な情報しか得ることができない環境下において，なるべく確信度の高い情報を信用するという知識を表している.

5. UvA Trilearn Base チームでの学習実験

(1)の目的関数と 4.5 で述べたヒューリスティクスを用いて学習実験を行う．学習実験では実際の試合時間である 6000 サイクル(10 分)を 1 試合とし，未学習の自分自身のチームを相手にして 50 試合，ゲームを行う．学習実験後に評価実験として，未学習の自分自身を相手に 30 試合，ゲームを行う．学習実験中の 50 エピソードの報酬の平均値 $\overline{r_{50}}$ とヒューリスティクスの重み$\{\omega_i\}$の変化とを，それぞれ Fig.4 と Fig.5 に示す．なお，UvA ベースの本チームの学習エージェント(MF)は 4 人存在するが，重みの推移の傾向において殆ど差がなかったため，Fig.5 ではそのうちの一人(MF_8)のデータだけを掲載している．なお，学習実験においては，重みの初期値は全て 0，T=10.0，ε=0.1 とし，評価実験においては T=0.1 と設定した.

Fig.5 から，学習を行うことによって U_1 と U_5 のヒューリスティクスが重要視されたことが分かる．これは，パスコースに敵がいないパス先(U_1) であって，かつ，そのパス先の位置情報の信頼度が高い場合(U_5)を優先して選ぶことを学習したと言える．また，重みの比率が一定であることから，パラメータ空間の強いアトラクタへ学習が収束していると推測される.

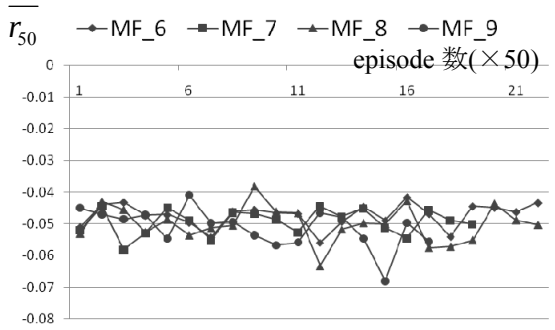

Fig.4 報酬の平均値 $\overline{r_{50}}$ (UvA ベースのチーム)

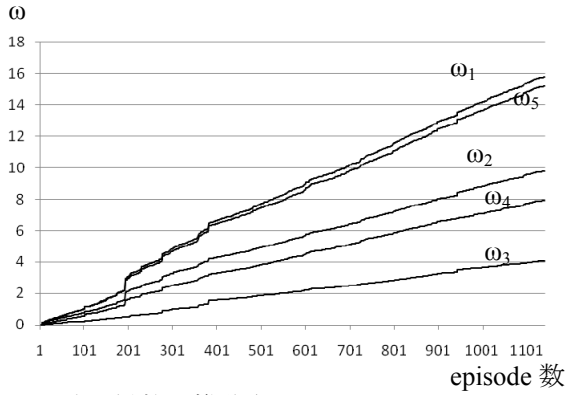

Fig.5 重み係数の推移(UvA ベースのチームの MF_8 の例)

一方，Fig.4 を見ると，報酬の平均値がそれほど上昇していないように見える．これは学習が進むにつれてチームが攻撃的になったのが原因である．攻撃的で敵陣の深い領域でプレーすることが多くなるほど，チームとしてのボールの保持時間そのものは短くなる．実際，学習後の試合を見るとチームは攻撃的になっており，勝率も確実に増加している．学習が進むにつれて攻撃的になるのは，パス先の位置情報の信頼度 U_5 を重要視するようになったからである．本実験で採用したチームの「ボールを持っているプレーヤーは主に敵ゴール方向を注視する」という特性を持っている（というように作成してある）．この特性により，信頼度の低いバックやサイドへのパスは稀となり，チャンスがあれば信頼度の高い前方のプレーヤーへとパスを行うようになったと解釈される．

また，評価実験 30 試合の勝敗は，未学習チーム同士の対戦における勝敗は一方の 17 勝 13 敗（総得点 32，総失点 26）であったが，学習チーム対未学習チームでは， 23 勝 7 敗 (総得点 : 40，総失点 : 13）であった．学習後のチームが積極的に攻撃を行って得点を挙げていることがわかる．

6. Helios チームでの学習実験

HELIOS における MF は 3 人である．HELIOS では自分の周りに敵がいればパスを行い，いなければドリブルを行う．このパス動作を本方式によるプログラムに置き換えた．また，HELIOS では信頼度の最も低い味方プレーヤーを積極的に見ることにより，常に周囲の味方プレーヤーの位置情報の信頼度を一定以上保持しようとする．したがって，後方の味方プレーヤーの信頼度も保たれており，UvA ベースのチームに比べて横方向や後方へのパスも出しやすい．一方，(9)の U_1 は，敵がパスコース上に近い位置にいてもリスクを覚悟してパスを出すという攻撃的な知識を表している．したがって，(9)をそのまま HELIOS へ用いると，その副作用として危険なバックパスや横パスを出す可能性が高くなる．インターセプトの巧みなチームが相手ではなおさらである．そこで，安全性の高いパスを出すための知識として，(9)の U_1 に換えて，次の U_1'を用いた．

$$U_1' = \begin{cases} 10.0\ (if\ \ diff > 30\ \ \&\ \ OppDist > AgentDist \\ \qquad or\ \ パス先が自分自身) \\ 9.0\ (if\ \ 20 < diff \le 30 \,\&\, OppDist > AgentDist) \\ 7.0\ (if\ \ 10 < diff \le 20 \,\&\, OppDist > AgentDist) \\ 4.0\ (if\ \ diff \le 10 \,\&\, OppDist > AgentDist) \\ 2.0 \quad (else) \end{cases} \tag{15}$$

(15)では,敵がパスコースから角度が 30 度以上，かつ，距離についても敵がパス先より遠い場合に限り最大値 10.0 を与え，パス先よりも近い位置に敵がいれば方向によらず最小値 2.0 を与える．これは(9)よりパスの安全度を高めた知識である．

その他の諸条件は 5 章の実験と同一に設定した．学習実験中における $\overline{r_{50}}$ と，重みの推移(MF_6 の例)を，それぞれ Fig.6 と Fig.7 に示す．

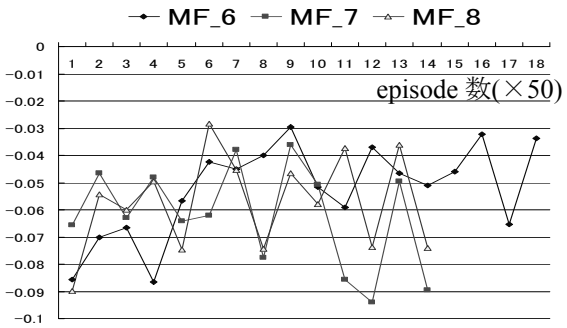

Fig.6 報酬の平均値 $\overline{r_{50}}$ (HELIOS ベースのチーム)

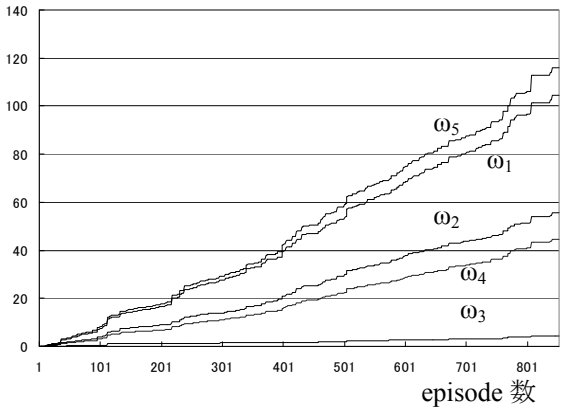

Fig.7 重み係数の推移(HELIOS ベースのチームの MF_6 の例)

Fig.7 での重みの推移は， UvA ベースのチームと同じ傾向が見られる．Fig.6 でも同じく，各 MF の報酬は学習回数とともに大きく増加しているとは言い難い．また，評価実験の結果は，未学習チーム同士の対戦における勝敗は一方の 1 勝 1 敗 28 分け（総得点 1，総失点 1）であったが，学習チーム対未学習チームでは， 3 勝 0 敗 27 分（総得点 3，総失点 0）と若干勝ち数が増えた．このように，報酬の平均値や勝敗・得失点だけからは学習の効果がわかりにくい．しかし，実際の試合では学習後のチームは敵陣によく攻め込んでおり，有利に試合を進めている．学習チームと未学習チームの試合でのボールのトレースの例を Fig.8 に示す．学習後のチームが敵陣に攻め込んでいるのがわかる．

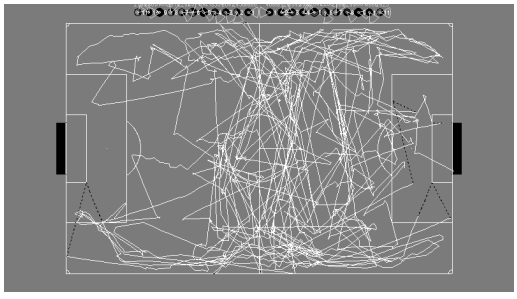

Fig.8 学習後のチーム（左のサイド）と未学習チームとの試合におけるボールのトレースの例

7. まとめ

RoboCup サッカーシミュレーション２Ｄリーグで使用されている公式サッカーシミュレータを用いて，マルチーエージェント系における強化学習の研究を行った．具体的には，フルゲームにおける MF プレーヤーのパス選択問題を，強化学習の一手法である方策勾配法を用いて学習する方式を提案した．報酬は自チームや各 MF のボール保持時間が増加するように与えた．この方式を UvA と HELIOS という 2 つの代表的なチームへ移植し，有効性を検証した．結果はともに有効なパスが増え，敵陣で攻撃を行う機会が増えることが観測された．特に前者の例では，未学習チームと対戦させると大きく勝ち越すようになった．

今後はパスを出すタスクだけではなく，それ以外のタスクにおける協調行動も学習させたい．例えば，深いサイドからのセンタリング，FW のポストプレー，複数人による守備的プレスやパス・インターセプト，敵プレーヤーのマーク，オフサイド・トラップ，ポジション取りなど，複数人の味方プレーヤーが関与する戦術レベルの各種攻撃／守備タスクをタスクごとに学習する．これらを学習後，試合中の場面に合わせて，タスクごとに学習した方策を切り替えることにより，協調行動の連鎖を実現できるのではないかと考えている．これらのタスクごとの方策の切り替えや，複数のタスクの組み合わせは，上記のような少数エージェントが関与する「戦術レベル」での行動ではなく，それよりも上のチーム全体が関与する「戦略レベル」での協調行動である．我々は本論文で述べた学習法の枠組みで，その上位レベルの方策も学習できるのではないかと期待している．さらに，対象としている学習問題を階層的にサブタスクに分割し，サブタスクごとに学習しておき，それらを統合する方策をさらに学習するという 2 段階方式の強化学習の一般的研究[32]へと繋げていきたいと考えている．

参考文献

[1] Weiss, G. and Sen, S. (Eds.): Adaption and Learning in Multi-agent Systems, Springer-Verlag, Germany(1996).
[2] Weiss, G. (Ed.): Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence , Sen, S. and Weiss, G.: Learn-ing in Multiagent System, pp. 259-208, The MIT Press(1999).
[3] 高玉圭樹：マルチエージェント学習－相互作用の謎に迫る－，コロナ社(2003).
[4] <http://www.robocup.org/02.html>
[5] 北野宏明，大沢英一，松原仁：なぜ今，RoboCup なのか？，bit, Vol.28, No.5, pp.22-27(1996).
[6] <http://sserver.sourceforge.net/>
[7] 野田五＋樹，國吉康夫：シミュレーション部門と Soccer Server，bit， Vol.28，No.5，pp.28-36(1996)
[8] 野田五＋樹：シミュレーションリーグとインフラ技術の技術的課題と展望，日本ロボット学会誌，Vol.20，No.1，pp.7-10(2002)
[9] Arai, S. and Miyazaki, K.: Learning Robust

Policies for Uncertain and Stochastic Multi-agent Domains, 7th International Symposium on Artificial Life and Robotics, pp.179-182(2002).
[10] Lovejoy, W.S.: A survey of algorithmic methods for partially observed Markov decision processes, Annals of Operations Research, vol.28, pp.47-66(1991).
[11] Kok, J.R.: UvA Trilearn, <http://remote.science.uva.nl/~jellekok/robocup/>, (accessed 2007-06-26)
[12] 秋山英久：ロボカップサッカーシミュレーション２Ｄリーグ必勝ガイド，秀和システム (2006).
[13] Sutton, R.S. and Barto, A.G.: Reinforcement Learning, The MIT Press, Massachusetts (1998).
[14] Kaelbling, L.P., Littman, M.L., and Moore, A.W.: Reinforcement Learning: A Survey, Journal of Aritificial Intelligence Research, Vol.4, pp.237-285(1996).
[15] H.Kitano (Ed.): RoboCup-97: Robot Soccer World Cup I , Andou, T.: Refinement of Soccer Agents' Positions Using Reinforcement Learning, pp.373-388, Springer-Verlag, Berlin,(1998).
[16] 熊田陽一郎，植田一博：予測能力を持つサッカーエージェントによる協調戦術の獲得，人工知能学会論文誌，Vol.16，No.1，pp.120-127(2001).
[17] Riedmiller, M. and Gabel, T.: On Experiences in a Complex and Competitive Gaming Domain -Reinforcement Learning Meets RoboCup-, Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games (CIG2007), pp.17-23(2007).
[18] Bredenfeld, A., Jacoff, A., Noda, I. , and Takahashi, Y. (Eds.): RoboCup 2005:Robot Soccer World Cup IX, Stone, P., Kuhlmann, G., Taylor, M.E., and Liu, Y.: Keepaway Soccer:From Machine Learning Testbed to Benchmark, pp.93-105, Springer-Verlag, New York,(2006).
[19] Lakemeyer, Sklar, Sorrenti, Takahashi (Eds.): RoboCup-2006:Robot Soccer World Cup X , Kalyanakrishnan, S., Liu, Y., and Stone, P. : Half Field Offence in RoboCup Soccer - A Multiagent Reinforcement Learning Case Study, Springer-Verlag(2007).
[20] Kitano, H. (Ed.): RoboCup-97:Robot Soccer World Cup I, Ohta, M.: Learning Cooperative Behaviors in RoboCup Agents, pp.412-419, Springer-Verlag, Berlin(1998).
[21] Igarashi, H., Nakamura, K., and Ishihara, S.: Learning of Soccer Player Agents Using a Policy Gradient Method: Coordination between Kicker and Receiver during Free Kicks, Pro. of 2008 International Joint Conference on Neural Networks (IJCNN 2008), Paper No. NN0040, pp.46-52(2008).
[22] 石原聖司，五十嵐治一：マルチエージェント系における行動学習への方策こう配法の適用-追跡問題-，電子情報通信学会論文誌 D-I，Vol.J87-D1，No.3，pp.390-397(2004).
[23] Williams, R.J. : Simple Statistical Gradient- Following Algorithms for Connectionist Reinforcement Learning, Machine Learning, vol.8, pp.229-256(1992).
[24] 木村元，山村雅幸，小林重信：部分観測マルコフ決定過程下での強化学習-確率的傾斜法による接近，人工知能学会誌,Vol.11,No.5, pp.761-768(1996).
[25] 五十嵐治一，石原聖司，木村昌臣：非マルコフ決定過程における強化学習-特徴的適正度の統計的性質-,電子情報通信学会論文誌 D, Vol. J90-D, No.9, pp.2271-2280(2007).
[26]五十嵐治一，石原聖司，木村昌臣：非マルコフ決定過程における方策勾配法の一考察：カーリングの事例，電子情報通信学会技術報告 NC2006-148，Vol.106，No.588，pp.179-184 (2007).
[27] Sutton, R.S., McAllester, D., Singh, S. and Mansour, Y. : Policy Gradient Methods for Reinforcement Learning with Function Approximation, Proc. of Advances in Neural Information Processing Systems 12 (NIPS' 99), pp.1057-1063(2000).
[28] Konda, V. R. and Tsitsiklis, J. N.: Actor-Critic Algorithms, Proc. of Advances in Neural Information Processing Systems 12 (NIPS' 99),pp.1008-1014 (2000).
[29] Kakade, S.: A natural policy gradient, Proc. of Advances in Neural Information Processing Systems 14 (NIPS' 01), pp.1531-1538(2002).
[30] Peters, J., and Schaal, S.: Policy Gradient Methods for Robotics, Proc.of the IEEE International Conference on Intelligent Robotics Systems (IROS 2006).
[31] Peshkin, L., Kim, K.E., Meuleau, N. and Kaelbling, L.P.: Learning to cooperative via policy search, Proc. of 16th Conference on Uncertainty in Artificial Intelligence (UAI2000), pp.489-496(2000).
[32]五十嵐治一，石原聖司：方策勾配法における状態空間の階層化の一考察，人工知能学会第 27 回 SIG-Challenge 研究会資料，pp.7-12 (2008).