2009－GI－21 (9)
2009／3／9

Keepawayタスクにおけるマルチエージェントの協調行動の学習

伊佐野勝人[*1] 秋山英久[*2] 片上大輔[*1] 新田克己[*1]

東京工業大学大学院総合理工学研究科[*1] 産業技術総合研究所[*2]

RoboCupサッカーシミュレーションのKeepawayタスクに関する多くの研究では，協調行動を獲得するためにボールを持ったプレイヤに強化学習を適用している．本研究では，従来研究の分析から，ボールを持ったプレイヤの能力を改善するために，新しい報酬関数を提案し，プレイヤがとりうる行動としてドリブルを追加した．さらに，ボールを持たないプレイヤにも強化学習を適用した．これらの拡張により，ボールを奪われる割合が減少し，ボールを持たないプレイヤがボールを持ったプレイヤの動きの変化に柔軟に対応できるようになった．この結果，10秒以下のキープ時間の発生回数が従来研究よりも大幅に減少した．

Learning Multiagent Coordinated Behavior in Keepaway Task

Shoto ISANO[*1] Hidehisa AKIAYAMA[*2] Daisuke KATAGAMI[*1] Katsumi NITTA[*1]

Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology*1

National Institute of Advanced Industrial Science and Technology*2

Most researches on Keepaway task of RoboCup have applied reinforcement learning to the player with a ball to get a coordinated behavior. In this paper, from an analysis of early studies, we suggested to add a new reward function and a new dribble skill to expand the ability of the player with a ball. In addition, we focused on the players without a ball and applied reinforcement learning to them too. As the effect of these techniques, the probabilities of interception and mistakes to pass a ball were decreased, and the player without a ball was able to cope flexibly with the change of the player with a ball behavior. As a result, the occurrences of keep time under 10 seconds were clearly lower than earlier studies.

1 はじめに

マルチエージェントシステムにおける個々の「エージェントの能力」[Bond 88]を評価するベンチマークとして，RoboCupサッカーシミュレータ[Noda 98]のKeepawayタスクが注目を集めている．Keepawayとは，限られた領域内でキーパというチームがテイカーという相手チームにボールを取られないようにボールをキープすることを目的としたタスクである．Keepawayではエージェント数と領域の大きさを任意に決めることができるため，問題の複雑さを調節することができる．キーパのボールキープ時間が各エピソード時間となるため，強化学習のテストベッドとして適している．

Keepawayタスクを扱った従来研究では，強化学習を適用することでハンドコーディングアルゴリズムよりも平均キープ時間が上昇したという報告がある[Stone 01, 荒井 06]．しかし，これらの研究では，強化学習の適用対象はキーパのボールを持ったエージェントのみで，そのエージェントの行動パターンもボール保持とパスだけであった．また，荒井らの手法では平均キープ時間は長いものの，エピソード継続時間が短い場合が多いなど，いくつかの改善すべき点が挙げられる．

本研究では，2つの先行研究のうち平均キープ時間の長かった荒井らの研究の問題点を指摘し，その改善を試みた．

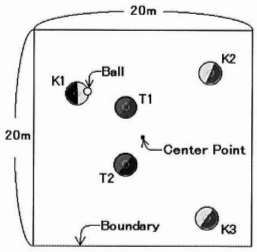

図1　3対2, 20m×20mのKeepaway

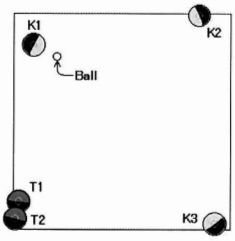

図2　Keepawayの初期配置

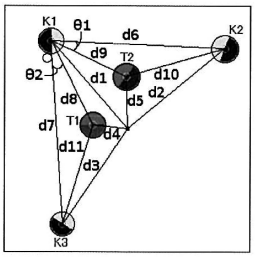

図3　パサーの状態変数

2　問題設定

2.1　Keepaway

Keepawayタスクの様子を図1に示す．本稿では，Stone，荒井らと同様に，3人のキーパと2人のテイカーが20m×20mの矩形領域内でボールを奪い合う設定とする．キーパがテイカーにボールを取られる，または，ボールが矩形領域の外へ出ると，そのエピソードは終了となる．エピソード開始時，テイカーは2人とも矩形領域の左下のコーナーに，キーパは残りのコーナーに1人ずつランダムに配置され，ボールは左上のコーナーの近くに置かれる（図2）．エージェントには移動量に制約を与えるためのスタミナパラメータが設定されているが，エピソード開始時に全エージェントのスタミナは最大値まで回復する．あるエピソードの終了状態が他のエピソードへ影響することは無い．

本稿では，各エージェントを識別するために，キーパに属するエージェントをボールから近い順にK1, K2, K3と呼ぶ．同様に，テイカーに属するエージェントをT1, T2と呼ぶ．さらに，K1がボールを制御可能な状態であるとき，K1をパサーと呼び，K1以外のキーパに属するエージェントをレシーバと呼ぶ．

2.2　従来研究

本節では，比較対象とする荒井らの研究[荒井 06]について説明し，その問題点を指摘する．

2.2.1　パサーのアルゴリズム

荒井らはパサーに強化学習を適用し，学習アルゴリズムとしてSarsa(λ)を用いた．状態変数として，11個の距離変数（図3 d1～d11）と2個の角度変数（図3 θ1, θ2）を設定し，さらに，1次元タイルコーディングによって状態数の削減を行っている．パサーが取りうる行動は3つのマクロ行動「ボール保持」「K2へパス」「K3へパス」の3種類である．パサーへの報酬は，エピソード終了時に負の報酬（罰）として与えられる．さらに，エピソード終了に対する各エージェントの責任の重さを反映するために，式(2.1)による報酬設計を提案している．

$$r_a = -\frac{1}{\mathrm{TaskEndTime} - \mathrm{LastActionTime}} \quad (2.1)$$

TaskEndTimeはエピソードが終了した時間を，LastActionTimeはエピソード内でパサーとして最後に行動を実行した時間を意味する．パサーとしての行動がエピソード終了時に近いほどTaskEndTimeとLastActionTimeとの差が短くなるので，エピソード終了に対するパサーとしての責任を各エージェントに上手く分配することができる．この報酬設計によって，逐次的な報酬を用いたStoneらの手法[Stone 01]よりも高い性能を示している．

2.2.2　レシーバのアルゴリズム

レシーバはハンドコーディングのアルゴリズムで制御される．このアルゴリズムでは，レシーバは，エージェントの密集度が低く，パサーからのパスコースがある位置へ移動するように設計されている．

2.3　従来研究の問題点

2.3.1　報酬設計の問題

パサーとT1の距離が一定値以下になると，T1にボールを奪われやすくなるだけでなく，パスコースの候補も少なくなるため，パサーがその場でボールを保持し続ける行動は危険である．パサーとT1の距離が一定値以下になる前にパサーがパスを出さなければ，エピソードが終了しやすくなる．よって，「ボール保持が危険である」状態をパサーは学習すべきである．しかし，荒井らの報酬関数ではタスク終了時

に報酬を与えているため，「ボール保持が危険である」ことをT1にボールを奪われたときにしか学習できない．

2.3.2 パサーの行動集合の問題

Stoneや荒井らの実験では，パサーが取りうるマクロ行動の種類が不十分であった．例えば，T1がパサーに接近した状況では，パサーはレシーバへパスを出すことでボールを奪われる危険を回避すべきである．しかし，T1やレシーバとの位置関係によっては成功が期待できるパスコースを，発見できない場合がある．Stoneや荒井らの設定では，そのような状況下であっても，その場にとどまってボールを保持することしかできなかった．その結果，T1によってボールを奪われやすくなったと考えられる．

3 提案手法

2.3節で指摘したパサーの問題点を改善するために，本研究では荒井らの手法に対して以下の変更を加える：

- 報酬関数の追加
- パサーが取りうるマクロ行動にドリブルを追加
- レシーバへの強化学習の適用

3.1 パサーに対する変更

3.1.1 報酬関数の追加

パサーとT1の距離が一定値以下になったときに「ボール保持が危険」であることをパサーが学習できれば，T1にボールを直接奪われて終了するエピソードが減少すると期待できる．本研究では，荒井らの手法に加えて，新たな報酬関数として式(3.1)と式(3.2)を導入する．

$$\mathrm{dist(P,T1)} < 5.0 \Rightarrow -1 \qquad (3.1)$$

$$r_a' = -\frac{\beta}{\mathrm{TaskEndTime} - \mathrm{LastActionTime}} \qquad (3.2)$$

dist(P, T1)はパサーとT1の距離である．式(3.1)は，パサーとT1の距離が5mより小さい場合に罰が与えられることを意味する．式(3.1)は条件が一致したすべての場合に適用され，エピソード途中でも報酬が与えられる．それに対して，式(3.2)はエピソード終了時のみに適用される．式(3.2)は式(2.1)の変形で，1以上の定数βが導入されている．βの値は，式(3.1)の報酬が式(3.2)の報酬を上回らないように設定する． 予備実験において，TaskEndTime − LastActionTimeの値はほとんどの場合10以下となったため，β=10とした．

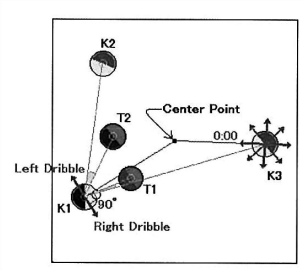

図4 パサーのドリブルとレシーバの移動

3.1.2 ドリブルの追加

T1によるボール奪取を回避し，パサーによるパスコースの発見を促すために，パサーのマクロ行動としてドリブルを追加する.ドリブルによるパサーの進行方向は，矩形領域の中心位置とパサーとを結ぶ線分に垂直な方向とする（図4）．一回のドリブル行動の継続時間は4サイクル（実時間で0.4）までとし，それ以上の時間を要する長距離のドリブルは禁止する．また，ドリブルによってボールが矩形領域外へ出てしまうことがないように制約を与えておく．

3.2 レシーバへの強化学習の適用

本研究では，提案手法によるパサーの動きの変化へ柔軟に対応することを目的として，レシーバにも強化学習を適用する．

3.2.1 レシーバの行動集合と状態表現

荒井らによるハンドコーディングアルゴリズムでは，矩形領域全体からレシーバの目標移動位置を選んでおり移動距離に制限がない．移動距離が大きい場合，移動途中にパスを受けることがある．この場合，パスを受けた位置がテイカーの近くであることや，ボールへの反応が遅れてトラップできないことが考えられるため，タスク継続が難しくなる．

Keepawayと同様に連続状態空間である水たまり問題へ強化学習を適用した研究では，エージェントの行動を上下左右への一定距離の移動のみとしている[Sutton 96]．本研究でもレシーバの行動にこの方法論を適用する．レシーバの移動方向は，レシーバの

位置から矩形領域の中心位置への角度を0度とした8方向へ限定する（図4）．レシーバの行動は，8方向への移動と，その場に留まりボール位置へ体を向ける行動を加えた全9種類とする．

状態変数は，従来研究の13個にレシーバ間の距離を追加して計14個とする．

3.2.2 報酬設計

Keepawayは，ある目標状態に到達するのではなく，「望ましい状態」を維持しながら無限に続けることが要請される連続タスクである．「望ましい状態」の報酬値を定めることは逐次的教師信号を与えることと等価である[荒井 06]．そのため，レシーバは常に周りの状態から判断してレシーブに適切なポジショニングを行う必要がある．そこで本研究では，レシーバの報酬関数として式(3.3)〜式(3.6)を設計した．dist(Self, C)はレシーバと領域の中心の距離，dist(Self, P)はレシーバとパサーの距離，dist(Self, R)はレシーバともう一人のレシーバの距離，そして，ang(Self, P, minT)はパサーを中心としたレシーバとテイカーの相対角度のうちもっとも小さい値(図3のθ1またはθ2)を意味する．

$$\mathrm{dist(Self,P)} < 10.0 \Rightarrow -1.0 \quad (3.3)$$

$$\mathrm{dist(Slef,R)} < 10.0 \Rightarrow -1.0 \quad (3.4)$$

$$\mathrm{dist(Slef,P)} > 10.0,\ 7.0 < \mathrm{dist(Self,C)} < 9.0 \Rightarrow +1.0 \quad (3.5)$$

$$\mathrm{ang(Slef,P,minT)} < 10.0 \Rightarrow -1.0 \quad (3.6)$$

まず，他のキーパとの距離を一定以上に保つために，予備実験で良い結果を得た10.0mを閾値として，他のキーパとの距離がこの値を下回った場合に-1の報酬を与えるように式(3.3)と式(3.4)を設定した．これは，テイカーがボールに追いつくまでの時間を確保するためである．またレシーバとパサーとの距離が10.0m以上離れていて，領域の中心からも離れている場合は+1の報酬を与えるように式(3.5)を設定した．さらにパスコース確保を学習することを意図して，ang(Self, P, minT)が10度以下なら罰を与えるように式(3.6)を設定した

3.3 1次元タイルコーディングの設定

レシーバの強化学習では，パサーよりも状態変数が1つ，行動数が6つ増加するため，状態行動対が3倍以上になる．状態数は1次元タイルコーディングによって削減されているが，状態行動対が増加すると，学習時間が非常に長くなることが予想される．

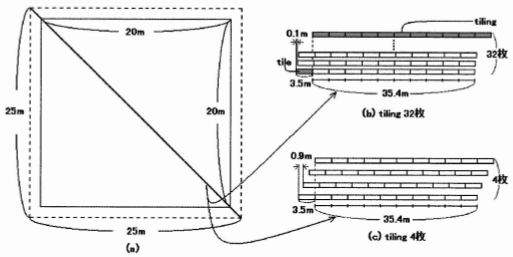

図5 タイルコーディングによる距離の状態近似

従来研究では，25m×25mの正方形の対角距離約35.4mを10分割しているため，タイル1枚の大きさは約3.5mになる（図5(a)）．さらに32枚のタイリングを重ねているため，精度は$3.5/32 \cong 0.1$[m]となる（図5(b)）．しかし，Keepawayにおける意思決定に対して，0.1m では精度が小さすぎることが実験的に分かっている．そこで，本研究ではタイリングの枚数を4枚にして，あらかじめ精度を粗くすることでタイル総数を減らすことにする．精度は$3.5/4 \cong 0.9$[m]となり，タイル枚数はパサーとレシーバ共に1/8に減少する（図5(c)）．

4 実験

4.1 実験設定

本実験では2.1節で定義したKeepawayに従って，20m×20mの領域で行う3対2のタスクを扱う．テイカーは2人ともボールを追う戦術をとる．マクロな個人スキルや状態取得の関数には，librcsc[秋山 06]を用いた．キーパの行動は，以下に示すlibrcscの行動ライブラリから選択される．

・パサー

StopBall()：自分とボールの現在の位置関係を維持する．

KickMultiStep()：目的の位置へ向けてボールを蹴る．

Dribble()：目的の位置へドリブルする．

・レシーバ

GoToPoint()：目的の位置へ移動する．

TurnToBall()：ボールのある方向へ身体を向ける．

Intercept()：ボールを追いかける．

Sarsa(λ)におけるλの値は，Stoneらや荒井らの実験，そして本研究の予備実験においても最も良い結果が得られたλ＝0を用いる．その他の学習パラメータは，荒井らの実験と同じ設定にするため，α = 0.125, γ = 0.95, ε＝0.01とする．エージェントによる観測誤差を可能な限り排除し，提案手法が学習に与える影響を観察するために，エージェントは常にすべての物体の情報を知覚できる設定とした．

4.2 実験手順

学習を行う8種類のプログラムを表1に示す．略称のRLPがパサー学習，HCRがハンドコーディングレシーバを表し，一番上の従来手法がこれに当たる．またdistT1は報酬関数の追加，dribはドリブルの追加，RLRはレシーバ学習を表し，3つの提案手法となる．変更数1～3は，それぞれ各提案手法を従来手法に組み込んだ数を表す．

2500エピソードを学習時間としてRLP_HCRを学習させる．ただし本研究の各提案手法の状態行動対は，dribの追加で約1.7倍，RLRの変更で3倍以上になるため，dribとRLRの入る実験では，学習時間をそれぞれ5000エピソードと10000エピソードにして学習を行う．

学習後は1000エピソードを5回試行し，キープ時間とエピソード終了原因を調べる．タスク終了原因は表2に示すように4タイプに分類される．判定はタスク終了時の状態から，決定木を用いて自動的に行われる．自動判定と観察による判定を比較した結果，90%以上の精度であったため信頼性は十分であると考えられる．

表1 実験の種類

変更数	略称	説明
	RLP_HCR	従来手法
1	RLP_HCR_distT1	報酬関数
	RLP_HCR_drib	ドリブル
	RLP_RLR	レシーバ学習
2	RLP_HCR_distT1_drib	報酬関数 + ドリブル
	RLP_RLR_distT1	報酬関数 + レシーバ学習
	RLP_RLR_drib	ドリブル + レシーバ学習
3	RLP_RLR_distT1_drib	報酬関数 + ドリブル + レシーバ学習

表2 タスクの終了原因

名前	状況
Near taker	T1に直接ボールを奪われる
Pass cut	takerにパスをカットされる
Area out	レシーブした場所が領域外
Trap miss	トラップをミスする

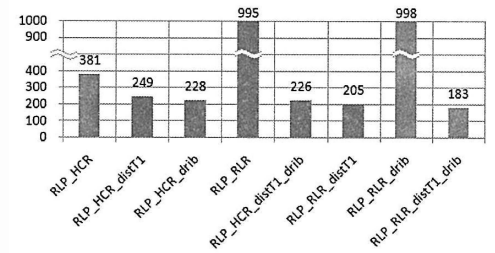

図6 10秒以下のエピソードの発生回数

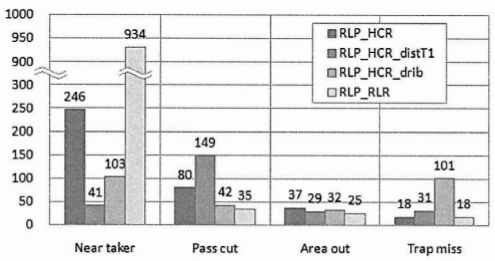

図7 10秒以下で終了したエピソードの終了原因（変更数1）

4.3 結果

4.3.1 10秒以下のキープ時間の発生回数

10秒以下のキープ時間になったエピソードの発生回数を図6に示す．従来手法のRLP_HCRでは発生回数が4割近くになっており多いことが分かる．各提案手法とその組み合わせでは，RLP_RLRとRLP_RLR_distT1以外，発生回数を減少させる効果が見られた．

4.3.2 タスク終了原因

従来手法と各提案手法の10秒以下のエピソードのタスク終了原因を図7に示す．RLP_HCRの主なタスク終了原因は，Near takerとPass cutであることが分かる．このことから荒井らの手法では，パサーとT1の距離が一定以下になったとき「ボール保持が危険」であると学習できていないことが確認できる．

・変更数が1つの場合

図7より，RLP_HCR_distT1では，Near takerが減少し，Pass cutが増加している．これは報酬関数の追加では，パサーがT1との距離が5.0m以下になったときにパスを出すため，T1にボールを直接奪われることはないが，パスコースが確保されていない状態でもパスを出すため，パスが失敗しやすくなったと考えられる．

RLP_HCR_dribでは，Near takerとPass cutが減少し，Trap missが増加した．これはドリブルの追加では，T1にボールを奪われる前にドリブルを実行してパスコースを確保するので，T1にボールを奪われにくくなり，パスが失敗しにくくなったと考えられる．しかし，ドリブルによるパサーの移動と同時に，レシーバも移動を開始するので，移動中にレシーバにパスを出すことになり，トラップをミスしやすくなったと考えられる．

RLP_RLRでは，Near takerが大幅に増加しており，RLP_HCRよりもボール保持の傾向が強まったと考えられる．図6に示す10秒以下のキープ時間のエピソードがほぼ10割であることから，学習によって性能が改善されなかったことが分かる．

・変更数が2つの場合

従来手法と各提案手法を2つ組み合わせた10秒以下のキープ時間のタスク終了原因を図 8 に示す．報酬関数の追加とドリブルの追加を組み合わせたRLP_HCR_distT1_dribでは，Near takerとPass cutが減少し，Trap missが増加した．これはドリブルを単体で追加した場合と同じ影響である．報酬関数の追加ではPass cutが増加したが，ドリブルと組み合わせることにより，増加することはなくなった．

報酬関数の追加とレシーバ学習を組み合わせたRLP_RLR_distT1では，Near takerが減少し，Pass cutが増加した．これは報酬関数を単体で追加した場合と同じ影響である．しかしレシーバの学習の効果により，Pass cutの増加率は報酬関数を単体で追加した場合よりも低くなった．

ドリブルの追加とレシーバの強化学習の適用を組み合わせたRLP_RLR_dribでは，RLP_RLRと同様，Near takerが大幅に増加した．ここでもRLP_HCRよりもボール保持の傾向が強まったと考えられる．

・変更数が3つの場合

従来手法と各提案手法を3つすべて組み合わせた10秒以下のキープ時間のタスク終了原因を図9に示す．3つの提案手法をすべて組み合わせたRLP_RLR_distT1_dribでは，Near takerが減少した．Pass cutは増加しているが僅かでしかない．このため3つの手法を組み合わせることで，負の影響が消え，効果だけが残った．

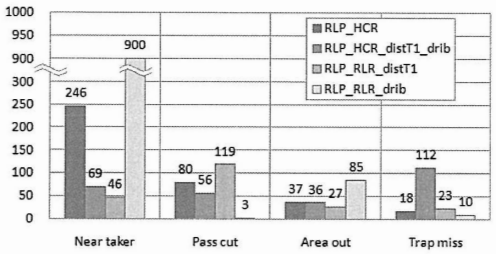

図8　10秒以下で終了したエピソードの終了原因（変更数2）

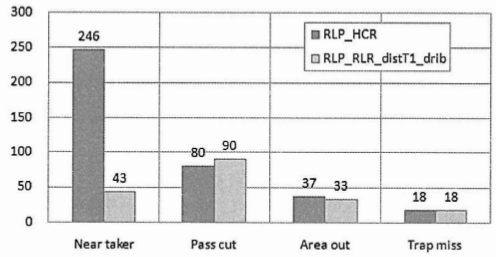

図9　10秒以下で終了したエピソードの終了原因（変更数3）

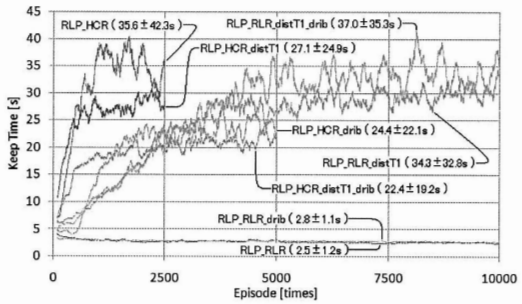

図10　学習曲線と平均キープ時間

4.3.3　タスク継続時間

100 エピソードごとの移動平均をとった学習曲線と平均キープ時間を図10に示す．荒井らの手法である RLP_HCR の学習曲線を明らかに上回るものはないが，本研究で加えた変更点をすべて組み合わせたRLP_RLR_distT1_drib は RLP_HCR と近い性能を獲得していることが分かる．Near taker が大幅に増加した RLP_RLR と RLP_RLR_drib では性能の改善がまったく見られなかった．

5 考察

5.1 タスク継続時間について

5.1.1 Near takerと10秒以下のエピソードの発生回数の相関関係

図7〜9のNear takerと図6の10秒以下で終了したエピソードの発生回数を比較すると、相関関係があることが分かる。このため，Near takerの割合を減少させることで，10秒以下で終了したエピソードの発生回数を抑えることができると考えられる．

5.1.2 タスク継続時間ごとの頻度

タスク継続時間ごとの頻度について，RLP_HCRとRLP_RLR_distT1_dribの違いを図11に示す．図6からも分かるように，RLP_RLR_distT1_dribはRLP_HCRに比べて，10秒以下のキープ時間が半数以下なっているのがここでも確認できる．また10〜90秒のキープ時間が占める割合もRLP_RLR_distT1_dribの方がRLP_HCRより多いことが分かる．しかし最高キープ時間については，RLP_RLR_distT1_dribが200秒台前半であるのに対し，図11には表示されていないが，RLP_HCRでは300秒以上のものも観測された．このことからRLP_RLR_distT1_dribは10〜90秒のキープ時間を安定して出せるのに対して，RLP_HCRはロングテール現象が起きており，10秒以下と100秒以上のキープ時間の割合が高いことが分かる．本研究の3つの提案手法の組み合わせが，10秒以下のエピソードが2割以下に減少したにもかかわらず，平均キープ時間が従来研究よりも上昇なかったのは，この従来研究のロングテール部分が影響しているためと考えられる．

5.2 各提案手法について

5.2.1 ドリブル実行のタイミング

ドリブルを追加して学習が確認できたプログラムについて，ドリブル実行時のパサーとT1の距離を測定した結果，6.0m〜8.0mに集中していることが分かった．またドリブル実行時の状況を観察したところ，パサーはドリブルを実行してT1を引きつけ，その間にレシーバに移動させてパスコースを作らせている振る舞いが観察された．この結果パスが通るようになり，Pass cutが減少したと考えられる．

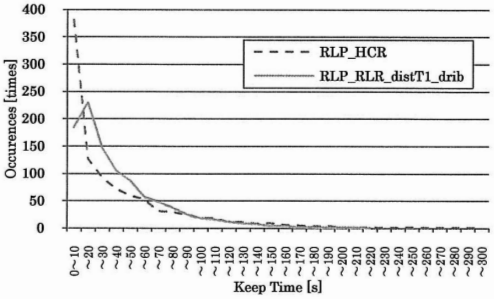

図11 タスク継続時間ごとの頻度の違い

5.2.2 レシーバへの強化学習の適用

RLRはdistT1と組み合わさらないと学習しないということが分かった．これはパサーが未学習のレシーバにパスを出す価値がないと初期段階で学習してボールを保持する傾向になるためと考えられる．distT1がない場合，パサーが「ボール保持が危険」であると学習できるのは，T1にボールを直接奪われたときのみである．distT1が組み込まれない限り，dribを追加したとしても，パスを出すようには学習することができない．そのため，レシーバが適切なポジショニングを学習したとしても，パサーはパスを出すことがないため，エピソード開始後すぐにT1にボールを奪われてエピソードが終了してしまうと考えられる．

5.2.3 各手法間の影響

distT1のPass cut増加の欠点をdribが埋め，distT1とdribに対応したレシーバの動きをRLRによって獲得し，RLRの学習をdistT1が促すというそれぞれの効果が影響し合うことで3つの手法がすべて機能した．この結果RLP_RLR_distT1_dribは，荒井らの手法であるRLP_HCRよりも10秒以下のキープ時間の発生回数を減少させ，さらに安定して10〜90秒のキープ時間を出すことができるようになった．

6 おわりに

本研究では，マルチエージェントの連続タスクとしてKeepawayを取り上げた．従来手法の問題点の分析から，パサーの学習において報酬関数とドリブルの追加を行い，レシーバに強化学習を適用した．これらの効果により，テイカーにボールを奪われてタスクが終了する割合の減少し，レシーバはパサーの動きに柔軟に対応できるようになった．各提案手法はそれぞれ負の影響が表れたが，すべての手法を

組み合わせることでその影響を解消し,3つの手法がすべて機能するようになった. この結果, 従来手法よりも 10 秒以下のキープ時間の発生回数を大幅に減少させ, 安定してキープ時間が10〜90秒になることを示した.

本研究では, レシーバの報酬関数に状態を組み込んでいるため, 20m×20m の領域で行う 3 対 2 の Keepaway タスクに依存度が高い. そのため今後の課題として, レシーバがボールを受けった場合に正の報酬を与えるなど, 領域の大きさやエージェント数に依存しない報酬設計を行う必要がある. また今回は10秒以下のキープ時間に着目して, 従来研究の問題点を解決するアプローチを取ったが, 今後は 100 秒以上のキープ時間に着目して, なぜ長時間キープが可能になるのかについても分析する必要がある.

参考文献

[Bond 88] Bond, A. H., Gasser, L. (eds.): Readings in Distributed Artificial Intelligence, Morgan Kaufmann Publishers, (1988).

[Noda 98] Noda, I., Matsubara, H., Hiraki, K., and Frank, I.: Soccer server : A tool for research on multiagent systems, pp. 233-250, (1998).

[Stone 01] Stone, P., Sutton, R. S.: Reinforcement Learning toward RoboCup Soccer, in Proceedings of 18th International Conference on Machine Learning, (2001).

[荒井 06] 荒井幸代, 田中信行: マルチエージェント連続タスクにおける報酬設計の実験的考察, 人工知能学会論文誌, pp.537-546, (2006).

[Sutton 96] Sutton, R. S.: Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding, Advances in Neural Information Processing Systems 8, pp.1038-1044, (1996).

[秋山 06] 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).