コンピュータビジョンと
イメージメディア 103－7
（1997．1．23）

マルチエージェント環境における部分空間同定法を用いたエージェントの判別と行動理解
—分散協調視覚システムにおける対象の行動理解法—

内部 英治　浅田 稔　細田 耕

uchibe@robotics.ccm.eng.osaka-u.ac.jp
大阪大学 工学部 電子制御機械工学科
〒565 大阪府吹田市山田丘 2-1

学習者以外に能動的に行動できるエージェントが存在する環境では，学習者がエージェントを含めた環境の変化を予測できない限り，適切な行動を学習によって獲得できない．本報告では，学習者の行動が他のエージェントに及ぼす影響を観測を通して推定することにより，エージェントの分類および行動戦略を識別する方法を提案する．エージェントのモデルを同定するために，部分空間同定法の一つである正準相関分析に対し，赤池の情報量規準を応用する．モデルを獲得した後に，得られた状態ベクトルに基づき強化学習を適用する．提案する手法をサッカーロボットに適用し，本手法の有効性を検証する．

Classification and Identification of the Agent Using Sub-State Space Identification in Multi-Agent Environment
— Understanding Behavior of Other Agents in Distributed Cooperative Vision System —

Eiji Uchibe, Minoru Asada, and Koh Hosoda

Dept. of Mech. Eng. for Computer-Controlled Machinery
Osaka University, 2-1, Yamadaoka, Suita, Osaka 565, Japan

This paper proposes a method for agent behavior classification which estimates the relations between the learner's behaviors and the other agents' ones in the environment through interactions using the method of system identification. In order to identify the model of each agent, Akaike's Information Criterion(AIC) is applied to the result of Canonical Variate Analysis(CVA). Next, reinforcement learning based on the estimated state vectors is applied to obtain the optimal behavior. The proposed method is applied to soccer playing robots. Unlike our previous work, the method can cope with a rolling ball. Computer simulations and real experiments are shown and the discussion is given.

1 はじめに

動的な実世界でタスクを遂行することを学ぶ自律ロボットを実現することは，ロボティクスと AI の中心課題の一つである．従来のコンピュータビジョンの分野では 2 次元の画像系列から距離や形状などの 3 次元情報を復元する方法がとられた．しかし，多大な計算コストをかけて獲得された環境表現がロボットにとって適切である保証はない．そこで我々は 3 次元再構成を行なわず目的行動を達成するための手法として，視覚に基づく強化学習に関する研究を行なってきた [7, 2].

しかし，マルチエージェント環境では，学習エージェントから観測した環境の変化がランダムになる場合があるため，通常の強化学習をそのまま適用できない．マルチエージェント環境での学習を困難にしている理由として，

A 他のエージェントが確率的に行動選択を行なっている．

B 他のエージェントは，学習エージェント (観測者) とは異なる知覚を持っている．

といったことが挙げられる．よって，エージェント間に明示的な通信がない限り，学習者は現在の状態から，次の状態を予測できない．

しかしながら，これまでのマルチエージェント環境における学習を取り扱った研究には，他者の行動に関する仮定が理想的なものが多かった．Littman [5] はゲーム理論を応用したマルチエージェント環境での強化学習法を提案し，1 対 1 の格子上のサッカーゲームに適用した．そこでは，相手の戦略が学習者に提示されていた．Sandholm and Crites [6] は IPD(iterated prisoner's dilemma) に強化学習を適用し，学習が成功するためには，十分な過去の観測量と行動が必要であることを示した．しかし，その履歴の長さを決定するのは，一般に困難な問題である．

これまで説明した通り，マルチエージェント環境で学習が収束するためには，他のエージェントの行動戦略が何らかの形で学習者に与えられている必要がある．しかし，エージェント間で明示的な通信がない場合には，学習者は全ての状況を観測できない部分観測問題のため，最適である保証はないが，学習者ができることは，自らの行動と観測を通して，自分の行動と相手の行動の関係を推定することだけである．

そこで本報告では，学習者の行動が他のエージェントに及ぼす影響を観測を通して推定する手法を提案する．ここで，我々は問題 **B** に重点をおくため，他者の行動戦略は変化しないと仮定する．赤池の情報量規準 (AIC) [1] を部分空間同定法と呼ばれるシステム同定の一手法である正準相関分析 (CVA) [4] の状態ベクトルの次元決定に適用する．ここで，予測の精度は必ずしも厳密である必要はなく，一般にモデルの正確さと制御則の設計はトレードオフである．そこで，獲得された状態ベクトルをもとに強化学習を適用し，モデルの不正確さを吸収する．提案する手法を 2 つの能動的なエージェントを含む環境に適用し，本手法の有効性を検証する．

2 エージェントの判別

2.1 正準変量解析法 (CVA)

先に述べたとおり，学習が成功するためには，学習者が自分自身の行動の結果，環境の次の状態をある程度予測できる必要がある．以下では，システム同定の手法を用いて，自分自身の行動と観測の結果から，観測量を予測できる状態ベクトルを発見し，それを用いて対象物を判別する手法について述べる．

近年，多入力多出力のシステムを同定する手法として，部分空間同定法が注目されている [8]．部分空間同定法は予測誤差法などの古典的な同定法とはことなり，状態空間の特定のパラメトリゼーションは用いないため，多入力多出力の線形システムの同定に適用できる．ここでは，正準変量解析法 (Canonical Variate Analysis, 以下 CVA) [4] を用いる．

いま，システムのモデルが

$$\begin{aligned} \boldsymbol{x}(t+1) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{w}(t), \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) + \boldsymbol{v}(t), \end{aligned} \quad (1)$$

として，与えられると仮定する．ここで，$\boldsymbol{u}(t) \in \Re^m$ を入力ベクトル (エージェントの行動に相当)，$\boldsymbol{y}(t) \in \Re^q$ を出力ベクトル (エージェントの観測に相当) である．また $\boldsymbol{A}, \boldsymbol{Q} \in \Re^{n\times n}$, $\boldsymbol{B} \in \Re^{n\times m}$, $\boldsymbol{C} \in \Re^{q\times n}$, $\boldsymbol{D} \in \Re^{q\times m}$, $\boldsymbol{S} \in \Re^{n\times q}$, $\boldsymbol{R} \in \Re^{q\times q}$ であり，$\boldsymbol{v}(t) \in \Re^q$

, $\boldsymbol{w}(t) \in \Re^n$ は平均値 0，共分散行列が

$$E\left\{\begin{bmatrix} \boldsymbol{w}(t) \\ \boldsymbol{v}(t) \end{bmatrix}\begin{bmatrix} \boldsymbol{w}^T(\tau) & \boldsymbol{v}^T(\tau) \end{bmatrix}\right\} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{S} \\ \boldsymbol{S}^T & \boldsymbol{R} \end{bmatrix}\delta_{t\tau},$$

で与えられる白色雑音ベクトルである．CVA は雑音を含む過去の入出力のシーケンスから，式 (1) と同値なシステム

$$\begin{bmatrix} \boldsymbol{\mu}(t+1) \\ \boldsymbol{y}(t) \end{bmatrix} = \boldsymbol{\Theta}\begin{bmatrix} \boldsymbol{\mu}(t) \\ \boldsymbol{u}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{T}^{-1}\boldsymbol{w}(t) \\ \boldsymbol{v}(t), \end{bmatrix}, \quad (2)$$

を求める．ここで

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T} & \boldsymbol{T}^{-1}\boldsymbol{B} \\ \boldsymbol{C}\boldsymbol{T} & \boldsymbol{D} \end{bmatrix}, \boldsymbol{x}(t) = \boldsymbol{T}\boldsymbol{\mu}(t), \quad (3)$$

である．$\boldsymbol{T}$ は適当な正則行列である．$\boldsymbol{\mu}(t)$ は入出力から作られた新しい状態ベクトルである．以降，$\boldsymbol{\mu}$ と $\boldsymbol{x}$ を同義に用いる．CVA の詳細は他の文献 [4] を参照されたい．ここでは，CVA による同定法を簡単に述べるに留める．

CVA アルゴリズム

1. 入出力のデータセット $\{\boldsymbol{u}(t), \boldsymbol{y}(t)\}, t = 1, \cdots N$ に対して，新しいベクトル

$$\boldsymbol{p}(t) = \begin{bmatrix} \boldsymbol{u}(t-1) \\ \vdots \\ \boldsymbol{u}(t-l) \\ \boldsymbol{y}(t-1) \\ \vdots \\ \boldsymbol{y}(t-l) \end{bmatrix}, \quad \boldsymbol{f}(t) = \begin{bmatrix} \boldsymbol{y}(t) \\ \boldsymbol{y}(t+1) \\ \vdots \\ \boldsymbol{y}(t+k-1) \end{bmatrix},$$

を構成する．以下，$N' := N-k-l+1$, $l' := l(m+q)$, $k' := kq$ とする．

2. 以下の共分散行列 $\hat{\boldsymbol{\Sigma}}_{pp}$, $\hat{\boldsymbol{\Sigma}}_{pf}$, $\hat{\boldsymbol{\Sigma}}_{ff}$ を計算する．簡単のため，$\hat{\boldsymbol{\Sigma}}_{pp}$, $\hat{\boldsymbol{\Sigma}}_{ff}$ は正則行列であると仮定する．

3. $\boldsymbol{\Sigma}_{pp}$, $\boldsymbol{\Sigma}_{ff}$ の固有値，固有ベクトルを求め，$\hat{\boldsymbol{\Sigma}}_{pp}^{-1/2}$, $\hat{\boldsymbol{\Sigma}}_{ff}^{-1/2}$ を計算する．

4. 以下の特異値分解

$$\hat{\boldsymbol{\Sigma}}_{pp}^{-1/2}\hat{\boldsymbol{\Sigma}}_{pf}\hat{\boldsymbol{\Sigma}}_{ff}^{-1/2} = \boldsymbol{U}_{aux}\boldsymbol{S}_{aux}\boldsymbol{V}_{aux}^T, \quad (4)$$
$$\boldsymbol{U}_{aux}\boldsymbol{U}_{aux}^T = \boldsymbol{I}_{l(m+q)}, \ \boldsymbol{V}_{aux}\boldsymbol{V}_{aux}^T = \boldsymbol{I}_{kq},$$

を計算し，行列 $\boldsymbol{U}$ を

$$\boldsymbol{U} := \boldsymbol{U}_{aux}^T\hat{\boldsymbol{\Sigma}}_{pp}^{-1/2}.$$

のように定義する．

5. n 次元ベクトル $\boldsymbol{\mu}(t)$ を

$$\boldsymbol{\mu}(t) = [\boldsymbol{I}_n \ 0]\boldsymbol{U}\boldsymbol{p}(t), \quad (5)$$

のように定義する．

6. パラメータ行列 $\boldsymbol{\Theta}$ を式 (2) に最小二乗法を適用して求める．

2.2 エージェントの分類，行動同定

状態ベクトルの次元 n と行列 $\boldsymbol{A}$ が観測者から見た対象の内部表現になる．そのため，状態ベクトル $\boldsymbol{x}$ の次元をいかに決定するかは重要であり，パラメータ数と推定精度のトレードオフを考慮して決定する必要がある．

ここでは，モデルの次元 n を決定するために，赤池の情報量規準 (AIC) を用いる．予測誤差を $\boldsymbol{\varepsilon}$，予測誤差の共分散行列を

$$\hat{\boldsymbol{R}} = \frac{1}{N-k-l+1}\sum_{t=l+1}^{N-k+1}\boldsymbol{\varepsilon}(t)\boldsymbol{\varepsilon}^T(t).$$

とする．このとき，情報量規準 $AIC(n)$ は

$$AIC(n) = (N-k-l+1)\log|\hat{\boldsymbol{R}}| + 2\lambda(n), \quad (6)$$

のように計算される．ここで

$$\lambda(n) = n(2p+m) + pm + \frac{1}{2}p(p+1). \quad (7)$$

である．AIC を最小にする n が最適な次元 n^* である．ここで

$$1 \le n \le \min(l(m+q), kq).$$

である．

3 強化学習

式 (2) で与えられる状態空間モデルが推定された後，学習者は強化学習を用いて，自分自身の行動を学習する．前節で次の状態を予測するのに十分な n

次元状態ベクトルが求められているため，環境はマルコフ性を満足しているとみなせる．

ここではQ学習 [9] を採用する．Q学習は確率的動的計画法に基づく強化学習法であり，エージェントを含めた環境がマルコフ過程としてモデルかできる時，環境とのインタラクションを繰り返すことにより，最適な行動を獲得できる．詳細は他の文献 [9] を参照して頂くとして，ここでは，アルゴリズムを以下に示すに留める．

Q 学習アルゴリズム

1. 全ての状態 x と行動 a の組に対して，$Q(x,a)$ を 0 に初期化する．
2. 現在の状態 x を観測する．
3. 現在の行動価値関数に基づき，行動 a を選択する．
4. 行動 a を実行する．環境は次の状態 x' に遷移し，報酬 r を生成する．
5. 行動価値関数を

$$Q_{t+1}(x,a) = (1-\alpha_t)Q_t(x,a) + \alpha_t(r + \gamma \max_{a' \in \boldsymbol{A}} Q_t(x',a')) \quad (8)$$

のように更新する．ここで α_t は学習率，γ は減衰係数である．

6. 2. に戻る．

4 実験結果

4.1 タスクと想定

提案する手法を図 1 に示す簡単なサッカーゲームに適用する．環境中には静止エージェントとしてゴールとライン，受動エージェントとしてボール，能動エージェントとして 2 台の移動台車が存在する．

各能動エージェントは TV カメラを持ち，そこから得られる画像情報のみを用いて行動する．従って，ボール，ゴール，他のエージェントに関する位置や重さといった量や，焦点距離などのカメラパラメータは未知とする．

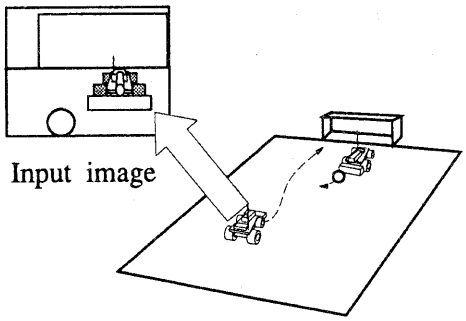


図 1: 実験環境

4.2 状態および行動空間

システムへの入力，すなわち台車の行動 $\boldsymbol{u}$ は

$$\boldsymbol{u}^T = [v \ \phi], \quad v, \phi \in \{-1, 0, 1\},$$

で表現される．ここで，v は台車の速度，ϕ は操舵角である．行動空間は，それぞれの要素を 3 つの部分行動に分割し，その組合せによって構成した．しかし，$\boldsymbol{u}^T = [0, \pm 1]$ の場合には，ロボットは実際には行動できないため，それら二つは除いた．結果として，7 つの行動で行動空間は構成される．

また，環境に対する自身の行動の効果は，視覚情報を通して獲得される．観測 (出力) ベクトルを図 2 に示す．ボールの場合は，画像上の重心位置 (x_c, y_c)

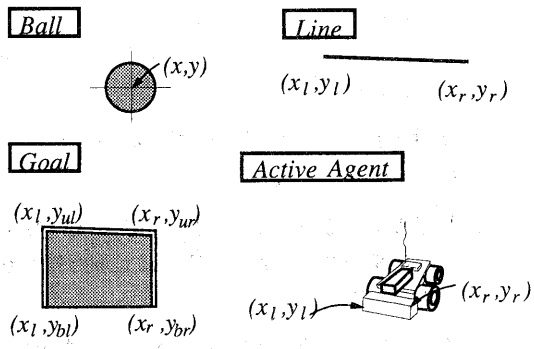


図 2: 画像特徴量

を，フィールドラインの場合は，両端の画像上での位置 (x_l, y_l) と (x_r, y_r) を用いる．ゴールの場合は，左上の x 座標 x_{ul} と左下の x_{bl} がほとんど等価であ

るため，$x_l = (x_{ul} + x_{bl})/2$ をかわりに用いる (x_r も同様)．エージェントの場合には，前面のプレートの対角の点の画像上の位置を用いて表現する．以上より，各物体に関する観測ベクトルの次元は，ボールが 2 次元，ラインが 4 次元，ゴールが 6 次元，エージェントが 4 次元である．

Q 学習を適用するために，状態空間を量子化する必要がある．状態ベクトル$\boldsymbol{\mu}$ の共分散行列は単位行列であるため，状態ベクトルの各要素 x_i を

$$x_i < -1, \quad -1 \leq x_i < 1 \quad 1 \leq x_i.$$

の 3 つに分割する．しかし，単一の行動が必ずしも一つの状態遷移に対応するわけではない．この問題は状態と行動のずれ問題 [2] と呼ばれる．この問題に対処するために，状態が変化するまで同じ行動を取り続けることにする．状態が変化した時，行動価値関数を更新する．

4.3 シミュレーション結果

CVA による推定に用いたラグと進みは $l = k = 5$ として，まず計算機実験を行なった．ライン，ゴール，ボールについて推定された状態ベクトルの次元を表 1 に示す．また，図 3 にボールの重心の画像上での y 座標の予測誤差を示す．時刻 40 と 230 の

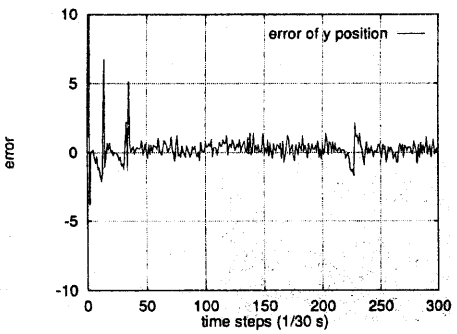


図 3: ボールの重心の画像上の位置の y 座標の予測誤差

時に，エージェントが強くボールを蹴っているため，予測誤差が大きくなっている．

次に，他のエージェントの識別結果を表 2 に示す．当然のことながら，ランダムに行動するエージェントは予測できない．それ以外の行動をするエージェントの場合は，ランダムエージェントと同じ次元で予測されるが，予測誤差はランダムエージェントの場合と比べてかなり小さくなっている．

表 1: 推定された状態ベクトルの次元

対象	n	$\log \lvert \boldsymbol{R} \rvert$	AIC
ライン	3	-2.14	-800
ゴール	2	-0.001	121
ボール	4	5.08×10^{-3}	64

表 2: 他のエージェントについての次元

戦略	n	$\log \lvert \boldsymbol{R} \rvert$	AIC
静止	3	-6.41	-589
ランダムウォーク	6	1.22	156
直進	4	-6.32	-194
左回りの前進	6	-0.463	79

シュート行動

図 4 にゆっくり動くボールをゴールにシュートする様子を示す．ここで学習中は，ボールはゴールの方からゆっくり動いてくるとする．ボールをゴールにシュートした時だけ報酬を 1 与え，それ以外は 0 とする．エージェントから出ている 2 本の線は，エージェントの視野を表している．

表 3 にこれまでの現在の画像情報のみを用いた手法 [2, 7] とのシュートの成功率の違いを示す．画像情報のみを用いた場合には，学習エージェントはボールやゴールの現在の画像情報のみを用いていたため，ボールが転がる場合には，適切な行動を獲得できない．言い換えると，環境が非マルコフ的であるため，行動価値関数が不安定になっている．

パス行動

パス行動は，動いている目標にボールを蹴り込む動作とみなせる．シュートするエージェントの戦略は，先に述べたシュート行動で獲得された行動で，固定である．他のエージェントにボールをパスできた時，報酬 1 を与え，エージェント間で衝突で発生し

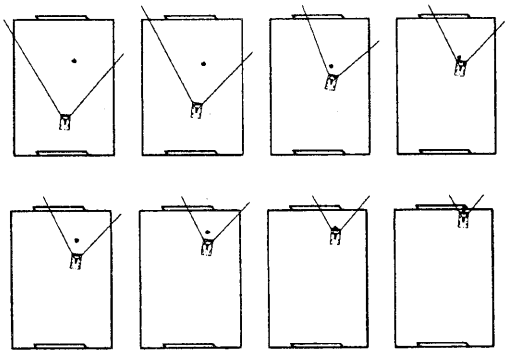


図 4: 転がるボールをシュートする様子

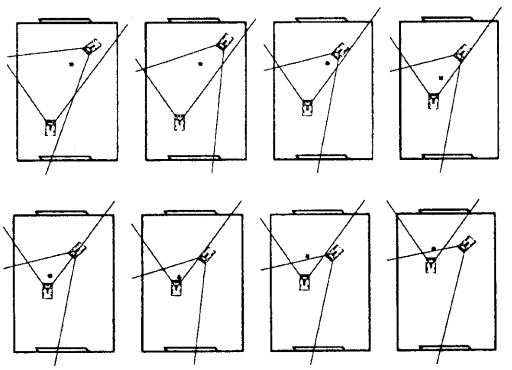


図 5: 他のエージェントにボールをパスする様子

た時には，−0.8 を与え，それ以外は，報酬を 0 とする．図 5 にパス行動の様子を示す．

表 3: パフォーマンスの比較

状態ベクトル	シュート行動の成功率 (%)	パス行動の成功率 (%)
従来法 [2, 7]	10.2	9.8
提案手法	78.5	53.2

4.4 実ロボットによる実験結果

図 6 に示す実システム [7] で実験を行なった．各ロボットに搭載された TV カメラから撮られた画像

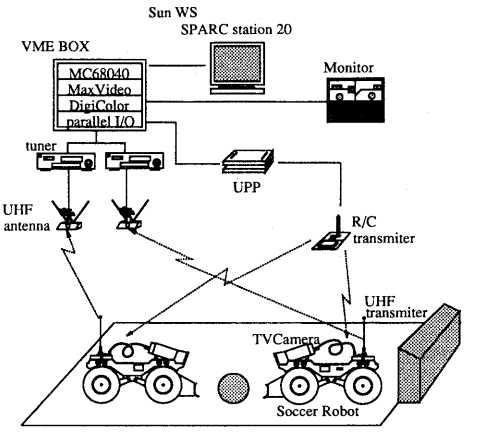
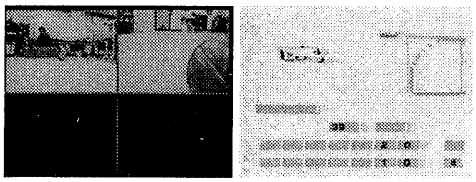


図 6: 実システム

は，ビデオ送信器でホストの UHF 受信器に送られ，一つの画像に統合された後，パイプライン型画像処理装置 (MaxVideo200) で処理される．処理の簡単化と高速化のため，ボール，ゴール，ロボットの前面プレートはそれぞれ赤，青，黄色に塗装されている．入力画像を図 7(a) に，処理画像を図 7(b) に示す．画像処理，状態同定，行動選択などは，ホスト CPU(MC68040) 上の OS (VxWorks) によって制御される．ホスト CPU はイーサネットを介して Sun ワークステーションに接続されている．ロボットの制御にはリモートブレインシステム [3] を採用し，市販のラジコン車を改造している．



(a) 入力画像　　(b) 処理画像

図 7: ロボットの入力画像と処理画像

実ロボットの場合は，コンピュータシミュレーションとは異なり，ロボットがランダムに行動するのは

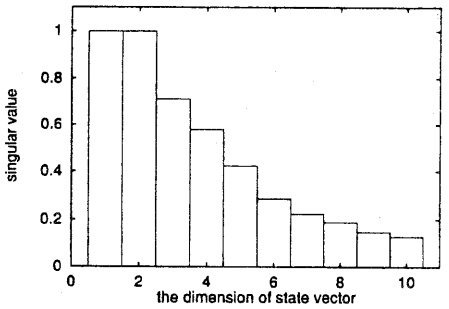


図 8: 状態ベクトルとその有効度の関係

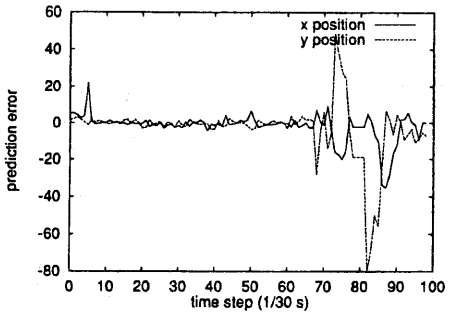


図 9: 相手エージェントに関する予測誤差

非常にコストがかかる．そのため，コンピュータシミュレーションによって獲得された結果を実ロボットに適用し，データ収集のための無駄な探索を減らした．実環境で収集したボールやゴールの次元を表 4 に示す．また，状態ベクトルと CVA アルゴリズム 3. の特異値 (正準相関係数) の関係を図 8 に示す．これから，推定された 4 次元の状態ベクトルの要素の内，最初の二つの要素は推定に同定度の影響を持ち，残りの二つは最初の二つと比較すると 6,7 割の影響力しかないことがわかる．

表 4: ボール，ゴールに関する状態ベクトルの次元

対象	n	$\log \lvert \boldsymbol{R} \rvert$	AIC
ゴール	3	−1.73	−817
ボール	4	1.88	284
移動エージェント	4	3.43	329

また，図 9 はロボットの前面プレートの右上の (x, y) の推定誤差である．70 から 90 ステップ間で誤差が大きくなっているのは，画像処理の失敗によりセンサ出力が真値と大きく異なっているためである．今回のタスクでは，相手エージェントに関する画像特徴量がそれほど変化しなかったため，推定された状態ベクトルはシミュレーションのときほど高次元にならなかった．

最後に，実環境で収集したデータをもとに推定をやり直し，さらに学習して獲得された行動を図 10 に示す．ただし，このときの学習はオフラインで行なわれる．シミュレーション結果をそのまま適用するよりも，行動は改善された．

4.5 部分観測問題

エージェントのセンシング能力は制限されている．そのため，全ての環境の状態を常に観測できるわけではない．例えば，相手エージェントが同一の行動をしていても，観測位置が異なる場合，対象の見え方は著しく異なり，同定結果は異なる．そのため，異なる観測地点で獲得された他のエージェントのモデルを統合する必要がある．

5 おわりに

本報告では，他のエージェントが存在する環境で強化学習を適用するための，他のエージェントの判別および行動を理解する手法を提案した．本手法は，予測誤差，状態ベクトルの次元，および，同定に必要な時間シーケンスの長さを考慮している．提案した手法をサッカーゲームに適用することにより，本手法の有効性を示した．

今後の方針としては，まず，状態空間の分節化の問題がある．今回は状態ベクトルの共分散行列が単位行列であることから，これまでの分割よりも恣意性を排除できたが，この方法でもロボットにとって最適である保証はない．また，高度なロボット間の協調，競合行動の学習に本手法が適用可能かどうかを検討する予定である．

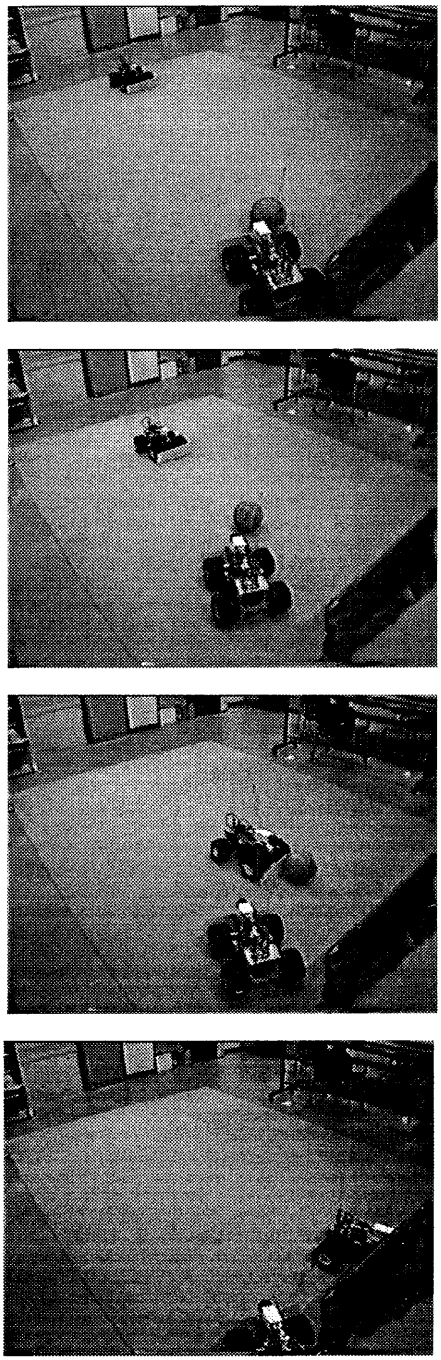


図 10: 獲得された行動

謝辞

本研究は日本学術振興会未来開拓学術研究推進事業「分散協調視覚による動的3次元状況理解」プロジェクトの一環として行った.

参考文献

[1] H. Akaike. A New Look on the Statistical Model Identification. *IEEE Trans. AC-19*, pp. 716–723, 1974.

[2] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-Based Reinforcement Learning for Purposive Behavior Acquisition. In *Proc. of IEEE International Conference on Robotics and Automation*, pp. 146–153, 1995.

[3] M. Inaba. Remote-Brained Robotics : Interfacing AI with Real World Behaviors. In *Preprints of ISRR'93*, Pitsuburg, 1993.

[4] W. E. Larimore. Canonical Variate Analysis in Identification, Filtering, and Adaptive Control. In *Proc. 29th IEEE Conference on Decision and Control*, pp. 596–604, Honolulu, Hawaii, December 1990.

[5] M. 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.

[6] T. W. Sandholm and R. H. Crites. On Multiagent Q-learning in a Semi-competitive Domain. In *Workshop Notes of Adaptation and Learning in Multiagent Systems Workshop, IJCAI-95*, 1995.

[7] E. Uchibe, M. Asada, and K. Hosoda. Behavior Coordination for a Mobile Robot Using Modular Reinforcement Learning. In *Proc. of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1996.

[8] P. Van Overschee and B. De Moor. A Unifying Theorem for Three Subspace System Identification Algorithms. *Automatica*, 31(12):1853–1864, 1995.

[9] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, May 1989.