社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B201-1 (5/4)

強化学習を用いたサッカーエージェントのボール保持行動獲得

A Reinforcement Learning for a Ball Holding Task in the RoboCup Soccer 2D Simulation

秋山英久 † 岡山智彦 ‡ 中島智晴 ‡
Hidehisa AKIYAMA†Tomohiko OKAYAMA‡Tomoharu NAKASHIMA‡
福岡大学 † 大阪府立大学 ‡
Fukuoka University†, Osaka Prefecture University‡
akym@fukuoka-u.ac.jp,{tomohiko.okayama@ci.cs,tomoharu.nakashima@kis}.osakafu-u.ac.jp

概 要

In this paper, we propose a reinforcement learning approach for a ball holding task in the RoboCup soccer 2D simulation environment. We applied the Sarsa(λ) algorithm and evaluated several reward design in order to acquire the ball holding action with practical performance. The experiment results show that our approach can outperform the hand-coded heuristic method.

1 はじめに

マルチエージェントシステムにおいて，集団の協調行動のパフォーマンスを向上させるには個々のエージェントの性能向上も重要である．特に，自律的な意思決定を行う敵対的なエージェントが存在する場合，環境は動的であり，最適な行動ルールをハンドコーディングで実装することは難しい．このような環境下でのタスクにおいては，試行錯誤を通じて最適な制御規則を自動的に獲得できる強化学習の適用が有望と考えられる．敵エージェントが存在する動的環境下での制御技術が要求される問題として，サッカーのような対戦型スポーツが考えられる．本研究では，サッカーにおける基本的なプレイヤの制御技術である，単体エージェントによるボール保持行動を扱う．実験環境として RoboCup サッカー 2D シミュレータを用い，強化学習によって実用的な性能のボール保持行動を獲得させる．

2 関連研究

RoboCup サッカー 2D シミュレーション環境において，エージェントの制御技術や戦術的行動の獲得を強化学習によって試みた研究が数多く行われている．Gabel ら[2]らや Nakashima ら[4] はボールを捕捉する動作の精度向上を強化学習によって実現している．さらに，Gabel らは，ボールを保持する敵エージェントに対する守備行動を実用的な性能で獲得することに成功している[3]．Stone ら[1]は複数の敵味方エージェントが存在する環境でボールを保持することを目標とする Keepaway タスクを提案し，複数エージェントの行動政策の獲得を試みている．Carbalho ら[5]は，1 対 1 でのドリブルタスクへの強化学習の適用している[5]．本稿では，サッカープレイヤとしてより基本的な技術であるボール保持タスクを取り上げ，動的環境での明確な目標状態を設定できないタスクにおいて，強化学習の適用によって実用的な性能を

得ることを目指す．

3 強化学習によるボール保持行動獲得

3.1 ボール保持タスク

本稿では，敵エージェントにボールを奪われないことを目的とした単体のエージェントによるボール保持行動をボール保持タスクと呼ぶ．本稿で扱うボール保持タスクは，ボールを保持するエージェントとボールを奪おうとする敵エージェントとの 1 対 1 での環境とする．ボールを保持するエージェントはその場から移動せず，ボールをキックする行動のみ可能とする．

RoboCup サッカー 2D シミュレータでは，エージェントの位置を中心とした半径 $kickable\_area$ の円周内部がキック可能領域となる（図 1）．エージェントのキック可能領域内にボールが存在すれば，そのエージェントはボールをキックすることができる．よって，本稿におけるボール保持タスクは，敵エージェントのキック可能領域内へボールを侵入させないように，ボール保持エージェントのキック可能領域内でボールを移動させることが目的となる．これを達成するためには，現在の観測状態に応じて，自分自身のキック可能領域内の適切な位置へボールを移動させなければならない．しかしながら，自律的に行動する敵エージェントを避けるための最適なボール移動位置は明らかではなく，ハンドコーディングでのルール実装は難しい．そこで，本稿では強化学習によってボール保持タスクの解決を試みる．

3.2 強化学習アルゴリズム

本稿では，強化学習アルゴリズムとして Sarsa(λ)[7] を使用する．Sarsa では 1 ステップ毎に直前の状態-行動対のみ価値を更新するのに対して，Sarsa(λ) では，数ステップ遡って過去の価値を更新することができる．ボール保持タスクでは，

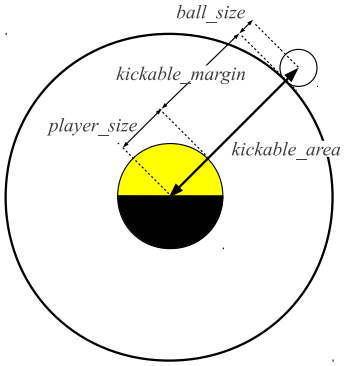

図 1: エージェントのキック可能領域．エージェントとボールとの間の距離が $kickable\_area$ 以下であれば，エージェントはボールを蹴ることができる．エージェントの半径 ($player\_size$) と $kickable\_margin$ の大きさはエージェントごとに異なる．

ステップ前の行動がボール保持の失敗に影響を与えると予想されるため，Sarsa(λ) が適していると考えられる．行動選択の手法として ϵ-greedy を用いる．

ボール保持タスクにおけるエピソードは，ボール保持エージェントのキック可能領域の外へボールが出てしまい，ボールの制御を失った時点で終了とする．RoboCup サッカー 2D シミュレータは離散時間シミュレータであるため，シミュレータの 1 シミュレーションステップを強化学習における 1 ステップとして扱う．

3.3 状態空間の設計

RoboCup サッカー 2D シミュレータは連続状態行動空間の環境である．強化学習をサッカーシミュレーション環境へ適用するために，状態空間を近似する手法として一次元タイルコーディング[7]を使用する．状態変数として以下の情報を用い，それぞれにタイリングを用意する．各状態変数で用いる座標系は，ボール保持エージェントの現在位置を原点とし，体の向きを x 軸正方向とする．

- ボール保持エージェントと敵エージェントとの距離

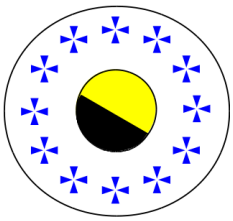

図 2: あらかじめ設定する 12 個のボール移動候補点.

- 敵エージェント位置の方向
- ボール保持エージェントとボールとの距離
- ボール位置の方向
- 敵エージェントの移動速度の大きさ
- 敵エージェントの移動速度の方向
- ボールの速度の大きさ
- ボールの速度の方向

3.4 行動空間の設計

サッカーシミュレーション環境では，状態空間と同様に行動空間も連続である．行動空間を離散化するために，エージェントのキック可能領域内にボールの移動候補位置をあらかじめ設定する．本稿ではこの候補位置を 12 箇所に限定した．まず，エージェントの体の正面を基準にエージェントの周囲を 12 分割する．分割した方向それぞれに，エージェントの体の中心から $ball\_size + kickable\_margin \times 0.72 + player\_size$ の距離だけ離れた位置にボールの移動候補位置を設定する (図 2).

3.5 報酬の設計

本稿では，次の 4 種類の報酬設計を用意し，数値実験によって獲得される政策の性能を調査する．

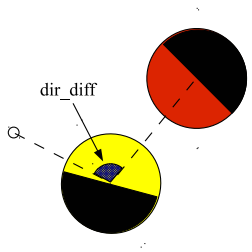

図 3: ボール保持エージェントを中心とし，ボール方向と敵方向のなす角の絶対値を dir_diff とする．

3.5.1 報酬設計 1: エピソード終了時の負の報酬のみ

エピソード終了時に，ボールの制御を失ったことへの罰則として報酬-1.0 を与える．

3.5.2 報酬設計 2: 保持成功ステップ数の反映および角度差によるステップごとの報酬

報酬設計 1 のように，エピソード終了時には負の報酬を与えるのは妥当と考えられる．しかし，長時間のボール保持を達成した場合には，ボール保持成功ステップ数の大きさに応じて罰則を小さくすることが望ましいと予想される．そこで，エピソード終了時に与える報酬を $-\frac{50.0}{steps}$ とすることで，ボール保持成功ステップ数に応じて負の報酬の絶対値を小さくする報酬設計を導入する．$steps$ はボール保持に成功したステップ数，すなわち，エピソードの長さである．

さらに，エピソード中のステップ毎に報酬を与える．エージェントがボール保持に成功している状態では，ボールは敵から遠い位置に存在する．そこで，図 3 に示される角度差 dir_diff の大きさに応じた報酬を設定する．ここでは，ステップ毎に $\frac{dir\_diff - 120.0}{15}$ を報酬として与える．

3.5.3 報酬設計 3: 保持成功ステップ数の反映およびステップごとに一定値の報酬

エピソード終了時の負の報酬として $-\frac{50.0}{steps}$ を使用する．また，エピソードのステップ毎に，敵エージェントがキック可能な状態であれば-1.0，キック不可能な状態であれば 1.0 の報酬を与える．

3.5.4 報酬設計 4: 累積エピソード数の反映

エピソードの実行回数を $episodes$ とすると，エピソード終了時の負の報酬として $-5.0 - \frac{\sqrt{episodes}}{20}$ を与える．

4 実験

4.1 実験設定

実験環境として，公式 RoboCup サッカー 2D シミュレータである rcssserver バージョン 15.0.1 を使用する．今回の実験では，観測誤差による影響を小さくするために，ノイズのない完全知覚情報を得られる fullstate 環境を用いる．ボール保持エージェントと敵エージェントの身体能力は同じとする．

本稿で提案した強化学習アルゴリズムに基づいたボール保持行動を RLHoldBall と呼ぶ．強化学習の各エピソードを実行中，ボール保持エージェントは常に RLHoldBall を実行する．敵エージェントは，ボールが自身のキック可能領域に存在すればボールをキックし，そうでなければボールを捕捉するための行動を実行する．ボール保持エージェントとは異なり，敵エージェントはボール捕捉のために移動動作を実行する．本来のシミュレータの仕様ではエージェントの移動には体力を消費する仕組みとなっているが，今回の実験では体力の消費は無いものとする．ボール保持エージェント，敵エージェントのいずれも agent2d-3.1.0[6]を使用する．

表 1: 使用した強化学習パラメータ．

α	0.125
λ	1
ϵ	0.01
$trace\_value$	1

4.1.1 エピソードごとの初期設定

エピソード開始時の初期状態として，次のようにボール，ボール保持エージェント，敵エージェントを配置する．

1. ボール保持エージェントをフィールドの中央に配置する．
2. フィールド中央からの距離が $ball\_size + kickable\_margin \times 0.72 + player\_size$ で方向がランダムな位置に，停止した状態でボールを配置する．
3. 敵エージェントをフィールド中央から 3.0m だけ離し，フィールド中央に対してランダムな方向に配置する．

エピソード開始後，ボール保持エージェントはボール保持行動を取り，敵エージェントはボール捕捉行動を取る．ボール保持エージェントのキック可能領域の外へボールが出た時点でエピソードが終了する．

4.1.2 強化学習パラメータの設定

今回の実験では，すべて 50000 回のエピソードを繰り返して学習を行う．強化学習のパラメータとして，Carvalho ら[5]と同じ設定を使用する (表 1)．α は学習率，λ はトレース減衰パラメータ，ϵ はランダムに行動を選択する確率である．適格度トレースの更新には入れ替え更新トレースを用いる．$trace\_value$ は入れ替え後の適格度トレース e の値である．

表 2: 状態変数の種類，それぞれの値の範囲，タイルの分割数．

状態変数	範囲	分割数
$dist(holder, opponent)$	[0.0, 3.0]	10
$angle(holder, opponent)$	[-180, 180]	12
$dist(holder, ball)$	[0.0, 3.0]	10
$angle(holder, ball)$	[-180, 180]	12
$velNorm(opponent)$	[0.0, 3.0]	10
$velAngle(opponent)$	[-180, 180]	12
$velNorm(ball)$	[0.0, 3.0]	10
$velAngle(ball)$	[-180, 180]	12

4.1.3 状態変数の設定

表 2 に使用する状態変数を示す．$dist(A, B)$ は A と B との距離，$angle(A, B)$ は A の体の方向を x 軸方向とした場合の B 位置の相対角度，$velNorm(A)$ は A の速度の大きさ，$velAngle(A)$ は A の速度の方向とする．各変数に対してタイリングは 32 層とする．RLHoldBall として，これらの状態変数のうち 6 個あるいは 8 個を使用する 2 種類の設定を用意する．状態変数の数が 6 の場合，ボールの位置速度情報は状態変数として使用されない．

4.1.4 他手法との比較

提案する強化学習アルゴリズムの性能を比較評価するために，以下の手法でもボール保持タスクを実行する．

- **RandomHoldBall**:
  ボールの移動先をランダムに決める．
- **HandCoding**:
  ハンドコーディングで設計した関数で移動候補点を評価し，もっとも高い評価値を得られた位置をボールの移動先に決定する．RoboCup2011 準優勝の HELIOS2011 が使用したものと同じものであ r ..
- **FarthestHoldBall**:
  敵からもっとも遠くなる位置でボールを保持する．

RandomHoldBall はボール移動候補位置として RLHoldBall と同じ位置を使用する．ただし，ボールの移動先は完全にランダムに決定される．HandCoding では，いくつかのルールとパラメータを人出で調整して評価関数を作成する．RLHoldBall と同様に複数の移動候補位置を生成し，評価関数が出力する評価値に基づいてボールの移動位置が決められる．FarthestHoldBall は，現在の敵位置からもっとも遠い位置を候補位置から選択し，ボールを移動させる．ボール移動位置とボール保持エージェントとの距離は RLHoldBall と同じ値を使用する．

4.2 実験結果

各手法で 50000 エピソードを実行した結果を図 4 に示す．このグラフはボール保持成功ステップ数の 1000 エピソード毎の移動平均を示している．RLHoldBall には報酬設計 2 を用いている．RLHoldBall は HandCoding を上回る性能を示しており，人手で設計した評価関数を用いるよりも高い性能を持つ行動政策を強化学習によって獲得できたことが分かる．

ボールの位置と速度を状態変数として追加し，状態変数の数を 6 から 8 に増やしても，結果に有意な差は見られない．これは，ボール保持行動の性能に関して，ボールの位置や速度は影響が小さい状態変数であったためと考えられる．

3.5 節の各報酬設計を使用した実験結果を図 5 に示す．グラフ中では，MinusOnly が報酬設計 1，PlusMinusDirDiff が報酬設計 2，PlusMinusKickable が報酬設計 3，PlusMinusSqrt が報酬設計 4 を表す．このグラフは保持成功ステップ数の 1000 エピソード毎の移動平均を示している．PlusMinusDirDiff と PlusMinusSqrt がほぼ同等の性能を示しており，他の報酬設計を上回る性能を得られた．

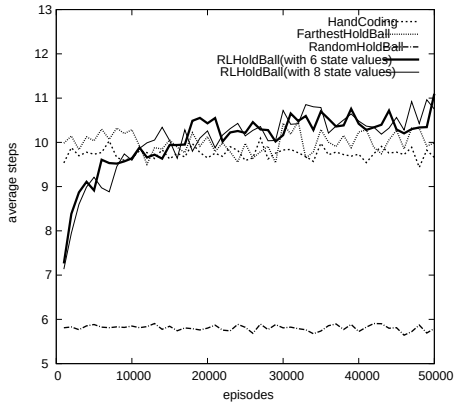

図 4: 報酬設計 2 と他手法との比較．保持成功ステップ数の 1000 エピソード毎の移動平均．

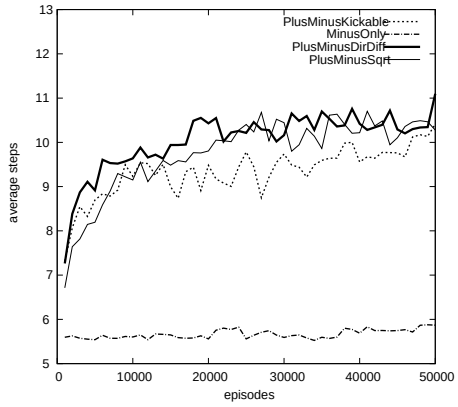

図 5: 各報酬設計の比較．保持成功ステップ数の 1000 エピソード毎の移動平均．

5 まとめと今後の課題

本稿では，RoboCup サッカー 2D シミュレーション環境において，強化学習を用いてボール保持行動を獲得させた．数値実験により，強化学習を用いることによって人出で設計した評価関数を用いる場合よりも高い性能が得られることを示した．今後の課題として，複数の敵エージェントへの対応，敵エージェントの特徴にや状況に応じた政策の獲得などが考えられる．

参考文献

[1] Peter Stone, Richard S. Sutton and Gregory Kuhlmann: Reinforcement Learning for RoboCup-Soccer Keepaway, Adaptive Behavior, 13(3), pp. 165-188, (2005).

[2] Thomas Gabel and Martin Riedmiller: Learning a Partial Behavior for a Competitive Robotic Soccer Agent, KI Zeitschrift, vol.20, pp. 18–23, (2006).

[3] Thomas Gabel, Martin Riedmiller and Florian Trost: A Case Study on Improving Defense Behavior in Soccer Simulation 2D: The NeuroHassle Approach. RobopCup 2008: Robot Soccer World Cup XII, pp. 61–72, (2008).

[4] Tomoharu Nakashima, Masayo Udo and Hisao Ishibuchi: A Fuzzy Reinforcement Learning for a Ball Interception Problem, RoboCup 2003: Robot Soccer World Cup VII, pp.559–567, (2004).

[5] Arthur Carvalho and Renato Oliveira: Reinforcement Learning for the Soccer Dribbling Task, Proceedings of the 2011 IEEE Conference on Computational Intelligence and Games, pp. 95–101, (2011).

[6] 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).

[7] 三上貞芳，皆川雅章: 強化学習，森北出版, (2001).