学振未来開拓「分散協調視覚」プロジェクト 複数視覚移動ロボットの行動学習

特集 プロジェクト研究

浅田　稔*

Cooperative Distributed Vision Project : Learning Cooperative Behaviors for Vision-Based Mobile Robots

Key Words : Distributed Vsion, Cooperation, Reinforcement Learning

1．はじめに

本稿では，平成8年度よりはじまった学術未来開拓事業「分散協調視覚による動的3次元環境理解」プロジェクトにおける大阪大学チームの研究概要を説明する．「分散協調視覚プロジェクト」全体に関しては，参考文献並びにホームページ[1]を参照して頂くとして，我々の研究グループでは，「視覚に基づくロボットの多種行動の獲得及び統合方式の開発」をめざしている．本研究は，視覚を備えた移動ロボット群に統一の取れたチームプレイ動作を行わせるためのロボット間の協調及び競合行動を視覚に基づいた強化学習によって獲得する手法を考案し，その有効性を実ロボットによる実験によって検証することを目的とする．特に，物理的身体性を持ったロボットに協調動作を行わせるには，いわゆる「アイ・コンタクト」のようなメッセージ通信なしのコミュニケーションを実現することが重要であると考え，そのための視覚機能や行動制御方式の研究を実施している．

具体的な問題設定としては，ロボットによるサッカー競技を取り上げ，ドリブル，シュートなどの個々のロボットの行動，パス，センタリングなどの複数ロボット間の協調行動，さらにはブロックなどの競争行動を視覚に基づいた学習によって獲得するための方式の開発を目指している．この問題は，敵，味方に分かれた多くのロボット群が存在するという環境の中で，他の個々のロボットの行動理解及びチームとしての行動戦略パターンの理解などといった高度な視覚認識の問題を含んでいる．検証手段として，筆者らは，ロボットによるサッカー競技会と研究集会「ロボカップ」を開催しており，世界中の研究者がこぞって参加している[2]．いかでは，まず最初に，行動学習法として利用している強化学習について説明する．次に，実際のロボットに適用する問題を指摘し，我々のアプローチを紹介する．

*Miroru ASADA
1953年10月1日生
1982年大阪大学大学院基礎工学研究科物理系専攻修了
現在，大阪大学・大学院工学研究科・知能・機能創成工学専攻，教授，工博，知能ロボット
TEL 06−6879−7347
FAX 06−6879−7348
E-Mail asada@aws-eng.osaka-u.ac.jp

2．強化学習の枠組

強化というと，アメリカの行動心理学者スキナーのスキナーボックスが思い出される．鼠を箱の中に入れ，その中にあるレバーを鼠がたまたま押すと，餌がもらえる実験で，一旦レバー押しを憶えると何回もレバーを押し続ける行動をとるそうである（図1参照）．このときレバーを押す行為に正の強化（餌，報酬，価値などなど）が与えられる．強化学習は，これを確率的動的計画法の枠組で定式化したものである．

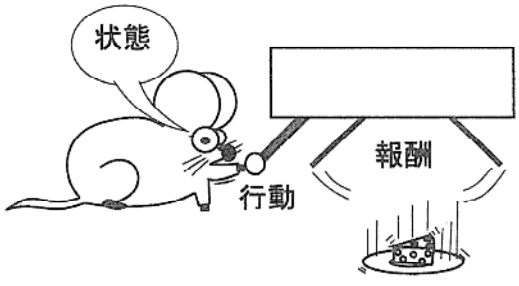


図1　スキナーの鼠箱

鼠は箱の中で，どこにいたり，レバーがどのように見えるかなどの状態($s \in S$：状態集合)が分かり，前に進んだり，レバーを押すなどの行動($a \in A$：行動集合)をとることができる．このとき，環境は厳密にはマルコフ過程としてモデル化され，現在の状態と鼠がとった行動により確率的に(うまく見えなかったり，脚を滑べらしたりするかもしれないので)次の状態($s' \in S$)遷移する．その結果報酬(r：例えばチーズ)が与えられる．状態遷移が既知であれば通常の動的計画法の枠組で最適行動が得られるが，未知のとき環境内で試行錯誤しながら，状態遷移と最適行動を推定しなければならない．これが確率的DPとか逐次的DPなどと呼ばれる結縁である．最も良く利用される強化学習法としてQ学習が有名で，状態sで行動aをとる行動価値関数Q(s, a)は，試行錯誤により，次式で更新される．

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

ここで，s'は，次状態，αは学習率で0と1の間の値をとる．γは，減衰率で，現在の行動が将来に渡ってどれくらい影響を及ぼすかを決めるパラメータで，0と1の間の値をとり，小さい程影響が少ない．行動選択は，学習の収束時間を決める要因の一つで，一旦憶えた成功例を何回も繰り返して上達させるか，別のアプローチを未経験のところから探すかのトレードオフがある．能動学習の観点からは前者が有利であるが，準最適解しか発見できない可能性が高くなる．

3．複数ロボット環境下での学習

強化学習を実際にロボットタスクに適用するには，様々な課題がある．まず，学習がうまく収束するような状態・行動空間を事前に用意しなければならない．また，通常の強化学習では，ゴール状態のみに報酬を設置する場合が多く，報酬が得られるまでに多大な学習時間を要する．更に，問題が複雑化したときに，どのようにスケールアップするかなどである．複数のロボットが協調・競合しあうサッカー競技のタスクでは，更に以下の課題がある．

A 他者の行動政策は，学習者にとって未知であり，センサから得られる瞬間の情報だけでは，次の状況を予測することは困難である．

B 特に学習の初期段階において，他者のランダムな行動戦略が，学習者の学習過程に悪影響を及ぼす．

学習を成功させるためには，学習者は他者の行動を自分自身の観測と行動を通して予測できる必要がある．従来研究では，十分にこれらの問題を解決していなかった．そこで本研究では，学習者の観測と行動を通して，学習者と他者の行動の関係を局所予測モデルとして推定し，その結果をもとに強化学習をおこなう手法を，また，マルチエージェント系での学習を安定にするための学習のスケジューリング法も提案し実験した．

環境には2台のロボットが存在しねそれぞれに対して異なるタスクを与える．環境はゴール，ボール，移動ロボットから構成され，学習者はそれぞれに対して局所予測モデルを構築する．各学習者は予測モデルを構築した後に，強化学習によって目的行動の学習を開始する．

4．学習アルゴリズム

4.1　アーキテクチャ

図2は各ロボットに与えられる行動獲得のためのアーキテクチャである．はじめに，学習者はセンサ情報だけでなく，学習者自身の行動のシーケンスから局所予測モデルを構築する．局所予測モデルは対象を次の運動が予測できるような状態ベクトルを推定する．次に推定された状態ベクトルをもとに，協調行動を獲得のための学習を開始する．

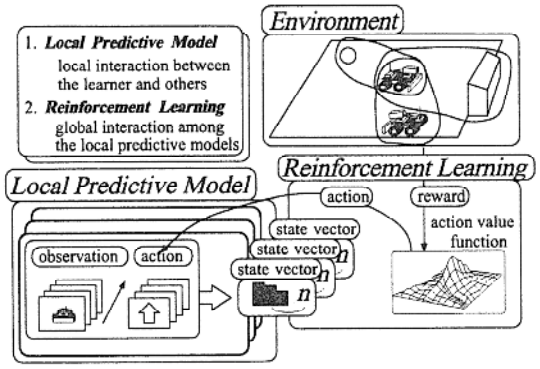


図 2　提案するアーキテクチュア

4.2　複数ロボットの学習のスケジュール

複数ロボットが存在する環境下で協調行動を獲得

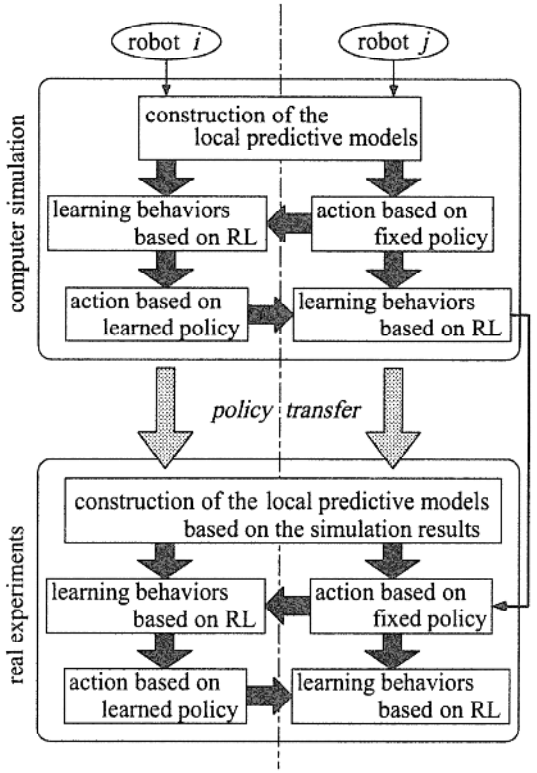


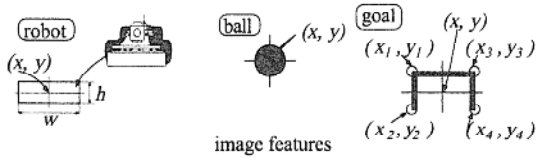
図 3　マルチエージェント環境における学習スケジュール

させるために，学習のスケジューリングをおこなった．図3に学習のスケジュールを示す．最初に局所予測モデルの構築を計算機上でおこなう．局所予測のモデルは，各ロボットが同時に推定し，その時の行動戦略はランダムである．次に，推定結果をもとに学習を開始するが，学習の初期段階でのランダム性を排除するため，学習するロボットを1台指定し，それ以外のロボットの行動戦略を固定する．学習ロボットの学習が終了した後で，別のロボットの学習を開始する．計算機シュミレーション上で局所予測モデルの更新と行動価値関数の更新を繰り返すことで，各ロボットは目的の行動を獲得する．

次に，学習結果を実ロボットに適用し，そのときの結果をもとに局所予測モデルを更新する．シュミレーションの結果を初期値とすることで，実環境での探索を短縮できる．局所予測モデルを構築した後で，シュミレーションと同様にして行動学習おこなう．

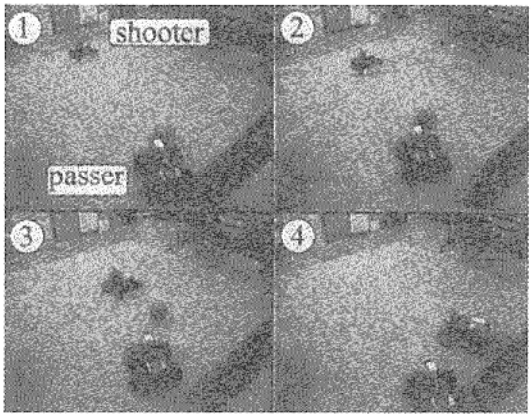
5．タ　ス　ク

提案手法を，2台のロボットが存在する環境下での，簡単なサッカーゲームに適用した(図5参照)．各ロボットはTVカメラを一つ搭載し，そこから得

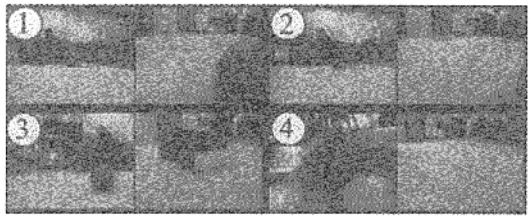


robot	ball	goal
area	area	area
center position	center position	center position
height	radius	4 corners
width		

図 4　ボール，ゴール，相手ロボットの画像特徴



(a) 鳥瞰図



(b) ロボットからの映像

図 5　獲得された行動

られる画像情報から環境の状況を観測する．モータコマンドとして，各ロボットはアクセルとステアリングの2自由度を待つ．また，各ロボットが観測できる画像特徴量(観測ベクトル)を図4に示す．結果として，ボール，ゴール，ロボットに関する観測ベクトルの次数はそれぞれ4，11，5となる．詳細は[3]を参照されたい．

6．実　験　結　果

最初にシューターとパッサーは，ボール，ゴール，そして互いの局所予測モデルを，計算機のシュミレーション上で構築する．次に，シューターを静止させた状況下で，パッサーは行動の学習を開始する．パッサーの学習が終了した時点で，パッサーの行動政策

を固定し，シューターの学習を開始する．パッサーは，ボールをシューターにパスしたときに報酬1を受け取り，シューターはボールをゴールにシュートしたときに，報酬1を受け取る．さらに，ロボット間で衝突が発生した場合，−0.3の報酬が与えられる．

計算機上での学習が終了した時点で，獲得された結果を実ロボットに適用する．実環境での局所予測モデルを再構築する為の行動戦略は，80％の確率でシュミレーションで獲得された行動戦略を用い，20％の確率で，ランダムに行動する．局所予測モデルを構築するために，実環境で100回の試行をおこなった．局所予測モデルが更新された後で，ロボットは行動価値関数を収集した実データをもとに洗練する．最後に，実環境でのパフォーマンスを計測するため，50回の試行をおこなった．図5に獲得された行動の例を示す．まず，パッサーがボールをシューターに向かってボールを蹴り，シューターはボールをゴールにシュートする．パッサーはボールを蹴った後は，シューターとの衝突を回避するための行動をしていることがわかる．

7．おわりに

本稿では，学振未来開拓推進事業「分散協調視覚プロジェクト」における大阪大学研究グループの研究テーマとして，強化学習を複数のロボットが存在する環境下に適用するための手法についてその概要と実験結果を示した．但し，ここで述べた状況は，2者の協調関係のみを対象にしているので，環境の下位のダイナミクスの推定とそのあとの強化学習により，協調行動を実現できたが，3者以上の場合，より複雑な相互作用が予想され，現手法では対応可能とは言えず，それらに対応するための手法の開発が，今後の課題である．現在，遺伝的手法を用いて，3台以上のロボットが存在する環境下で，協調，競合行動を学習させる実験を行っている．未来開拓事業の最終年度(平成12年度)には，より多くのロボットが協調して相手チームと戦う模様をお伝えしたい．何しろ，ロボカップの最終目標は，「11台のヒューマノイドチームによるワールドカップチャンピオンチームの打破」だから…

参考文献

1) http://vision.kuee.kyoto-u.ucjp/CDVPR/index-jp.html
2) 浅田 稔.「サッカーロボットの夢：ーロボカップ：実ロボットリーグ—」．生産と技術，Vol. 50, No.2, pp.73-75, 1998.
3) E. Uchibe, M. Asada, and K. Hosoda. “state space construction for behavior acquisition in multi agent environments with vision and action”. In Proc. of ICCV 98, pages 870-875, 1998.