博士学位論文

自律型移動ロボットの協調行動における確率的行動決定手法に関する研究

指導教官　　石井 和男

北住　祐一

目次

図一覧

表一覧

第一章

序論

第一章　序論

1.1　研究背景

近年の自律型移動ロボットの発展に伴い, 日本におけるロボット市場の拡大は大きな期待が寄せられている．Fig. 1 に示すように新エネルギー・産業技術総合開発局（NEDO）は 2010 年 4 月 23 日に発表した“2035 年に向けたロボット市場の推計[1]”において，2025 年頃にサービスロボット部門が製造ロボット部門に匹敵する代表分野へと成長し, ロボット市場をさらに拡大させると期待している．ロボット市場の拡大を達成するうえで, ロボットに求められる重要な要素の一つとしてロボットの“自律性”があげられる．一般的に自律性を備えたロボットは，“自律ロボット”と呼ばれ, 特にロボット自身が移動機構を持ち, 自身の判断により行動するロボットは“自律型移動ロボット”と呼ばれる．ここで，自律ロボットの定義として，George A. Bakey によって述べられた「実世界の中で実体を持ち，（その多い少ないはあるにしても）外部からの明示的な人間の制御なしで，自分でタスクを実行することが出来る知的な機械」という定義[3]を引用し, 自律型移動ロボットの定義を「実世界の中で実体を持ち,（その多い少ないはあるにしても）外部からの明示的な人間の制御なしで, 自分の行動と判断能力によってタスクを実行することが出来る知的な移動する機械」として拡張する．

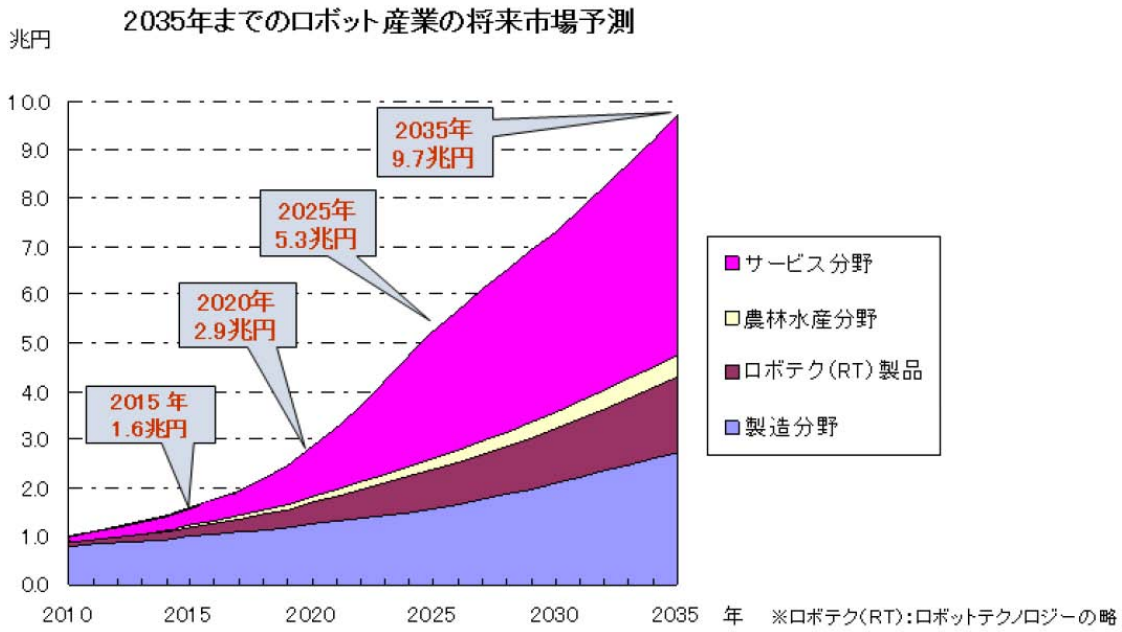


Fig. 1　The robot-market prediction by NEDO[1]

世界で初めて登場した自律型移動ロボットはスタンフォード国際研究所が1966年から72年にかけて開発したShakeyと言われている[2][3]．Fig. 2に示すShakeyは，外界センサとしてカメラとレーザレンジセンサを備え，無線通信を介して取得情報を外部のコンピュータへ送信し，“熟考型アプローチ”によって周辺環境の認識と次状態における最適な行動の推論を行った．Shakeyプロジェクトから45年が経過した現在では，Boston Dynamics社が開発したBigDogやHONDAのASIMO, iRobot社のRoombaなど, Fig. 2に示すように数多くの自律型移動ロボットが登場し，自律型移動ロボットはSF：Scientific Fictionの産物ではなくなった [4]-[9]．自律型移動ロボットの開発は，Shakeyから始まる45年間で大きな飛躍を遂げたと言えるが，開発の歴史を通して自律型移動ロボットは，大きく分けて以下の三つの能力が求められていると考えられる．

・未知の環境への適応能力
・複雑な作業の達成能力
・人間や他のロボットとの協調能力

自律型移動ロボットにこれらの能力を獲得させるためには，ロボットが動作する環境やロボットに求められる作業に内在する“不確実性”に対処する必要がある．本研究において述べる不確実性とは，ロボットがどんな環境で行動するのか，ロボットが遭遇した状況において何が最適な行動なのかなど，環境や状況判断に関する情報が事前に得られないことを示す．この不確実性に対する古典的なアプローチとしては，設計者が“起こりうる全ての状況”を事前に想定し，その全てをロボットにプログラムする方法があげられる．しかし，昨今のロボティクス分野の発展に伴い, 自律型移動ロボットに求められる動作環境や作業内容に含まれる不確実性は高まっており，設計者が想定可能な不確実性と，ロボットが対処すべき不確実性の規模には徐々に開きが生じている．ここで，自律型移動ロボット開発における不確実性をFig. 3のように定義する．Fig. 3に示すように，自律型移動ロボット開発における不確実性と，設計者によって事前の対処が可能な領域の差を“（設計者が事前に）想定困難な領域：Unexpected Area”とした．自律型移動ロボット開発は，“想定困難な領域”に対して，いかに有効なアプローチを行うかが重要な課題と言える． “想定困難な領域”への対処のように, 技術的に困難な課題への研究促進の手段として，ランドマークプロジェクトが挙げられる．よく知られた例としてアポロ計画等が挙げられるが，ロボティクス分野に繋がる人工知能研究においては, コンピュータチェスによって探索アルゴリズムやアーキテクチャが生み出され, ランドマークプロジェクトが技術促進に大きく貢献したと言える[46]．近年では，ロボット技術の発展を目的として，人工知能に加え，実時間性，実環境性，マルチエージェント性，動的要素などを含めたランドマークプロジェクト「ロボカップ」が世界規模で開催され，競技会を通じた技術促進が盛んに行われている．ロボカップサッカー中型リーグ（MSL：Middle Size Leagueと記す）は，サッカーという動的かつ瞬間的な判断力が必要となる難しい課題であり, 複数の自律型移動ロボットを用いた自律分散制御システムを構築する必要がある．

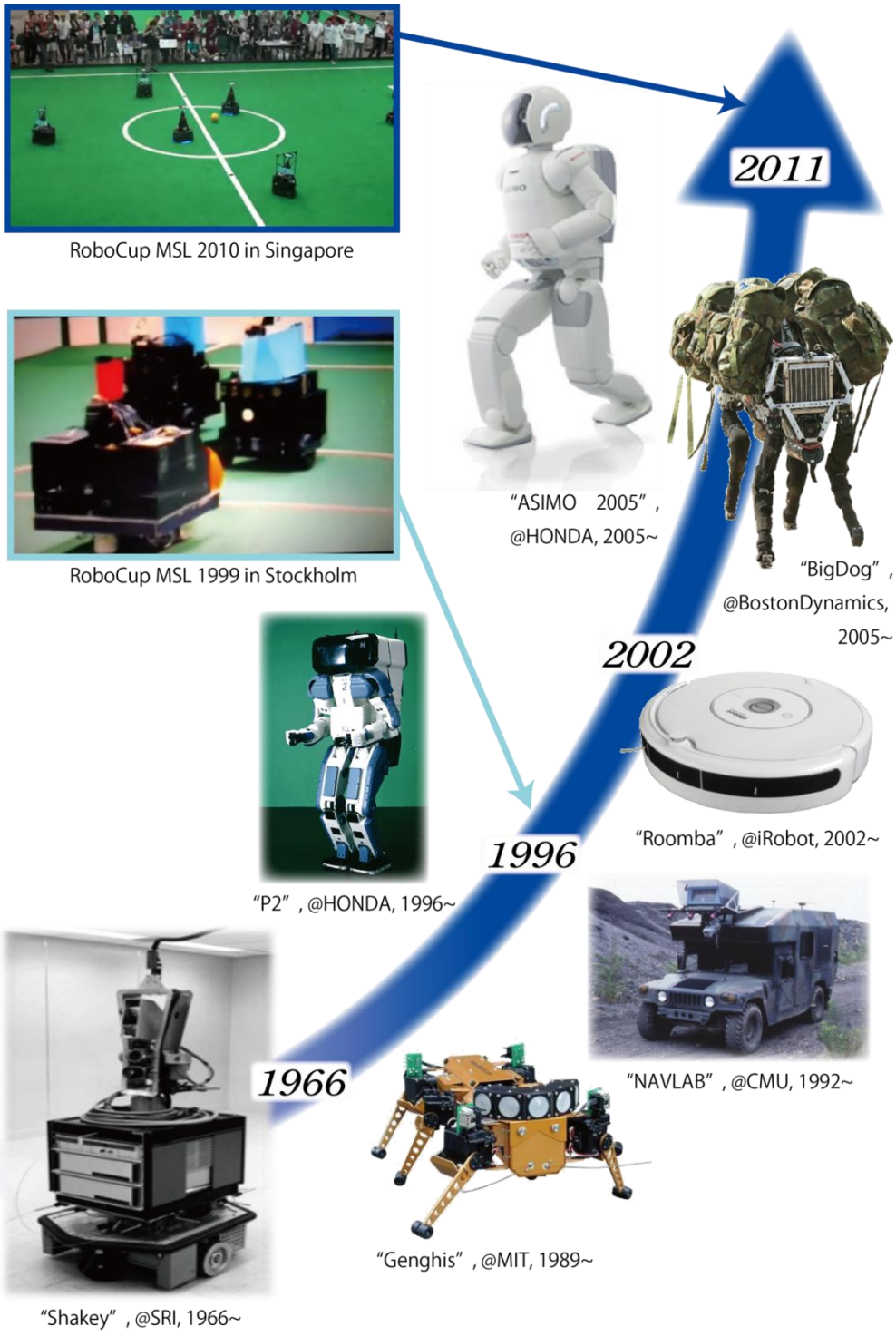


Fig. 2 Development History of Autonomous Mobile Robot

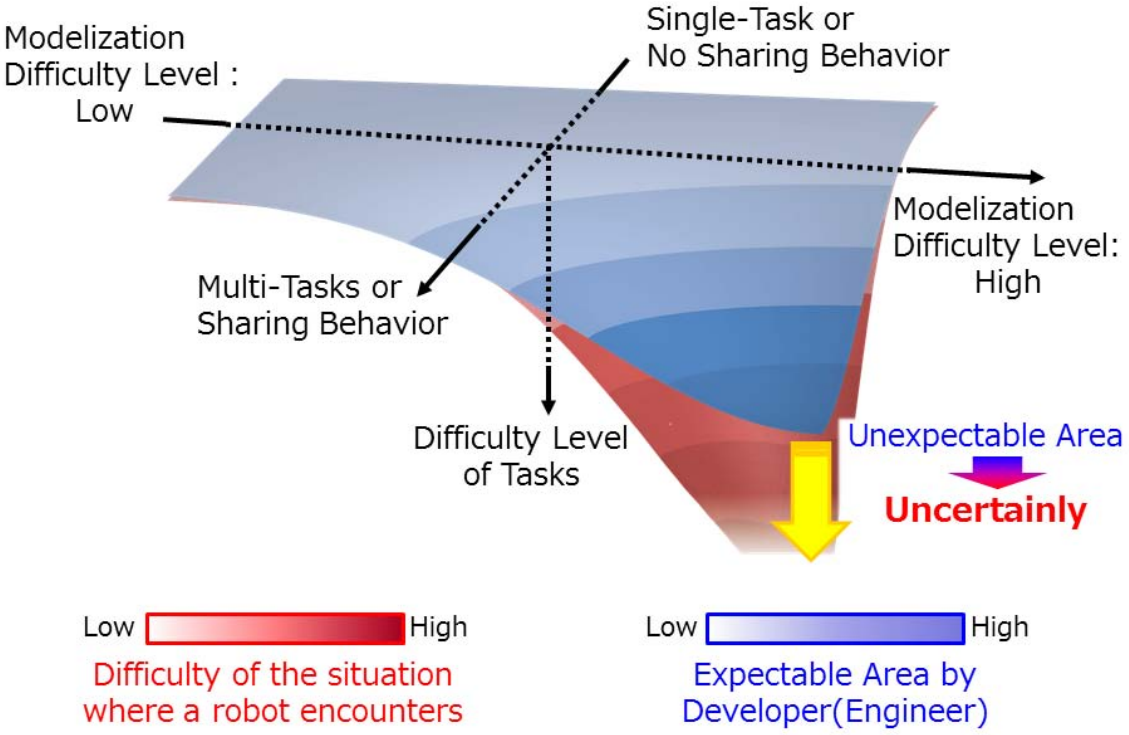


Fig. 3 Uncertainly Distribution in Autonomous Mobile Robot Operation Area

MSL では，ロボットの動作環境として“屋内の特設サッカーフィールドという既知環境（KnownEnvironment ）”を設定しているが，サッカーの試合を行う上では，サッカーフィールド内を複数のロボットやボールが移動し，周辺物体の位置が動的に変化することから，未知環境（Unknown Environment）としての要素も多く含まれている．従って MSL におけるロボットの動作環境は既知環境と未知環境の中間であると定義できる．各ロボットは，動的に変化する周辺環境に応じて，ボール取得やドリブル，障害物回避，シュートなどの様々な行動パターンから，自身にとって最適な行動を判断する必要がある．さらにロボット同士によるパス行動といった協調行動が求められるため，作業の複雑性（Multi Tasks or Sharing Behavior）は高い．これらの要素から，MSL におけるゴールを Fig. 3 に示すように配置した．MSL のゴールを目標点としてロボット工学技術全体を促進させることで，ロボットは動的環境への適応能力や複雑性を伴う作業への適応能力を獲得することとなり，実時間実環境下においてサッカーを十分に行えるロボットの能力は，搬送ロボットや清掃ロボット，空港や駅，デパート等，大勢の人が行き交う環境や家庭内環境におけるサービスロボットなど，様々なロボティクス分野において，十分な周辺環境認識能力，適切な行動判断能力，さらに他のロボットや使用者であるヒトとの協調動作能力等，自律型移動ロボット分野における大きな飛躍をもたらすと期待出来る．

1.2 本研究の目的

本研究では，MSLを検証用プラットフォームとして用い，ロボットの動作環境が含む不確実性への対処として強化学習を用いた行動獲得手法と確率ロボティクスを用いた情報の信頼性向上，及びロボットの協調行動における確率的行動手決定手法について提案する．

自律型移動ロボットが自律的に行動を獲得するためのアルゴリズムとして，強化学習が挙げられる．強化学習では，ロボットが周辺環境を探索し，環境に対して自身が持つ様々な行動パターンを試行することで，その結果得られた状態をフィードバックし，行動を最適化する[56]．特に，ロボットの動作環境に関する詳しい事前情報を必要とせず，ロボットによる試行錯誤によって行動を獲得可能である．つまり，強化学習は“環境情報に関する詳しい情報が無い”という不確実性に対して，自律的に適応させる学習法である．しかし，強化学習を用いた行動獲得では，一般的に膨大な数の学習回数が必要であり，実際のロボットを用いた試行錯誤による強化学習は現実的ではない[49]．また，この問題に対し，コンピュータ上のシミュレーション空間において強化学習を行い，学習結果を実機に適用する手法が提案されているが，量子化誤差の影響やマルコフ性の欠如によって致命的な性能低下を招く危険性があり[10]，自律型移動ロボットを対象とした強化学習研究の多くが，シミュレーション実験によるものとなっている．また，実機を用いた研究事例においても，センサ情報をおおまかに分割することで状態数を減少する手法[48]-[53]のように，状態数を可能な限り減らすことで学習時の負担を軽減させる手法を提案している．しかし，これらの手法は動的かつ複雑な環境を大雑把に観測していることになり，ロボットが獲得する行動は反射的なものになりやすい．つまり，これまで行われてきた自律型移動ロボットにおける強化学習研究は，

①理想的な環境・条件下におけるシミュレーション実験

②環境に関する観測情報を著しく簡易化した実機実験

に留まっている．そこで，本研究では，MSLにおけるゴールキーパロボットを対象とし，様々な入射角度，及び速度によって向かってくるボールに対する守備行動の学習について追究する．状態の定義として，絶対座標系上におけるロボットの位置と守備対象であるボールの位置・速度等を0.1[m, m/s]単位（守備対象であるボール直径の約半分）で離散化した状態を定義した．状態数が膨大であるため，学習はコンピュータ上に表現したシミュレーションによって行い，シミュレーションによって得られた学習結果を実ロボットの行動結果に反映させることで，実機のロボットに対する強化学習の適応可能性を評価した．本研究における強化学習研究の特徴は，ロボットの動作環境を細かく離散化することで量子化誤差の影響を抑え，実ロボットの観測情報を大雑把に分割することなく観測情報に応じた行動を獲得・実現できる点である．評価実験では，強化学習における行動罰や報酬のルール，方策の違いが学習結果に与える影響と学習結果を実機に適用した際の守備行動についての評価を行った．

次に, 実時間実環境下で動作する自律型移動ロボットが自身の行動を正しく実行するためには, 観測情報から自身の状態を正確に推定することが重要である．しかし，実時間実環境下におけるロボットの観測情報には，磁場や照明条件，移動に伴う環境の動的な変化などの様々な要因により誤差が含まれる．特にロボットの自己位置情報は，障害物回避や経路計画，行動判断アルゴリズムにおいて最も重要かつ基本的な情報であり, 自己位置情報に生じる誤差はロボット単体の基本的な行動を始め，複数のロボットによる協調行動の実現等において妨げとなる．このように不確実性を含んだ推定情報から自身の正確な状態を推定するためには, 情報に含まれる不確実性を数値的に表現し，情報の持つ信頼性を評価する必要がある．ここで，不確実性を数理的に表現する手法の一つとして確率ロボティクスがあげられる．確率ロボティクスでは，情報の信頼性を確率分布によって表現することで, 誤差が含まれる観測情報から信頼性の高い情報を推定することが可能である[68][69]．本研究では，MSLロボットにおける自己位置情報に焦点を当て，複数のロボット間で共有した情報に基づいた自己位置推定アルゴリズムを提案した．提案手法では，画像処理による自己位置推定を基本とし，MSLのフィールドに 1 つだけ存在するボールを全ロボットにおける共通のランドマークとしてランドマーク情報を共有する. さらに複数のロボットからのランドマーク情報を統合してランドマーク位置に関する推定精度を向上させるとともに, 各ロボットにランドマーク情報をフィードバックさせ自己位置推定の精度を高める．

1.3 本論文の構成

本論文では 2 章において研究対象である RoboCup サッカー中型リーグの歴史や，実機検証のためのプラットフォームとして用いたサッカーロボット“Musashi”について述べる．3 章では，動的環境に対するロボットの自律的行動獲得の手段として，ゴールキーパロボットを対象とし，Q 学習による守備行動獲得アルゴリズムの提案を行った. 研究対象であるゴールキーパロボットは連続空間内において守備行動を行うが, 本章では連続空間を離散化したシミュレーション空間内において守備行動を学習し, その学習結果について動力学シミュレーションと実機を用いて評価した．また，学習アルゴリズムの評価に関して，報酬関数や方策が行動獲得に与える影響について検討する． 4 章では，複数のサッカーロボットが観測した情報に基づき，ランドマークの位置を確率的情報共有によって推定する手法と, 推定したランドマークの位置を基準とした自己位置推定アルゴリズムについて提案する．提案手法の有効性を検証するため，実際にサッカーロボット 5 台を用いた実験を行い，その結果について議論する．5 章では，MSL におけるパス行動に着目し，周辺環境や戦略を確率分布と条件付けによって表現することで，環境が動的に変化する MSL の環境において適切なパスの目標位置を決定する手法について提案する．提案手法の評価では，動力学シミュレーションを用いて，プラットフォームとして利用したサッカーロボットの行動アルゴリズムと提案したパス行動を追加したアルゴリズムの比較実験を行った.

最後に本論文において提案した強化学習による行動獲得と確率的情報統合手法について考察を行い，結論を述べる.

第二章

サッカーロボット“Musashi”

第二章　サッカーロボット”Musashi”

2.1　ロボカップサッカー中型リーグにおける協調行動研究の歴史

自律型移動ロボットの研究開発に逸早く着目し,現在では世界規模で展開している自律型移動ロボット競技会の一つに RoboCup が挙げられる[11]-[15]．RoboCup は，サッカーを題材とした RoboCup Soccer や，災害救助用ロボットツール開発を目的とした RoboCup Rescue，若手研究者の育成を目的とした RoboCup Junior，人間との共生を目指す RoboCup @home など多くのプロジェクトから構成される．特に，RoboCup Soccer 中型リーグ(MSL)は，人間とサッカーを行える程度の大きさを持つロボット(最大寸法:500x500x800[mm])と RoboCup で最も大きなフィールド(フィールド寸法 : 18x12[m])を用いており，観客は迫力ある試合を見て，楽しむことが出来る．MSL では，各々のチームから 5 台のロボットを試合にエントリー出来る．それらのロボットは自律分散型制御でなければならない．それゆえ，全てのロボットは動的環境下で行動し，強調動作を行うことが求められる．13 年間続くロボカッププロジェクトでは，このマルチエージェントシステムに関わる問題に対し，多くのチームが様々な研究が進めている．

i.　1999 年当時のロボカップ中型リーグ

今日までの13年間を評価し，今後の動向を予想するため，1999年ストックホルムで開催された第3回世界大会についての批評を基に1999年当時のMSLを評価する[16]．第3回大会はシミュレーションリーグ，小型リーグ，中型リーグ(MSL)の3つのリーグから構成された．当時のMSLは，各チームからロボットを4台ずつエントリーし，計8台で試合を展開した．試合用フィールドの寸法は9.0x5.0[m]であり，コーナーキック，ゴールキックなどのセットプレイは公式ルールとして適用していない．また，RoboCup MSLにおける技術発展を促すテクニカルチャレンジが，ロボット単体によるボールコントロール(ボール探索，取得，パス，キック)を課題として設定していたことからも，当時のMSL全体はサッカーを行うための基本的な行動の実現が達成されていない．2000年当時のSérgio Monteiroらは，MSLにおける協調行動実現のために，以下のような提案をしている．

- 外部 PC，もしくはゴールキーパをオンラインコーチとして，各味方ロボットからの情報を集め，コーチとなるエージェントが次の行動を定め，チームメンバーに指令を出す
- 各ロボットに独自の行動(前衛左の FW と前衛右の FW などのような役割を与える)を行わせる．
- ワールドモデルを行動決定アルゴリズムのサポートとして導入し，障害物検出やパス軌道生成などを簡略化する．

また，行動決定については状態マシンや行動決定樹形図を用いることが進められている．例としてマルコフ決定過程を用いた確率的行動生成法やファジー評価法，システム動的基底(Dynamic Systems Based)などが挙げられている[17][18][19]．また，学習アルゴリズムの現状についても述べており，1999年当時，MSLではほとんどのチームにおいてソフトウェアが複雑化するという理由から学習アルゴリズムを導入せず，多くのチームが事前に作成したマップを用いて行動生成を行っていたと述べられている．少数ではあるが，学習アルゴリズムが用いられている例として強化学習の適用があり，敵ロボットの特定の行動に対して行動を分類する．

ii. 2005 年当時のロボカップ中型リーグ

2005年は日本・大阪において第9回世界大会が行われ，この頃には既にRoboCup Rescue，RoboCup Juniorなどが世界大会の正式種目として組み込まれており，SoccerではHumanoidリーグにおいて自律型二足歩行ロボットがサッカーとしての試合展開を実現し始めていた．
MSLでは，試合にエントリーできるロボット数が4台から6台となり，フィールドサイズを12x8[m]に拡張した環境で試合を展開した．この頃から，MSLではいくつかのチームがロボット6台による“組織戦”を実際の試合で実現するようになった．以下に2005年世界大会において2連覇優勝を果たした慶応大学のチーム「EIGEN」が用いていた協調行動アルゴリズムについて紹介する．続いて，学習アルゴリズムを用いた協調行動創発の一例としてInstitute for Systems and Roboticsのチーム「ISocRob」から提案されたミニマックス価値反復法について紹介する．

ii-i. タスク達成度評価に基づく役割決定手法

藤井らは，エージェントが実行する攻撃，守備，補助の三つのタスクに対し，それぞれのエージェント自身が行動の達成度を自己評価し，他の味方ロボットと評価値を比較することによって，自身の役割を決定する手法を提案した[20]. Fig. 4に藤井らによって提案された手法のコンセプトを示す．各エージェントは，実行中の行動に対し，式(2.1)と式(2.2)を用いてその行動と状態に関する自己評価値を計算する．

$$SystemSatisfaction_i = \sum_{k=1}^{m} Evaluate_i\left(Agent^{k}_{priority}\right) + \sum_{j=1}^{n} E_i\left(Agent_j\right) \tag{2.1}$$

$$\begin{aligned} &if \quad Evaluate_i(Agent_j) > Evaluate_i(Agent_{own}) \\ &\qquad E_i\left(Agent_j\right) = Evaluate_i\left(Agent_j\right) \\ &else \quad Evaluate_i(Agent_j) \leqq Evaluate_i(Agent_{own}) \\ &\qquad E_i\left(Agent_j\right) = 0 \end{aligned} \tag{2.2}$$

ここで,

Self-Evaluation：各エージェントが各自算出する自身の行動に関する自己評価値

System Satisfaction$_i$：目的達成のための評価指針と満足度．各エージェントによって異なる値

System Objective$_i$：各エージェントにおいて共通な目的達成を望む度合い

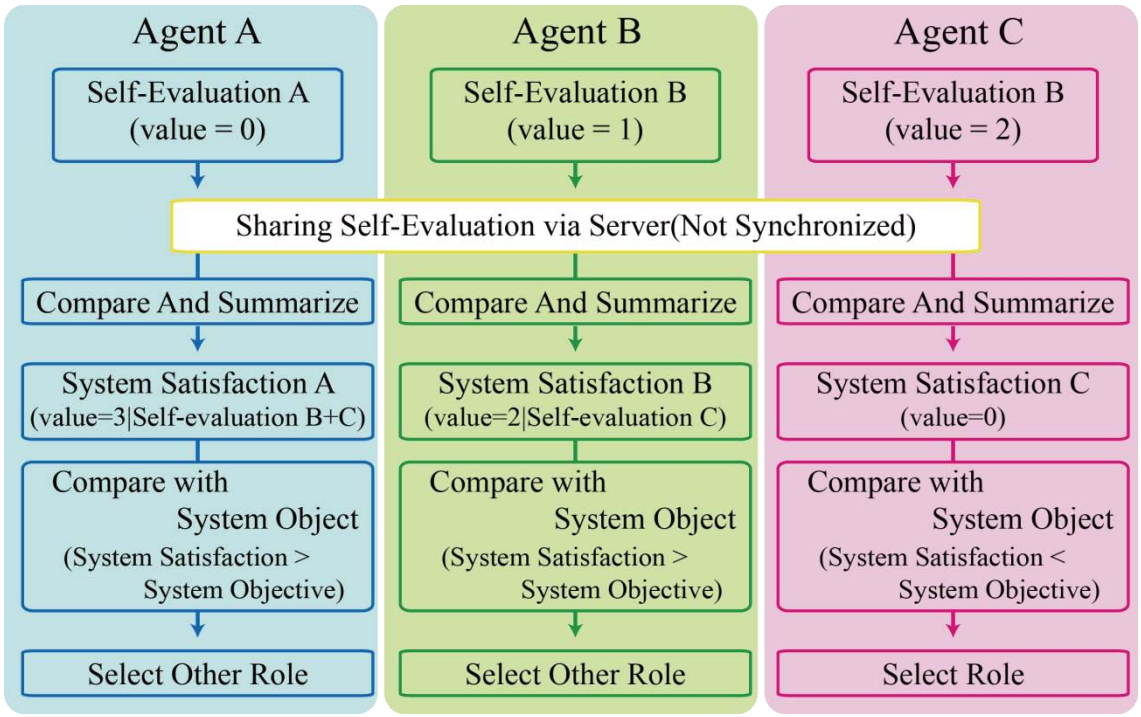


Fig. 4 Concept and the flow of the proposed method

式(2.1)と式(2.2)によって表される自己評価値はマルチエージェントシステムとして達成すべき目標が単一である場合を想定している．しかし，RoboCup MSLのような環境では，各エージェントが達成すべき目標は複数存在する. 例として得点を得ることをチームとして優先すべき場合や失点を防ぐことを優先すべき場合などが上げられる. 藤井らはこのように複数の達成目標が存在するマルチエージェントシステムに対応するため, 式(2.3)によって達成すべき目標選択手法を提案した.

$$v_i = SystemObjetive_i - SystemSatisfaction_i \tag{2.3}$$

v_i：エージェントの自己評価値と目的達成状況との差．目的に対する“達成願望レベル.”

藤井らの手法は, シミュレーション環境中において各エージェントが評価値に応じて役割を更新する結果を実現している．また，藤井らが提案した協調行動アルゴリズムがMSLにおいてチーム戦略として効果的であることは，EIGENが2004・2005年と2連覇を成し遂げている事実からも明らかである.

ii-ii.　ミニマックス価値反復法を用いた行動決定手法

Gonçalo NetoらはMSLの試合を“2人の零和ゲーム(Two-Person, Zero-Sum Game)”と呼ばれる確率的ゲームのモデルに当てはめ，ナッシュ均衡の探索と動的計画法により，チームが最悪のケース(失点により敗北する状況)に陥ることを防ぐ行動方策の創発手法を提案した[21].

Gonçalo Netoらの手法では，行動創発のためのアルゴリズムとして強化学習法が用いられている．報酬の最大化に基づいて行動選択を学習する強化学習法は，マルコフ決定過程(MDP)が成立する環境において適用可能である. 強化学習の基本的なアルゴリズムとして動的計画法を挙げると，報酬最大化に基づく状態価値マップの更新式は式(2.4)で表される.

$$V_{k+1}(s)=\max_{a}\sum_{s'}T(s,a,s')\left[R(s,a,s')+\gamma V_{k}(s')\right] \tag{2.4}$$

$V(s)$：状態価値　　T：遷移関数

R：報酬関数　　A,a：実行可能な行動集合

O,o：他の行動集合　　γ：割引率

s・s': 現在の状態・行動実行前の状態

式(2.4)は，ベルマン最適方程式に基づいた価値反復法としても知られている．また，MSLの試合のように複数のプレイヤーが1つの関連した報酬構造を持つようなゲーム体系をマトリクスゲームと呼ぶ. マトリクスゲームは，ゲームに参加しているプレイヤー全員が少なくとも一つのナッシュ均衡を持つという性質を持つ. ナッシュ均衡を保つ戦略を用いることで最悪のケースを回避可能であることが保障される．Gonçalo Netoらが提案した手法は，マトリクスゲームにおけるナッシュ均衡を見つけるための手段として式(2.5)に示すミニマックス演算子を用いている.

$$\max_{\sigma\in PD(A)}\min_{o\in O}\sum_{a\in A}\sigma(a)R(a,o) \tag{2.5}$$

σ：戦略

多重状態環境下で多数のエージェントを扱う確率的ゲームは，マトリクスゲームとMDPの拡張と考えることできるため，マトリクスゲームとMDPの両特性を持つ．重要な特性の一つとしてゲームの各中間点においてナッシュ均衡の存在を保障していることが挙げられる. ゲームの各中間点でナッシュ均衡を見つけることは，最適方策を見出すことと同じである．Gonçalo Netoらの手法では，このようなマトリクスゲームとMDPの両特性を持つゲームにおける最適方策を式(2.6)と式(2.7)により導出する.

$$V^{*}(s)=\max_{\pi\in PD(A)}\min_{o\in O}\sum_{a\in A}\pi(a)Q(s,a,o) \tag{2.6}$$

$$Q(s,a,o)=\sum_{s'}R(s,a,o,s')+\gamma T(s,a,o,s')V^{*}(s') \tag{2.7}$$

式(2.6)と式(2.7)による最適方策の導出は，確率的ゲームに参加するプレイヤー数が2人である場合に限定される. したがって, MSLの試合モデルを2人の零和ゲームに当てはめることで式(2.6), (2.7)による最適方策の導出が可能となる．2人の零和ゲームとは，2人のプレイヤーの内，片方の利益が他者の損失となり，全体の損得の和が常にゼロである場合に，自分の利益を最大にするような手を選ぶゲームを指す．Gonçalo Netoらは，ゲームに参加するプレイヤーをMSLにおけるエージェント単位ではなく，チーム単位で考えることでサッカーの試合を2人のプレイヤーによるゲームとして扱った. また, 報酬体系に関しても, “何点獲得したか”ではなく, “得点したか(+1), 失点したか(-1)”という採点を行い, 報酬は常に全てのチームメイトに同等の報酬を更新することで，MSLの試合モデルを2人の零和ゲームのモデルに当てはめた.

Gonçalo Netoらは，提案したアルゴリズムの検証として，シミュレーション上において2台対2台のロボットによる試合(i.e. Aチームは式(2.6)，(2.7)によってミニマックス価値反復法による行動創発を行い，Bチームはランダムに行動を選択する)形式の評価を行った. シミュレーション結果は, 各エージェントがディフェンス重視の戦略選択を行ったと報告している. またシミュレーション上における試合結果は, ミニマックス価値反復法を用いたチームとランダムに行動を選択するチームとの対戦において, ミニマックス価値反復法を用いたチームが約90％の確率で勝利したことも報告している.

iii. 2010 年におけるロボカップ中型リーグ

第 1 回世界大会以来 13 年目となる 2010 年 6 月は，シンガポールにおいて第 13 回世界大会の開催を予定している. MSL では, 2007 年に追加されたルールによりフィールドサイズは 18x12[m] となり，2008 年に追加されたルールによりそれまでランドマークとして与えられていたゴールの色(青・黄)が廃止され，人間のサッカーと同じように敵・味方のゴールが全く同等の形状・色(白)となった. 2009 年には，各セットプレイ時にロボット 2 台によるパス動作から試合を始めることが義務付けられた. 2010 年は, これまで用いられてきたオレンジ色の公式球の色を廃止し, 大会時に委員会が提示する色のボールを用いて試合を行うことが義務付けられている.

次節に 2010 年における協調行動研究の最先端例として，2009 年の第 12 回世界大会における MSL の優勝チーム「RFC Stuttgart」と同大会で 2 位を獲得したチーム「Tech United Eindhoven 」の研究事例を述べる.

iii-i. RFC Stuttgartの協調行動研究例

Fig. 5 に RFC Stuttgart(以下，RFC と記す)の開発したロボット“RFCBot”のシステム構成図を示す[22]. RFCBot はセンサデータと味方ロボットと共有したデータ，センサデータの履歴情報を基にワールドモデルを構築する．共有する情報とは，検出した物体の位置情報と，物体位置情報の共有により構築したワールドモデルを基に決定した各ロボットの役割への同意(相互確認)情

報を指す. RFC はワールドモデル構築において, 全味方ロボットと情報を共有することにより, ノイズを低減している. 精度の高いワールドモデルの獲得は, 複数台のエージェントによる高度な協調行動の実現や, 円滑な役割決定の実現を意味すると考えられる. 各ロボットはワールドモデルを基に, Player Decisions 部においてアタッカー, ディフェンダー, サポーターなどの 6 種類の役割を決定する. Navigator Actions 部では, 決定された役割に応じてドリブル, ボールゲットなど 8 種類の行動を決定する. Pilot Motor control 部では, 行動決定に従いモータへコマンドを送信する役割を担う. RFC は Fig. 5 に示すソフトウェアアーキテクチャにより, 精度の高いワールドモデルを構築することに成功していると考えられる. 事実, 2009 年の第 12 回世界大会では, 正確なポジショニングにより各セットプレイ時のパスワークを迅速かつ円滑に成功させている.

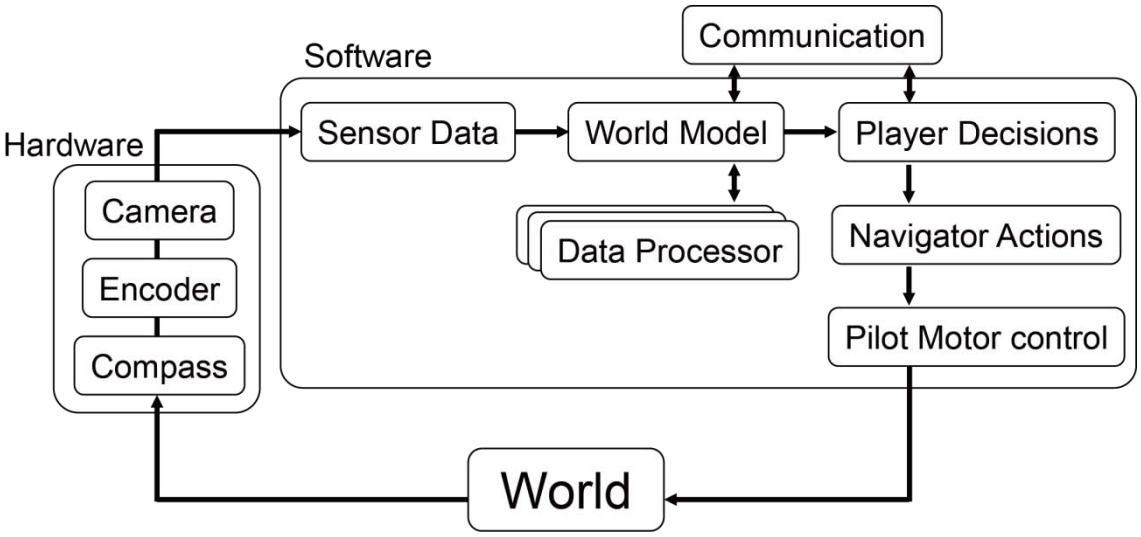


Fig. 5 Architecture and depicted data flow of the RFC Stuttgart robot software

iii-ii. Tech United Eindhovenの協調行動研究例

Fig. 6 に Tech United Eindhoven (以下, Tech United と記す)が開発したロボット“TURTLE”のシステム構成図を示す[23]. TURTLE のソフトウェア構成は Vision, World model, Motion のモジュールから構成され, Vision, World model モジュールは 30[Hz], Motion モジュールは 1000[Hz]で実行される. Tech United が 2010 年に行った新たな取り組みとして, World model モジュールにおける仲間ロボットとの情報共有が挙げられる. World model モジュールでは, 仲間ロボットと共有した情報を用いることにより, フィールド上に存在する全ロボットに関して敵ロボットと味方ロボットを区別し, 同時に, 敵・味方を含めた全ロボットの自己位置と移動速度を推定している. 各ロボットの状態推定には, 仮説樹形図とカルマンフィルタを用いて推定を行い, 推定確率に応じて各ノードの刈り込みを行っている. また, Tech United は推定したロボットの状態推定値を基にファジー推論を用いたパスルートの推定や, ベジェ曲線を用いた“敵ロボットへの衝突回避を考慮した経路計画”などを開発している.

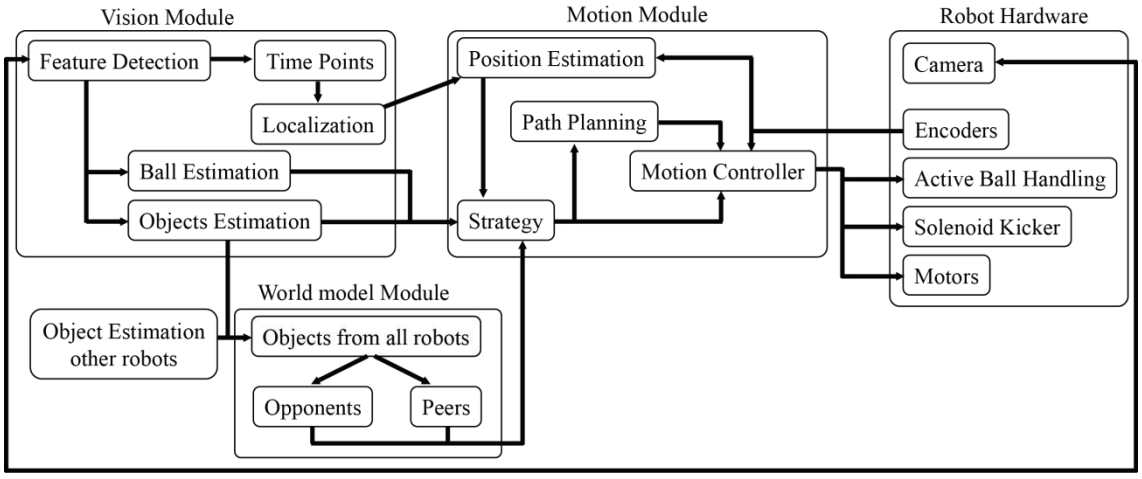


Fig. 6 Schematic representation of the robot's hardware and software modules and their int erconnections

iv. 予想される今後の動向

RFC と Tech United の例より，2010 年において各チームは，ロボット間で自己位置情報や実行中の役割，検出した物体情報などを共有化し，高精度なワールドモデルを構築することを基礎としてロボットの役割・行動決定を行っている．これは 2007 年の追加ルールによりフィールドサイズが拡張され，各ロボットが搭載する全方位カメラの情報だけではフィールド全体の状況把握に限界が生じたこと，そして MSL がロボット単体の性能向上だけでは，ゲームに勝てないほど技術的に発達したことが原因として挙げられる．

2009 年の第 12 回世界大会では，最大速度 5.0[m/s]で走行するロボットが正確なパスワークとボールコントロールにより，ゴールキーパとゴール上方バーの間(約 20[cm]の空間)へのシュートを成功させる場面なども見られた．人間の走る速さを 7.0[m/s]程度(50[m]走を 7.0[sec]で走る場合を想定)とすると，サッカーという持久力を要する競技に必要な移動速度として 5.0[m/s]という速度は十分であろう．

サッカーを行う上で，ロボット単体の運動能力が人間のレベルに達しつつあるとすると，今後，ロボット同士による協調行動が MSL の中心となるテーマである．実現可能なアプローチとして，味方ロボット間での情報共有により敵・味方ロボットの判別を行い，マンツーマンディフェンスによって失点を防ぐ協調行動の実現や，ワールドモデルを用いてパスワーク計画を算出することで，パスワークによってディフェンダーを回避する行動などの実現が挙げられる．

また近年，世界大会においてヨーロッパ諸国のチームがランキング上位を占める中，日本国内の MSL に目を向けると，残念ながら参加チーム総数が減少傾向にある．MSL は第 1 回大会から既に 13 年が経過しており，技術的に新規参入が難しいリーグである．また，大型のロボットを複

数台開発しなければならない背景や, 実験のために広大なフィールドを用意する必要があることからも新たに参入するのは困難である. 一方で, 様々なリーグを持つ RoboCup Project の中でも, MSL は，人間とロボットによるサッカーの試合実現について，技術的に最も近いリーグであると言える．このように高い研究価値を持つ研究分野を継続的に展開していくためには, 新規参入チームへの技術開示や既に参戦しているチーム同士の技術交流が盛んに行われる必要がある．

2.2 サッカー中型リーグチーム“Hibikino-Musashi”

“Hibikino-Musashi”は北九州学術研究都市を拠点としている Robocup サッカー中型リーグのチームである．九州工業大学大学院生命体工学研究科，北九州市立大学国際環境工学部，北九州学術推進機構ロボット開発支援室の三つの組織によって構成されており､強力なキック力を持つロボットが特徴である．大会成績は,06 年北九州日本大会優勝，ブレーメン世界大会ベスト 8, 07 年大阪日本大会準優勝，アトランタ世界大会 4 位である．

2.3 サッカーロボット“Musashi”

Musashi は，北九州学術研究都市の MSL プロジェクト“Hibikino-Musashi”において開発されたサッカーロボットである．MSL では，瞬時に周辺の状況を把握し，即座に任意の方向へ移動することで, より効率良く試合を進めることが出来る観点から，多くのロボットが全方位移動機構と全方位カメラを用いており，Musashi においても同様の機構を採用している．

2.3.1 全方位移動機構運動学

“Musashi”は全方位移動機構を搭載しており，120[deg]間隔でDCモータ，ギアボックス，オムニホイールの駆動系コンポーネントを三基，ロボット本体底部のプレート上に配置している．“Musashi”の移動機構部を2次元平面上で移動する剛体としたとき，“Musashi”の自由度は直進移動，並進移動，回転の3自由度である．以下に各車輪の移動速度からロボットの各方向への移動速度を計算する過程を示す．Fig. 7に地面上を“Musashi”が移動する際の，各車輪の速度ベクトルと“Musashi”の移動速度ベクトルの関係図を示す．

“Musashi”が移動する平面を絶対座標系(O_w-$x_w y_w$)とし，“Musashi”の重心を原点とした移動座標系(O_m-$x_m y_m$)とする．以下，添え字wは絶対座標系を表し，添え字mは移動座標系を表すとする．絶対座標系に対する移動座標系の傾きをϕ, 各車輪の半径をr, 各車輪の角速度をω_i, 速度ベクトルをv_i, “Musashi”の絶対座標系に対する移動速度をV_wとし，“Musashi”の重心から車輪までの距離をLとする．

ここで，各車輪の速度ベクトルは以下の式(2.8)より式(2.9)～(2.11)によって表される．

$$v_i = r\omega_i \tag{2.8}$$

$$r\omega_1 = -\frac{1}{2}\dot{x}_m + \frac{\sqrt{3}}{2}\dot{y}_m + L\dot{\phi} \tag{2.9}$$

$$r\omega_2 = -\frac{1}{2}\dot{x}_m - \frac{\sqrt{3}}{2}\dot{y}_m + L\dot{\phi} \tag{2.10}$$

$$r\omega_3 = \dot{x}_m + L\dot{\phi} \tag{2.11}$$

“Musashi”の各方向の速度ベクトルをx'_w，y'_wとし，回転速度ベクトルをϕ'_wとし，このときの重心速度ベクトルP'_wを

$$P'_w = \begin{bmatrix} x'_w & y'_w & \phi'_w \end{bmatrix}^T \tag{2.12}$$

と定義する．ここで，絶対座標系から移動座標系への変換を以下の行列によって行う．

$${}^wT_m = \begin{bmatrix} {}^wC_m & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi & x_m \\ \sin\phi & con\phi & y_m \\ 0 & 0 & 1 \end{bmatrix} \tag{2.13}$$

これより，移動座標系での重心速度ベクトルP'_mは,

$$P'_w = {}^wT_m P'_m \tag{2.14}$$

とあらわされる．ここで各車輪の移動速度qと角速度ω_iの関係は式(2.15)であらわされる．

$$q = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} r\omega_1 \\ r\omega_2 \\ r\omega_3 \end{bmatrix} \tag{2.15}$$

これらの関係式をベクトル表現を用いてまとめると

$$P'_m = J_m q \tag{2.16}$$

となる．ここでJ_mはヤコビアンであり，以下のようにあらわすことができる．

$$J_m = \begin{bmatrix} -1/3 & -1/3 & 2/3 \\ \sqrt{3}/3 & -\sqrt{3}/3 & 0 \\ 1/3L & 1/3L & 1/3L \end{bmatrix} \tag{2.17}$$

よって，各車輪の速度は，

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\cos\phi - \frac{\sqrt{3}}{2}\sin\phi & -\frac{1}{2}\sin\phi + \frac{\sqrt{3}}{2}\cos\phi & L \\ -\frac{1}{2}\cos\phi + \frac{\sqrt{3}}{2}\sin\phi & -\frac{1}{2}\sin\phi - \frac{\sqrt{3}}{2}\cos\phi & L \\ \cos\phi & \sin\phi & L \end{bmatrix} \begin{bmatrix} \dot{x}_w \\ \dot{y}_m \\ \dot{\phi} \end{bmatrix} \tag{2.18}$$

によって，求めることが出来る．この式から各車輪の移動速度だけで“Musashi”の各方向へ移動する移動速度を制御することが出来る[24].

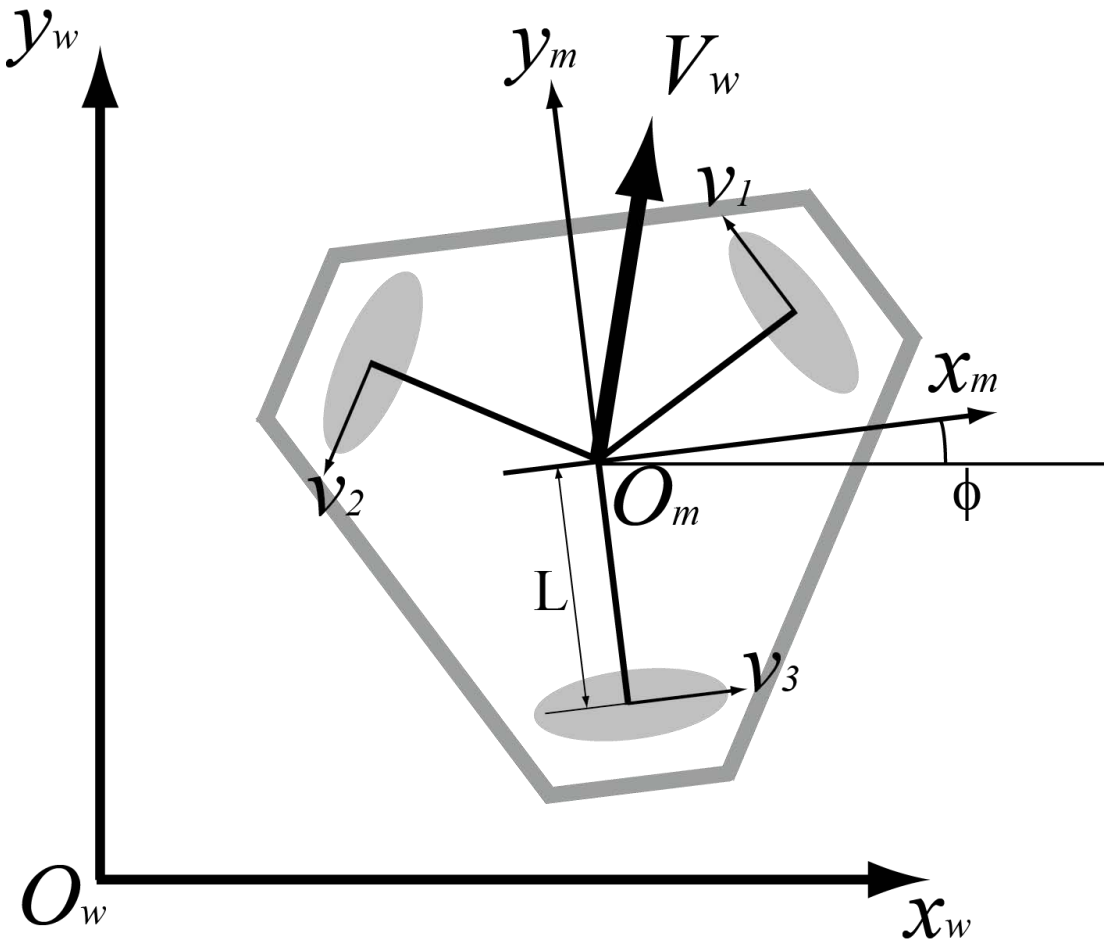
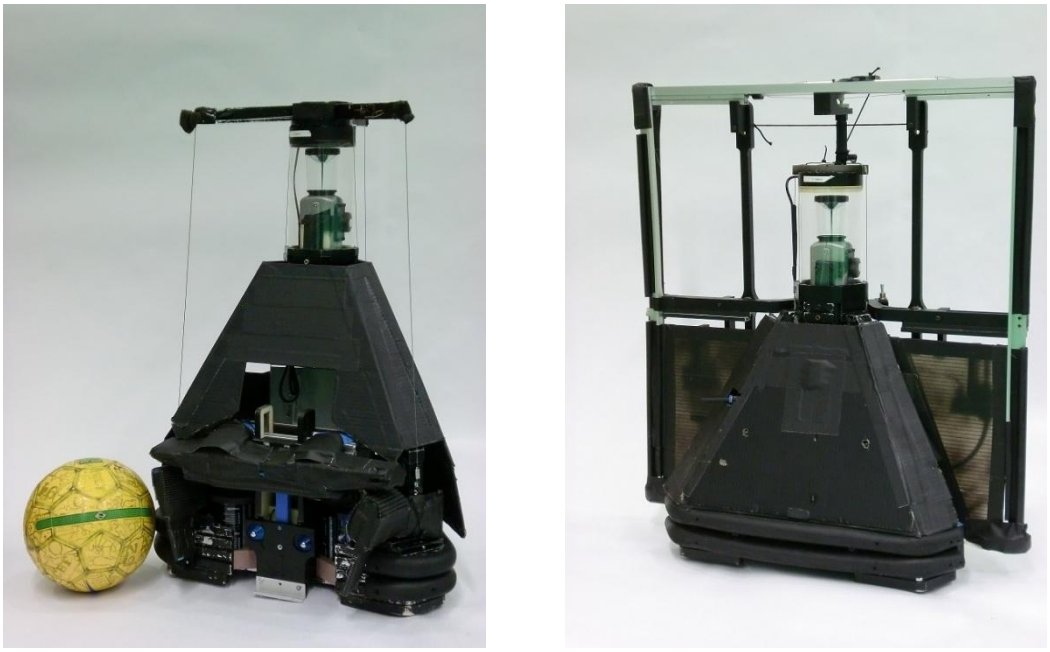


Fig. 7 Dynamics of Omni Directional Moving Mechanism

2.3.2 ハードウェア構造

Fig. 8に示す“Musashi”は，信頼性や頑強性，メンテナンス性の向上を目的とし，各機能を容易に着脱出来るモジュール単位に分けたモジュール構造を持つ[25]-[29]. Table 1にMusashiの仕様，Fig. 9にMusashiのモジュール構成を示す．モジュール構造の採用により，何らかのトラブルによってロボットが故障した際においても，操作者が故障したモジュールを新しいモジュールに交換することで，円滑にロボットの機能を修復することが出来る．

Fig. 9に示すようにMusashiの全方位移動機構は，全方位車輪と70 [W] DCモータ，500 [pulse/rot.] のエンコーダを1モジュールとする“Drive Module”を120 [deg] 間隔で配置した“Base Module”により実現しており，最大3.0 [m/s] での移動が可能である．



(a) (b)

Fig. 8 Field Player Robot and Goal Keeper Robot of “Hibikino-Musashi”

Table 1 Specification of Soccer Robot "Musashi"

Item	*Details*
Dimensions	Triangle × Hegiht 520[mm] × 800[mm]
Weight	24[kg]
Motor	Maxon DC Motor 70[w] × 3
Wheel	Omni wheel × 3
Motor Driver	Faulhaber MCDC 2805 × 3
Battery	Main Battery
	Li-Polymer Battery 25.9[V], 2000[mAh]
	or Ni-H Battery 25.9[V], 2800[mAh]
	Kicking Device Battery
	Li-Ion Battery 14.4[V], 8000[mAh]
Sensor	Omni Camera, Digital Cmpus

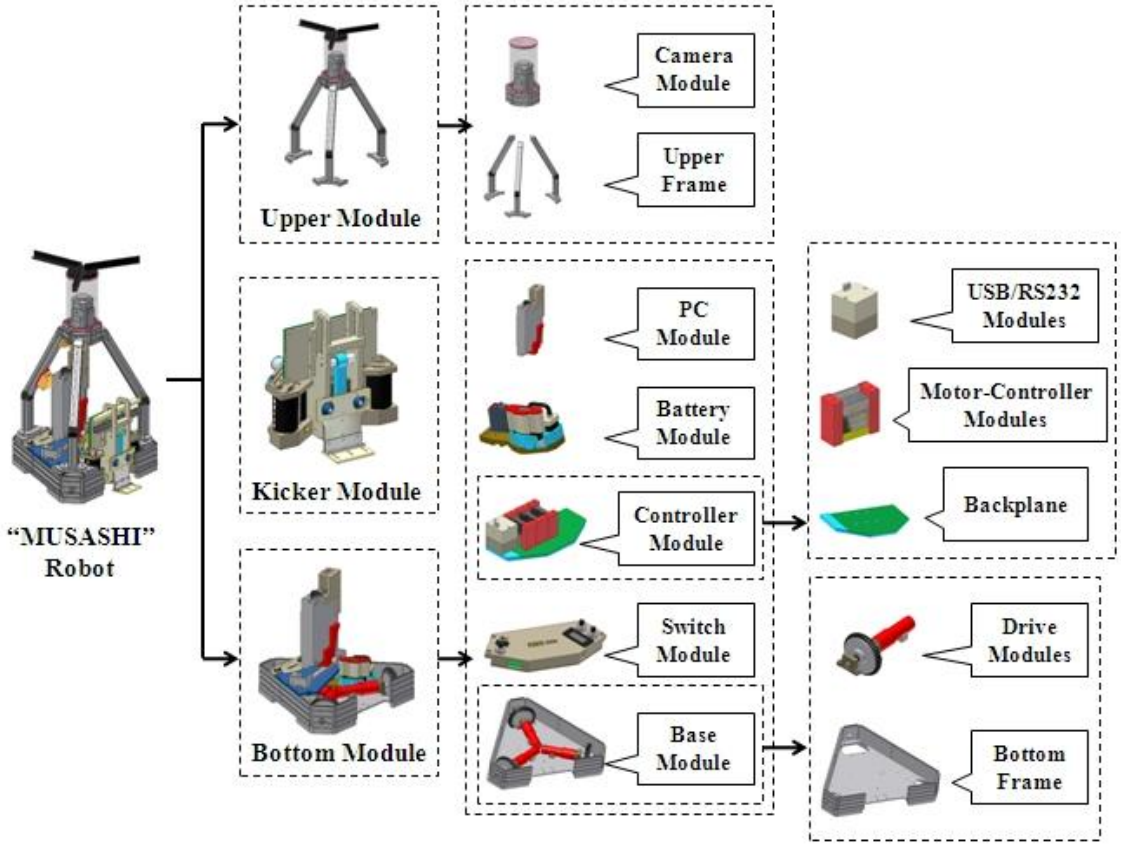
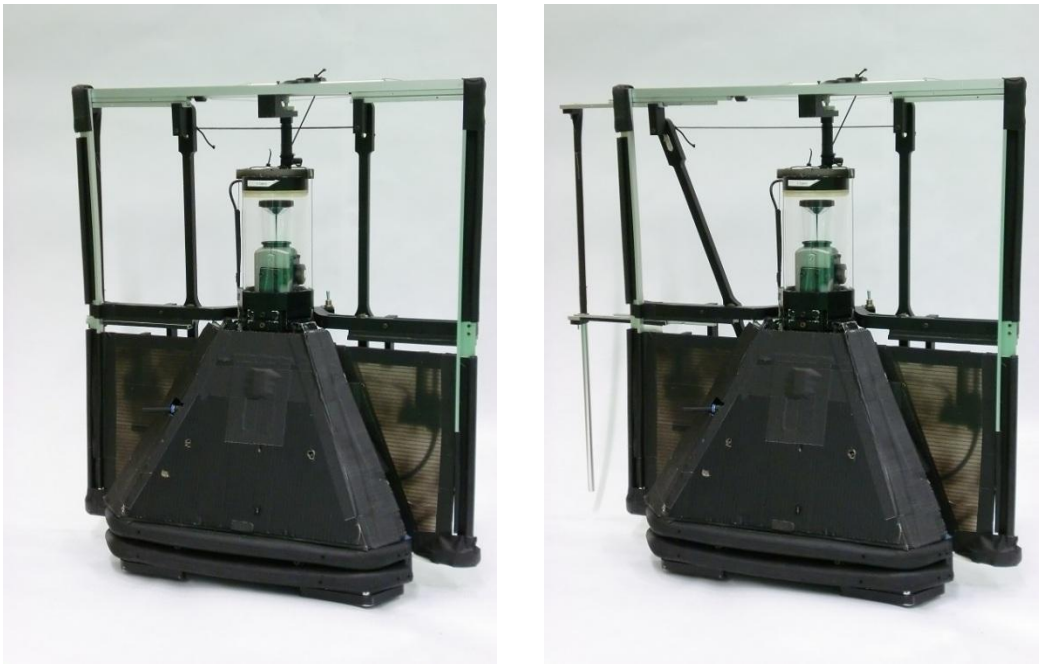


Fig. 9 Module structure of "Musashi".

各 Drive Module への速度制御は，ロボットに搭載された PC から RS232C によって送られる命令を基に，PID 制御により実行される．また，外界センサとして全方位カメラ，電子コンパスを搭載しており，IEEE1394 と USB を介してそれぞれ 30[fps]と 30[Hz]で PC へ情報を送信する．ロボット本体用バッテリには Ni-H バッテリ（25.9[V] / 2800[mAh]）を用い，キック装置用のバッテリとして Li-Ion バッテリ（14.4[V] / 8000[mAh]）を用いた．本バッテリによるロボットの稼働時間は約 30[min]である．ボールを蹴るための機構（以下，キック機構）にはソレノイド方式を用いており，ソレノイドの駆動には，ロボット本体に搭載された Li-Ion バッテリ 14.4[V]を 3 基の DCDC コンバータにより 90.0[V]へ昇圧し，コンデンサに充電することで出力を得ている．ソレノイド方式を用いることにより，最大 10[m]までの範囲において電流制御によるシュート飛距離の調整が可能となった[30][31].

< ゴールキーパロボット用 守備アーム >

Musashi のゴールキーパロボット[32]は，ゴールを守備する上で本体の面積を大きく取ることが効果的であることから，最大幅 74.0[cm]（50.0x50.0[cm]四方の対角線），最大高さ 80.0[cm]の守備フレームを搭載している．近年の RoboCup 中型リーグでは，5.0[m/s]以上の速度でシュートを放つロボットや，ループシュートを放つロボットなどが登場しており，Drive Module によるロボットの移動のみではゴールを守備出来ない場合が想定される．このような状況に伴い，RoboCup 中型リーグの公式ルールでは，ゴールキーパロボットに限り，守備行動のため 1.0[sec]間のみ本体寸法を 10.0[cm]拡張し，4.0[sec]の待機の後，再び 1.0[sec]間本体寸法拡張出来るルールが設けられている[45]．ルールに従い，Musashi は Fig. 10 のように展開式の守備アームを搭載している．Musashi における守備アームはプッシュ型のエアシリンダを用いて上方向と左右方向に，それぞれ上方向：10.0[cm] / 左右方向：14.0[cm]，本体を拡張するためのフレームを展開する．上方向への守備アーム展開に関しては，斜め情報前方に向けた超音波センサを設け，センサの取得情報により守備アームの展開を制御する．また，左右方向はロボットに搭載している PC と PIC により行う．



(a) Close　　(b)Right-Side Open

Fig. 10 The expansion transformation of the Goalkeeper defense arm

2.3.3　　ソフトウェア構造

Fig. 11 に“Musashi”のソフトウェア構造を示す．“Musashi”のソフトウェアでは，通信部，画像処理部，自己位置推定部，行動選択部等の各要素は，スレッド毎に実行される．まず通信部において，Kick off，Throw in といった試合進行の指令を受けとり，さらに全ロボット間で群内情報（Inner Group Information : IGI）を共有する[57]-[61]．次に画像処理部において，色抽出法に基づいてサッカーフィールド上の各対象物が検出される．ここではボール（黄色），障害物（黒色），そしてフィールド（緑色）と白線（白色）が認識され，それぞれの対象物との相対距離や角度も観測される[33][34]．自己位置推定部では，これらの検出された対象物情報に基づき，モンテカルロ自己位置同定法により，電子コンパスの情報と白線との相対距離・角度情報から自己位置が推定される．ここで，ボールや障害物の絶対位置は，ロボットの自己位置を基準にロボットと各対象物との相対距離から求められる．最後に，行動決定部において，各チームメイトの自己位置情報，ボールとの相対距離，相手チームロボットの位置等に依存して各ロボットの行動が決定される．以下に各スレッドの詳細を述べる．

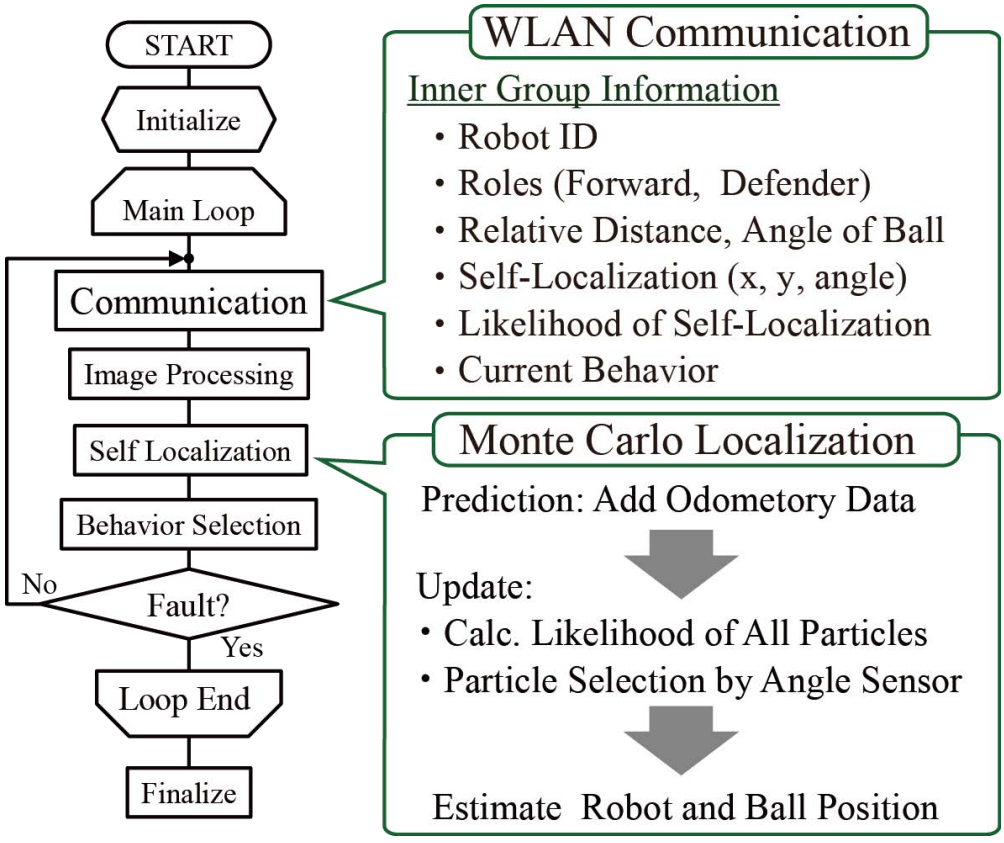


Fig. 11 Process flowchart of "Musashi".

I 通信部

"Musashi"は搭載した無線LANを介し，試合指示の受け取りや，チームメイトロボット間におけるIGI情報の相互通信を行う[57]-[61]．MSLでは人間の審判が下した判断をレフェリーボックスと呼ばれるPCによって各ロボットに搭載されたPCへ送られており，通信部ではレフェリーボックスより送られたKick Off，Throw In等の試合指示を受け取る．また，チームメイト間における相互通信では，Fig. 11に示すように，"情報のタグ"，"実行中の役割（FWやDFなど)"，"ボールとの相対距離と角度"，"自己位置情報"，"自己位置の尤度"，そして"実行中の行動（ボール取得動作，守備動作など)"が共有される．

II　　画像処理部

“Musashi”の“Camera Module”は全方位ミラーとIEEE1394デジタルカメラから構成される．カメラからの画像は30[fps]でYUV形式によりPCへ送信され，さらにHSV形式に変換された画像が生成される．各対象色（黄色，黒，緑，白）は，YUVとHSV形式の各画像に対し，対象色に合わせた上限と下限の閾値を設けることで抽出する[33]．例としてボールの色である黄色を抽出する場合を挙げると，ボールを構成する黄色のみが抽出されるよう，操作者が画像を目視しながら手動により閾値を設定する．YUV形式とHSV形式の画像は，異なる色空間により表現されるため，両画像に対して各閾値を設定した画像の論理積をとることで，画像に含まれるノイズが軽減される．

また，ボールや白線の誤認識を避けるため，白色や黄色の要素は，緑色の領域内に存在するもののみを対象とした．領域内外の判断は，まず白線とボールの領域を埋めるように緑色要素への膨張処理を施し，膨張処理後の緑色要素と，白色や黄色要素との論理積の出力によって判断した．各閾値の設定については，今後，サッカーフィールド上の照明環境に合わせて閾値を自動で調整するアルゴリズムの開発を行っている[34]．

検出された対象物に関する相対距離・角度の算出は，全方位カメラに関する相対距離・角度情報を参照することで求められる．これらの処理により，ロボットは，ボール，障害物，そして自己位置同定に用いられる白線に関する相対位置情報を認識する．

III　　自己位置同定部

ロボットによるサッカーでは, ボールを取得する動作からロボット同士によるパスに至るまで, 各ロボットがサッカーフィールド上における自己位置を認識する能力が重要となる．“Musashi”における自己位置推定は, 画像処理部において求めた白線との相対距離・角度に関する情報と方位センサの情報に基づき, モンテカルロ自己位置同定法(Monte Carlo Localization; 以下MCL)を用いて推定する[36]-[39]．ここでフィールドモデルには，フィールド中心を原点とし，x軸が自身のゴール方向を示す直交座標系を用いた．

ここで時刻tにおけるロボットの状態を，ロボットの位置$(r^t_x,\ r^t_y)$と方位${}_a\theta^t$ から$x_t = [r^t_x,\ r^t_y,\ {}^a\theta^t]^T$のように表す．

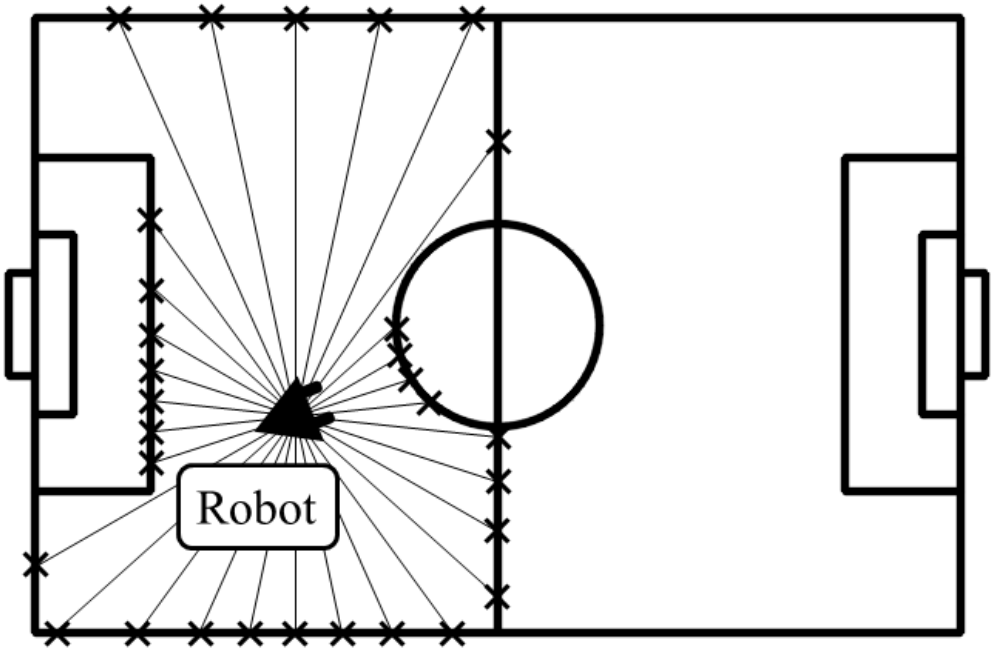


Fig. 12　Scanning lines image and result of self-localization.

事前確率密度$p(x_t|y_1 \ldots y_t)$は, 時刻tにおけるロボットの状態x_tと観測情報y_tの履歴から求められる．ここで時刻tにおける観測情報y_tはFig. 12に示されるような，ロボットを中心として6度毎に配置された60本の走査線によって得られた“最も近い白線までの距離情報”群を示している．

MCLにおける確率分布はN個のパーティクル（粒子）によって近似される．“Musashi”のシステムでは, 各パーティクルはロボットの状態と60個の白線との相対距離情報を含んでいる．ロボットの現在状態x_tは，事前状態x_{t-1}と制御入力u_{t-1}による条件付き確率密度$p(x_t|\ x_{t-1},\ u_{t-1})$によって表現される．ここで，Musashiでは制御入力u_{t-1}としてオドメトリ情報を用いている．MCLはマルコフ決定過程に従うことからロボットの現在状態に関する確率密度は式(2.19)から得られる．

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{u}_{t-1}) \cdot p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_{t-1} \tag{2.19}$$

次に式(2.20)に示すように事後確率密度$p(x_t|y_1 \ldots y_t)$は，白線との相対距離から求めた尤度$p(y_t|x_t)$を用いてベイズの定理から更新される．

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t)\cdot p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1})} \tag{2.20}$$

最終的なロボットの位置x^p_tはロボットの状態x_tと尤度に基づいた重み付き平均により求められる．“Musashi”の自己位置推定部では，上記の処理を繰り返すことにより実時間での自己位置推定を行う．繰り返し計算におけるパーティクルの再分布では，尤度に基づいて再分布の範囲を定め，自己位置x^p_tの周辺にパーティクルを分布する．

提案手法では，ロボットの方位${}_a\theta^t$はロボットの状態x_tから求められる．しかし，MSLのサッカーフィールドは直線と円弧の白線により構成され，左右対称の形状を持つ．従って観測データから推定したロボットの状態には方位が異なる二か所の位置が推定される可能性がある．一方で，ロボットの方向を特定するためのセンサとして地磁気を観測する方位センサが挙げられるが，磁気センサはサッカーフィールド周辺に配電管等が存在する場合，それらから発生する局所磁場の影響により誤差を含む可能性がある．“Musashi”における自己位置推定では，磁気センサが含む誤差を考慮し，方位センサが示す方位に対し±30度以上の方位差を示すパーティクルの尤度を0にすることで推定精度の向上を図っている．

IV　　行動判断部

“Musashi”の行動決定部では，IGIの一部であるボールとの相対距離に関する情報を比較することで自身の役割（FW，DF）を決定している．例として，あるロボットがチーム内で最もボールに近い場所に位置している場合，そのロボットはFWを担い，他のロボットは距離に応じてDFを担当するといった行動が挙げられる．実際の試合では，各役割は式(2.21)に基づいて動的に変化する[39][39].

$$f(d^i,\theta^i_g) = \{(k\cdot d^i\cdot\theta^i_g)/\pi\} + d^i \tag{2.21}$$

ここでk，d^i，θ^i_gは，それぞれ係数，ボールとの相対距離，ロボットから見た“ボールとゴールが成す角度”を示している．式(2.21)により求められた値が小さい場合は，そのロボットの状況がオフェンス動作に適していることを示している．実試合では，この式に基づく役割の設定は，相手チームの特徴等に合わせた操作者の主観により定めている．

オフェンス行動は主に次の行動から構成している．“Ball-Get(ボール取得のための動作)”，“Dribble(ボールをゴール方向へ運ぶ動作)”，“Avoidance(ロボットが障害物を回避する動作)”，“Shoot”である．ボール取得動作を円滑に行うため，Ball-Get動作ではロボットの角速度に対し，ロボットとボール間の相対角度に基づいたPD制御を用いた．さらにボールを取得する瞬間にボールを弾かないよう，ロボットの移動速度はボールとの相対距離に応じて減速する．

DribbleとAvoidance動作にはファジィポテンシャル法を用いた[41]. ファジィポテンシャル法では，様々な状況に考慮するためのルールを，ポテンシャルメンバーシップ関数（Potential Membership Function; 以下，PMF）によって表現する．ここでPMFの横軸はロボットの方位を表現しており，縦軸は各方位の優先度を表現するグレードを表している．これらのPMFを統合した結果，最も高いグレードを示した方位をロボットが進むべき方位とした．

PMFを定義するための主なルールを以下に示す．Musashiのシステムでは，ロボットの角速度と速度に対し，それぞれPMFを設計した．速度に関するPMFでは，ゴール方向に関するグレードを1とし，ロボットの周辺に障害物が存在する場合はその方向に関するグレードを下げるものとした．この際，グレード1はロボットが最高速度で走行することを意味している．角速度に関するPMFでは，相手ゴール方向のグレードを0とし，自身のゴール方向のグレードを1とした．つまり，角速度に関するPMFは相手ゴール方向を頂点とした三角形の形状を成している．速度に関するPMFと同様，グレード1はロボットが最大角速度で回転することを意味している．また，障害物の回避を考慮し，障害物が存在する方向のグレードを高くするものとした．これらのルールの設定により，ロボットは基本的には相手ゴールの方向へ最高速度で移動し，進行方向上に障害物が存在する場合には，障害物を避けつつゴールを目指すことが出来る．また，これらのルールに従ってボール取得動作を行った場合，ロボットはゴール方向を向いた姿勢でボールへアプローチする．このアプローチ動作により，ボールを取得した後にロボットが円滑にシュート行動に移ることが出来る．

第三章

強化学習による
ゴールキーパロボットの行動獲得

第三章　強化学習によるゴールキーパロボットの行動獲得

3.1　はじめに

本章では強化学習の一つである Q 学習を用いて試行錯誤的に行動を獲得させる手法について提案し，シミュレーションと実機を用いて評価する． RoboCup サッカー中型リーグに用いられる自律型移動ロボットに必要な研究課題としては，RoboCup プロジェクト提唱時に，提唱者である浅田稔，北野宏明らによって，実時間画像処理，行動学習，協調行動の獲得等が提示されている[46]．また，2000 年には Sergio Monteiro らによって RoboCup のようなマルチエージェントシステムを発展させていく上では，ボールの動きや相手チームの動作に依存する環境に対し，反応的・熟考的かつ予測的な制御が重要になると述べている[16]．特にロボット単体において最も基本的かつ重要な要素としてロボットの行動決定があげられ，各ロボットの行動決定が発達するにつれて，他のチームメイトロボットの動向や相手チームのロボットの行動等を考慮した協調行動へと発展すると考えられる．自律型ロボットにおいて行動を自律的に決定するためには，環境の観測，観測情報の解析，行動計画の立案，行動の実行といったいくつかの段階を踏むことが求められる[47]．これらの行動決定における各段階において，設計者が事前にロボットの全行動を設計する場合，どのような情報を対象として観測するのか，ロボットの動作環境においてロボットが遭遇する状況にはどんなものが想定されるか，その状況において最適な行動は何かなどを設計者が可能な限り想定する必要がある．ロボットの動作環境が既知から未知へ，静的環境から動的環境へと複雑になるにつれて，設計者による想定と準備は困難になる．RoboCup サッカー中型リーグのように，複数のロボットによるセットプレイなど，複数のタスクが存在する環境では，環境やタスクが複雑であり，それらすべてに対処するプログラムを設計者が全て想定することは困難である．人間や他の動物が行動を決定する際には，「パブロフの犬」に示されるように，過去の経験や，周囲の状況，他者との関係などを情報として適切な行動を決定すると考えられる．そこで，本章では強化学習を用いて，過去の経験や周辺情報を基に，自律型移動ロボットの自律的な行動学習について述べる．

このような行動学習に関しては，浅田らが行ったように強化学習によって RoboCup におけるシュート行動を学習した研究がある[48][49]．また浅田らは，実ロボットに強化学習を適用する場合に，状態空間の構成と複雑なタスクに対する学習について課題を定義している[51]．加藤らは PWS（Power Wheeled Steering）方式の移動機構を用いたロボットによって，ゴールキーパの守備行動を学習している[52][53].

3.2 強化学習の基礎

強化学習では、エージェントと呼ばれる学習器が外部環境との相互作用型学習によって自身の行動評価と、次の行動選択を自律的に行うことが可能である。また、外部環境との相互作用型学習であることから外乱に対するロバスト性が高く、自律型ロボットの行動決定手法として適した機械学習法だと言える。車の運転や道具の用い方などに関する人間の学習は、自らの能動的な動作と動作の結果得られる外部環境の変化との相互作用によって促進されると考えることが出来る。このような学習構造を数理モデル化し、エージェントと呼ばれる学習器に自律動作を学習させる機械学習法を強化学習と呼ぶ。強化学習は、エージェントと外部環境との相互作用によってエージェントがどのように行動すべきかを学習する学習構造を持つ。

具体的には、数値化された状態(State)によって学習器が選択すべき行動(Action)を決定し、行動の結果得られた新たな環境情報に対してなんらかの報酬(Reward)を与えることで自らの選択した行動を評価するという一連のプロセスを繰り返すことで学習する。強化学習の学習構造を Fig. 13 に示す。

Fig. 13 Structure of Reinforcement Learning

Fig. 13 から強化学習における学習プロセスには方策、報酬関数、価値関数と呼ばれる 3 つの構成要素が存在することがわかる。

方策は、ある状態においてエージェントが次にどのような行動を選択すべきかを示す。方策を決定する代表的な手法としてグリーディ手法、ε グリーディ手法、ソフトマックス行動選択などが挙げられる。

報酬関数は、エージェントが遭遇したある状態を数値化した報酬として表し、ある時点におけるエージェントの状態がどの程度望ましいものであるかを与える。強化学習においてエージェントは最終的に受け取る総報酬を最大化するよう方策を立てるため、報酬をどのように与えるかによってエージェントの学習を意図する方向へ仕向けることが出来る。

価値関数は、ある状態における行動が、将来的にどのような総報酬をもたらすかを示している。価値関数は報酬によって更新され、ある状態における行動が最終的に低い総報酬しか得られなかった場合には、選択した行動が低い価値しか持たなかったものとして更新される。価値関数は強化学習における行動を決定するため、強化学習の中核をなす要素と言える。強化学習は、方策、報酬関数、価値観数の三つの要素によってエージェントと環境間における学習を行っている。強化学習を用いて実問題を解くための基本的な解法として、動的計画法、モンテカルロ法、TD(Temporal Difference)学習法の三つの解法がある[2]。

I. 知識利用と探索動作

強化学習においてエージェントは、価値に関する自らの知識を利用して行動する性質(知識利用)と、価値とは無関係により高い総報酬を目指して低い価値の行動についても探索する性質(探索動作)が求められる。価値に関する自らの知識を利用して行動する手法を、エージェントが価値に対して貪欲であるという見方からグリーディ手法と呼ぶ。しかしグリーディ手法では、これまでの探索範囲の中から高い価値を持つ行動を選択するため、未探索領域に高い価値がある状態があってもこれを探索しないので、一般にグローバルな最適解には到達できない。

知識利用と探索動作という二つの相反する必要性を満たすために、グリーディ手法に確率 ε で価値とは無関係にランダムな行動を選択させ探査を行わせる手法を ε グリーディ手法と呼ぶ。

ε グリーディ手法における探索動作は価値に無関係に確率 ε で行われるため、将来的にも高い総報酬が得られないことが明白である行動も同じ確率で選択してしまうという欠点が上げられる。式(3.1)に示すギブス関数(ボルツマン分布)を用いて各行動の価値によって行動選択確率を変化させることで、知識利用と探索動作を効率よく実現させることが出来る。この手法を Softmax 行動選択と呼ぶ。

$$P_i = \frac{e^{Q_t(a_i)/\tau}}{\sum_{b=1}^{n} e^{Q_t(a_b)/\tau}} \qquad i = 1,2,3\cdots \tag{3.1}$$

P_i：行動a_iに関する行動選択確率

$Q_i(a_i)$：行動a_iの行動価値関数

τ：温度

温度τは選択確率を変化させる正定数のパラメータであり、温度τが高い場合には全ての行動がほぼ同程度に起こるが、温度τが低くなるにつれて行動選択の差が大きくなる。Fig. 14 に行動 a1、a2、a3 の行動価値関数値を固定し、Episode 数が増加するにしたがって温度 を減少させた場合の行動選択確率の遷移を示す。Fig. 14 より、学習初期には低い価値を持つ行動 (a3) に関しても行動が選択される可能性をもつが、Episode 数が増加するにしたがって最も高い価値を持つ行動が 100%の確率で選択されるグリーディな行動へ遷移してゆく様子がわかる。

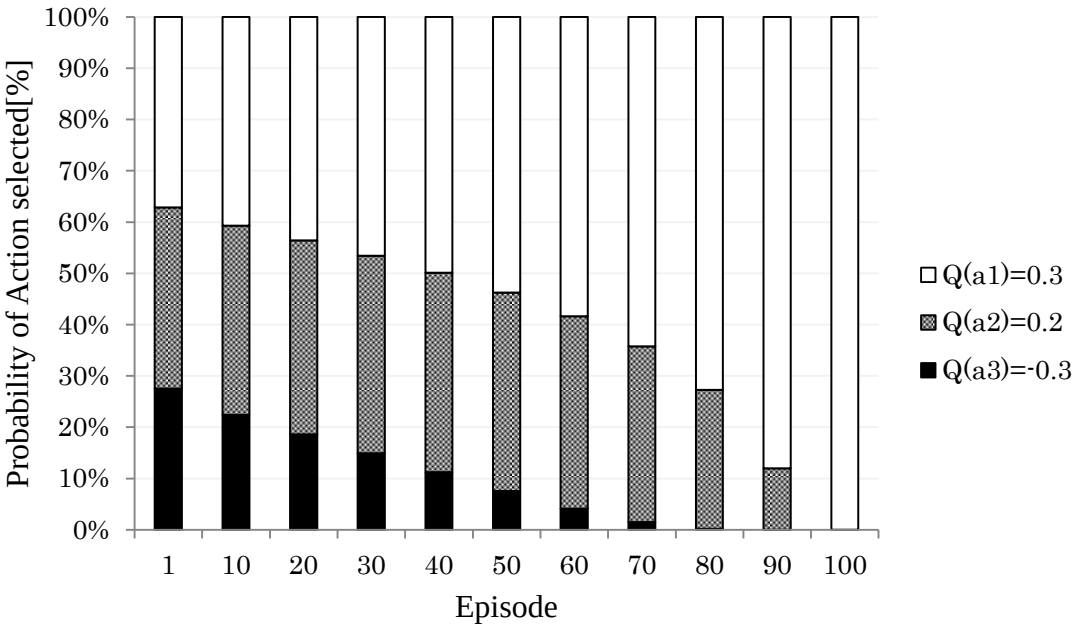


Fig. 14　Softmax action selection rule

II. 動的計画法 (Dynamic Programming；DP 法)

動的計画法(DP 法)は、エージェントが活動する範囲を有限のものとし、活動環境の完全なモデルが、マルコフ決定過程(MDP)として与えられている場合のみ、最適方策を計算できるアルゴリズムである。実問題においては、環境の完全なモデルを得ることは非現実的であり、DP 法の中で行われる方策反復や価値反復といった手法それぞれが膨大なメモリと計算量を必要とすることから DP 法は実用的ではないと言える。

しかしながら DP 法は強化学習において重要な理論を持つ手法であり、後に続くモンテカルロ法、TD 学習法なども異なる手法によって DP 法とほぼ同様の目的を達成する手法となっている。また DP 法は、現在の状態価値を後に得られる状態価値の推定量から更新する手法(ブートストラップ)を用いており、他の強化学習手法においてもブートストラップは頻繁に用いられる概念となっている。

III. モンテカルロ法

DP 法が予め与えられた完全な環境モデルを基に行動を計画するのに対し、モンテカルロ法ではエージェントの経験に基づいて学習を進める。この経験とは 1 Episode 終了時に得られた総報酬の結果を意味し、経験を基に状態価値と行動価値を推定する。

モンテカルロ法では、1Episode あたりの報酬結果から価値を推定するため“どのような行動を実行したとしても最終的には目的は達成される”という前提が必要となる。また価値推定のため

に 1Episode の全情報を記録する膨大なメモリ領域が必要であることや学習の収束までに膨大な学習時間がかかること、1Episode レベルで学習するため 1step レベルでの学習に適さないなどの問題から DP 法同様に実用的ではないと言える。

IV. 時間的差分学習法 (Temporal Difference Learning；TD 学習法)

TD 学習法は、DP 法とモンテカルロ法の利点をそれぞれ取り入れた学習手法である。具体的には、DP 法より“他の推定量を用いて現在の推定量を更新するブートストラップの概念”を取り入れ、モンテカルロ法より“完全な環境のモデルを必要としない、エージェントによる直接学習の概念”が取り入れられている。このことから TD 学習法は、環境の完全なモデルを必要としない、1[Step]レベルでの学習が可能な強化学習法であり、自律型ロボットの制御に最も適した機械学習法であると言える。一般的には TD 学習は、アルゴリズムが複雑で分析が困難であるというデメリットが知られているが、このデメリットを克服する手法として Q 学習が提案されている。本研究では、自律型ロボットへの適用を想定したシミュレーションを行うという目的から強化学習のシミュレーションに TD 学習アルゴリズムを用いた。TD 学習の最も基本的なアルゴリズムを式(3.2)に示す。

$$V(s_t) \leftarrow V(S_t) + \alpha[r_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{3.2}$$

$V(S_i)$：時刻tにおける価値関数

r_{t+1}：時刻tの次状態における報酬

α：ステップサイズパラメータ

αは1回の更新における学習の更新率を示す。式(3.2)のr_{t+1}からモンテカルロ法と同様、報酬を基に直接学習を行っていることがわかる。また、$V(S_{t+1})$からDP法と同様にブートストラップが用いられていることがわかる。式(3.2)によって示したTD学習アルゴリズムを制御問題に当てはめた手法として方策オン型TD制御と方策オフ型TD制御があげられる。

V. 方策オン型 TD 制御法 (State-Action-Reward-State-Action；Sarsa 法)

TD 学習法を制御問題に当てはめた場合は、エージェントがある方策のもとで状態 におけるすべての行動 に対して評価する必要がある。したがって Sarsa 法は行動価値関数の更新によって学習する。式(3.3)に Sarsa 法の行動価値関数更新式を示す。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{3.3}$$

Sarsa 法では状態 から次の状態に移る度に行動価値関数の更新が行われる。行動価値関数の更新は次状態 と行動 によって行われる。行動 は方策に基づいて決定される行動なので、Sarsa 法における行動価値関数の更新は方策に基づいた行動 によって行われることから方策オン型 TD 制御と呼ばれる。

VI. 方策オフ型 TD 制御法 (Q 学習法)

Q 学習も Sarsa 法と同様に制御問題を扱うため行動価値関数 Q(s_t, a_t)を用いる。式(3.4)に Q 学習における行動価値関数の更新式を示す。

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \tag{3.4}$$

Q 学習では行動価値関数の更新が、次状態 s_{t+1} と次状態 s_{t+1} における最大の価値をもつ行動α によって行われる。したがって方策に基づいた行動 a_{t+1} とは無関係に行動価値関数の更新が行われることから、方策オフ型 TD 制御と呼ばれる。Q 学習は max $Q(s_{t+1}, a)$ を用いたことでアルゴリズムの解析が容易になり、TD 学習のデメリットを克服した[56]。本研究では、解析が容易であるという理由から行動価値関数の更新アルゴリズムとして Q 学習を用いた。

3.3 ゴールキーパロボットへのQ学習の適用

RoboCup サッカー中型リーグにおけるゴールキーパロボットのタスクとして，自身のゴール方向に向かってくるボールに対して, ボールがゴールラインを超えないようボールとゴールの間に入って行動する動作があげられる. 学習環境としては，サッカーフィールド上に守備対象であるゴールとボールが一つずつ存在する環境を想定する. ボールは直線的な軌道でゴール内へ向かうものとし，ボールがキーパの正面に接触する状態，もしくはゴール外にボールが出る状態までを 1[episode]として守備行動を学習する．ゴールキーパロボットは，サッカーフィールドの絶対座標系におけるロボットの位置とボールの位置，ボールの速度，守備アームの状態を観測するものとする．本章において行う Q 学習では，これらの状態を状態変数とし，離散空間内で行う．

また, 連続空間を離散空間に置き換えて強化学習を行う場合, Bellman が“次元の呪い（the curse of dimensionality）”と呼ぶ問題を考慮しなければならない．“次元の呪い”とは，多くの状態変数を用いて離散空間を連続空間へ近似することで, 必要な計算量が指数関数的に増加する問題である．ゴールキーパロボットに求められるタスクでは，ロボットの位置（x,y,q），ボールの位置（x,y），ボールの速度（vx，vy），守備アームの状態（Open / Close）の 8 次元の状態変数を定義する必要がある. 本章では次元の呪いへの対処として, ゴールキーパロボットの方位を固定都市, 守備行動を左右方向のみと限定することで状態変数の増大を避けるものとした.

なお，次元の呪いは，学習処理を実行するハードウェアの能力に依存するものと考えられ，学習に用いるコンピュータの性能が向上するにつれて, より多くの状態変数を定義することが出来る．本章において示すゴールキーパの守備行動学習の目的は，状態変数や報酬関数を適切に定義することで設計者の意図する行動が強化学習によって獲得できることを示すことにある.

3.3.1 状態変数の定義

状態変数として，ロボットの位置[x_r][y_r]，ボールの位置[x_b][y_b]，ボールの速度[v_x][v_y]，守備アームの状態[s_a]を定義する．ここで，x_r はロボットの位置の x 座標，y_r はロボットの位置の y 座標，x_b はボールの位置の x 座標，y_b はボールの位置の y 座標とする．また，v_x はボールの移動速度の x 方向成分を表し，v_y はボールの移動速度の y 方向成分を表す．s_a は守備アームの状態を表している．ゴールキーパロボットの行動は横移動のみに限定するため，ゴールキーパロボットの行動範囲はゴールの正面 0.1[m]x2.0[m]の範囲を想定し，0.1[m]単位で離散化する．また，ゴールキーパロボットの移動速度は 1.0[m/s]とした．ボールの移動範囲はゴールの正面 6.0[m]x3.0[m]を想定し，0.1[m]単位で離散化するものとした．ボールの移動速度としては，2011 年現在の Hibikino-Musashi におけるゴールキーパロボットでは対処が困難である 8.0[m/s]以上でゴールへ向かってくるボールを想定し，ボールの移動速度の x 方向成分を{0.0[m/s] < v_x ≦10.0[m/s]}，y 方向成分を{-5.0[m/s]≦ v_y ≦5.0[m/s]}の範囲とし，それぞれ 1.0[m/s]で離散化した．守備アームの状態については，s_a=0 を守備アームの収納状態，s_a =1 を左方向へ守備アームを展開した状態，s_a =2 を右方向へ守備アームを展開した状態とした．守備アームの状態変数は，行動に依存して収納状態から右方向への展開，もしくは左方向への展開へ状態が遷移するものとする．以上の状態変数の定義により各状態変数は，ロボットの位置[x_r][y_r]=[1][20]，ボールの位置[x_b][y_b]=[60][30]，ボールの速度[v_x][v_y]=[10][10]，守備アームの状態[s_a]=[3]となり，7 次元（10,800,000）によって表現する．

3.3.2 行動の定義

強化学習においてエージェントは環境から取得した情報によって，自身の行動 a を実行する．本章において対象としたゴールキーパロボットは，全方位移動機構を持つため，前後左右への並進移動に加え，回転行動や並進と回転を同時に行う移動が可能であるが，本章において扱う強化学習では次元の呪いを考慮し，ゴールキーパロボットの行動を a = 0（停止），a = 1（左方向並進移動），a = 2（右方向並進移動）の 3 種類とした．また守備アームの展開は「ロボット本体の移動のみではボールを守備出来ない場合に展開」するものとし，上記した 3 種類の行動に加えて a = 3（左方向並進移動＋左方向への守備アーム展開），a = 4（右方向並進移動＋右方向への守備アーム展開）という 2 種類の行動を定義した．ここで本章において行う強化学習は，マルコフ決定過程(Markov Decision Process：MDP)に従うものとし，移動におけるロボットの加速度等は考慮しないものとした．

3.3.3 方策の定義

本章において扱うゴールキーパロボットの Q 学習では，より最適な行動の探索と方策による行動獲得結果の比較を目的とし，ε-greedy 手法と softmax 手法の 2 つの方策による行動学習結果について検証する．ここで，ε-greedy 手法では，常に一定の確率で探索行動を行うため学習の収束に膨大な時間を要することから，式(3.5)に示すように学習回数に応じて定数εを減衰させることで，学習初期には探索行動を伴う行動を選択し，学習後期にはグリーディに行動を選択するものとした．式(3.5)においてεはランダム行動を行う閾値となる定数を表し，$\varepsilon_{\max}$は定数εの最大値を示す．またξは減衰係数を表しており，kはエピソード数，$k_{\max}$は最大学習回数を示す．本章において述べる学習では，$\varepsilon = 0.4$，ξ=0.5と設定して行った．

ε-greedy 手法において定数εを減衰させることは softmax 手法と同様の効果をもたらすことになるが，両者の違いは定数εの減衰が行動価値関数 $Q(s, a)$に基づかない点と学習回数に基づいた減衰が線形に減衰するか，非線形に減衰するかという点にある

$$\varepsilon = \varepsilon_{\max} - \xi \frac{\varepsilon_{\max} \cdot k}{k_{\max}} \tag{3.5}$$

3.3.4 報酬の定義

ゴールキーパロボットにおける基本的な守備行動として, ゴール方向へ向かってくるボールとゴールの間にロボットが移動することで, ボールの進行を阻害しゴールを守備する行動があげられる．この守備行動は，ロボットが守備範囲内においてボールとの相対距離を小さくするように移動すると言える．従って，学習の際はボールが x=0.0 の位置（以降，エンドラインと記す）を超えた際にエピソードを終了するものとし，その時のボール位置 y_b とロボット位置 y_r の相対距離Δy に応じて，報酬を与える．Δy = 0.0[m]の時，ロボットはボールを正面で止めたこととなり，最大の報酬 $r^{\max}$ = 1.0 を与えるものとする．報酬はΔy の一次関数とし，$|\Delta y|$が増加する程報酬が減少する. 報酬が与えられる範囲として, 本章では Fig. 15 に示すように{-1.0[m] ≦ Δy ≦1.0[m]}と設定した.

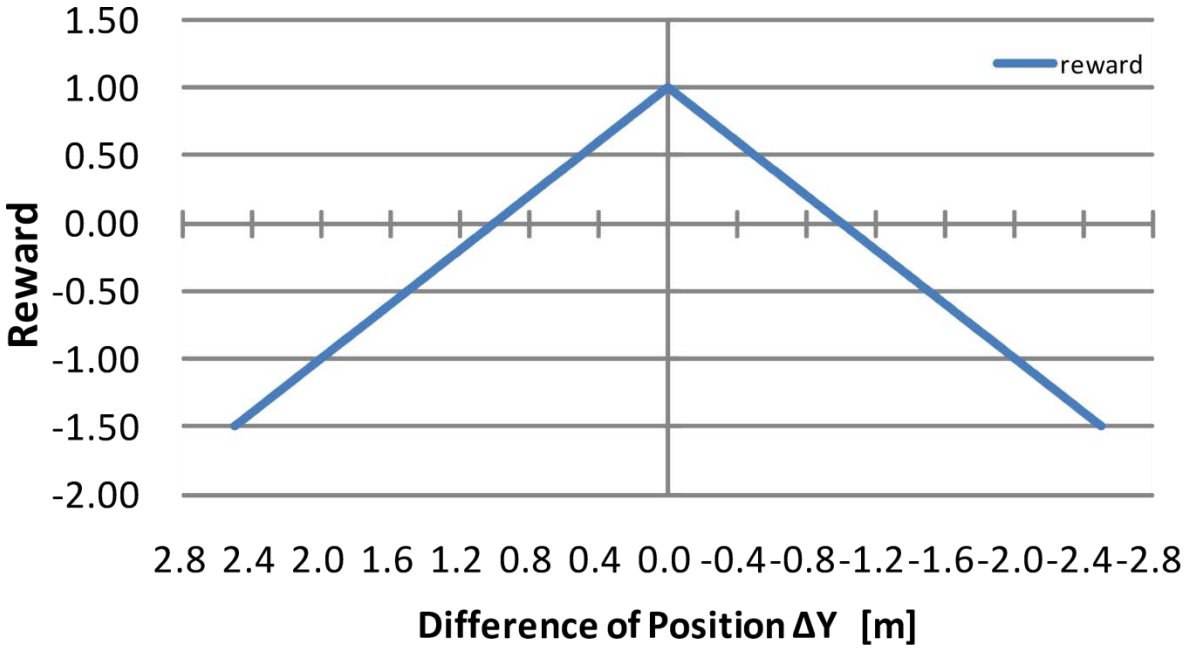


Fig. 15 Liner Reward Function

設計者が意図するゴールキーパロボットの守備行動として，「ゴールから遠く離れるような守備行動は避ける」という条件を付加した．これは，守備範囲をゴール範囲(幅：2.0[m])内のみと限定することで，最小限の動きによってボールの進行を防ぐことが出来るという考えに基づいて設定した．従って，ゴールキーパロボットの位置 y_b が，ゴール範囲を越えて遷移した際には，r_{t+1}= -1.0 が与えられるものとした．このような負の報酬を，本章では“罰”と呼ぶ．

実際のゴールキーパロボットは本体に搭載されたバッテリにより行動するが，行動を最低限に抑えることでバッテリの消費を抑えることが出来る．そこで，「バッテリの消費を抑えるため，最短経路によってゴールを守備する」という条件を加え，この条件を実現するために，右方向並進移動，左方向並進移動を行う毎に r_{t+1} = -1.0*e*-6 の罰を与えた．本章では行動毎に与えられる罰を“行動罰”と呼ぶ．また，実際のゴールキーパロボットにおいて，守備アームの展開は空気圧シリンダによって行われるため，空気圧の消費を考慮し左方向並進移動＋左方向守備アーム展開，右方向並進運動＋右方向守備アーム展開の行動には，r_{t+1} = -1.5*e*-6 の行動罰を与えた．

強化学習においてエージェントは, 基本的な行動選択として高い価値を示す行動を選択するため，上記の条件に基づいた報酬の付加により，「定められた範囲内において，最小限の行動によりゴール守備動作を行い，守備アームの展開は必要最低限に行う」という行動の獲得が期待できる．報酬関数を式(3.6)～(3.9)に示す.

$$r_{t+1} = r^{\max} - \Delta y \quad (x = 0.0) \tag{3.6}$$

$$r_{t+1} = -1.0 \quad (y_r < -1.0 \,\|\, y_r > 1.0) \tag{3.7}$$

$$r_{t+1} = -1.0e-6 \quad (a = 1 \,\|\, a = 2) \tag{3.8}$$

$$r_{t+1} = -1.0e-6 \quad (a = 3 \,\|\, a = 4) \tag{3.9}$$

3.4　Q学習による行動獲得評価実験

3.4.1　シミュレーション実験の条件

実環境を離散化したシミュレーション環境において, ゴールキーパロボットの行動獲得について評価する．ゴールキーパの初期位置，ボールの初期位置，ボールの移動速度はエピソード毎にランダムに与えるものとし，以下の条件によってエピソードを終了するものとした.

・ロボットがボールと接触する

・ボールがエンドラインを越える

・ロボットが守備範囲を越える

各エピソード開始時のゴールキーパロボットの初期位置は，y_rの状態数である 20 通りからランダムに決定する．またボールの初期位置の x 座標は，ゴールの正面から 6.0[m]前方の地点(x = 6.0)に固定とし，y 座標は y_b の状態数である 30 通りからランダムに決定する．またボールが進む方向に関しても y_b の状態数である 30 通りからランダムに決定するものとし，ボールの軌道は計 900 通りから定まる(Fig. 17)．ボールの速度は，1.0[m/s]~10.0[m/s]まで 1.0[m/s]毎に 10 通りを想定し，各エピソードにおいてランダムに決定されるものとした．決定されたボールの速度は，x 方向成分と y 方向成分にそれぞれ分けられ，状態変数 v_x，v_y とする．従ってボールの移動パターンは計 9000 通りとした．本章ではこれらのパターンを“シュートパターン”と呼ぶ.

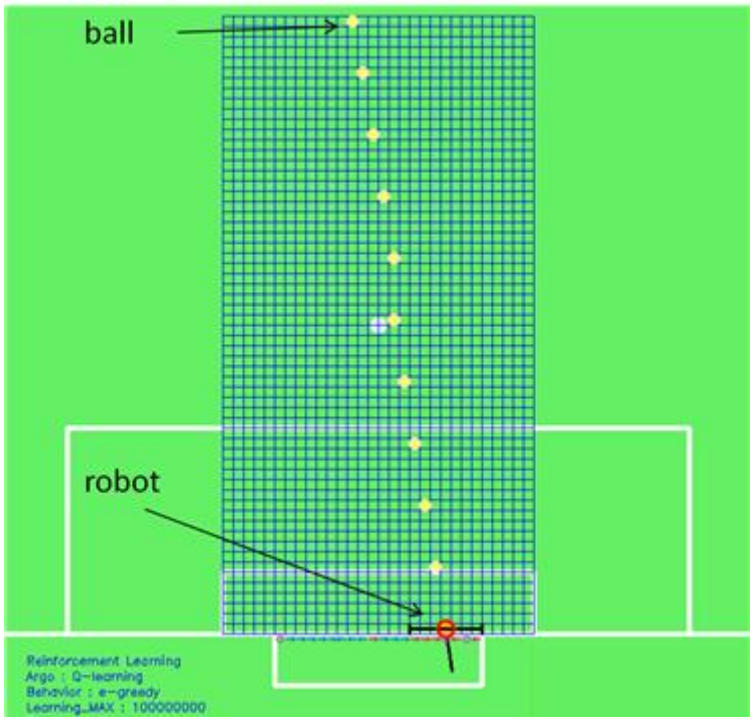
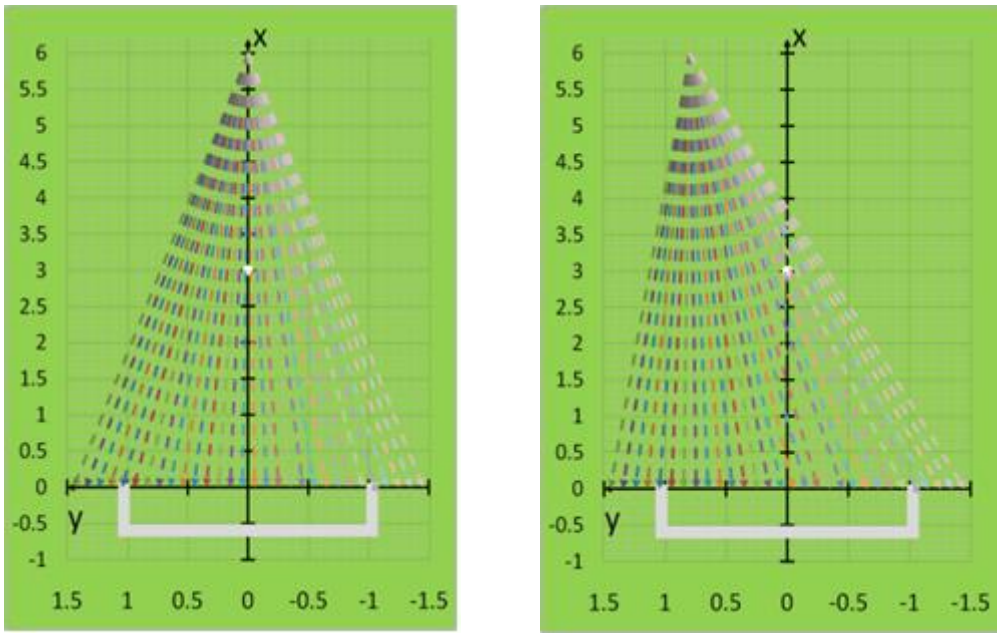


Fig. 16 Developed Reinforcement Learning Simulator



(a) : Initial point $(x_b, y_b) = (6.0, 0.0)$ (b) : Initial point $(x_b, y_b) = (6.0, 0.7)$

Fig. 17 Examples of Ball Trajectories

3.4.2　守備アーム展開を伴わない守備行動の獲得実験

本節では, 守備アームを用いない場合のゴールキーパロボットの守備行動獲得実験について述べる．報酬は式(3.6)~(3.9)に基づいて与えられるものとした．また，行動罰の有無に対する各エピソードの総走行距離から, 守備行動の最適化について行動罰が与える効果について評価を行った. 本実験おける最大エピソード数は 50,000,000 回(5 千万回)とし, エピソード 10 万回毎に, 1000 通りのランダムなシュートパターンに対する守備回数を評価値とした. 守備行動の学習結果については，1000 通りのシュートパターンに対する守備率(Success Rate of Saving)，行動の最適化については平均走行距離(Odometory)，学習の収束については各状態における行動価値 Q の最大値の総和(Sum of Q-max)によって評価を行った．実験条件を Table 2，実験結果を Fig. 18~Fig. 20 に示す.

Table 2　Condition of Experiment

Item	*Details*
Robot Position	$[x_r][y_r]$ = [1][20]
Ball Position	$[x_b][y_b]$ = [60][30]
Ball Speed	$[v_x][v_y]$ = [10][10]
Arm Condition	$[s_a]$ = [3]
Action	$[a]$ = [5]: Stop / Right-Move / Left-Move / Right-Move with Arm / Left-Move with Arm
Learning rate	α = 0.1
Discount Rate	γ = 0.5
Policy	ε-greedy
ε value	$\varepsilon = 0.4$
Damping coefficient	$\xi = 0.8$
Reward	$r_{t+1} = r^{\max} - \Delta y \quad (x = 0.0)$
Penalty	$r_{t+1} = -1.0 \quad (y_r < -1.0 \,\|\| \quad y_r > 1.0)$
Moving Penalty	$r_{t+1} = -1.0e-6 \quad (a = 1 \,\|\| a = 2)$ $r_{t+1} = -1.0e-6 \quad (a = 3 \,\|\| a = 4)$
Episode	50,000,000

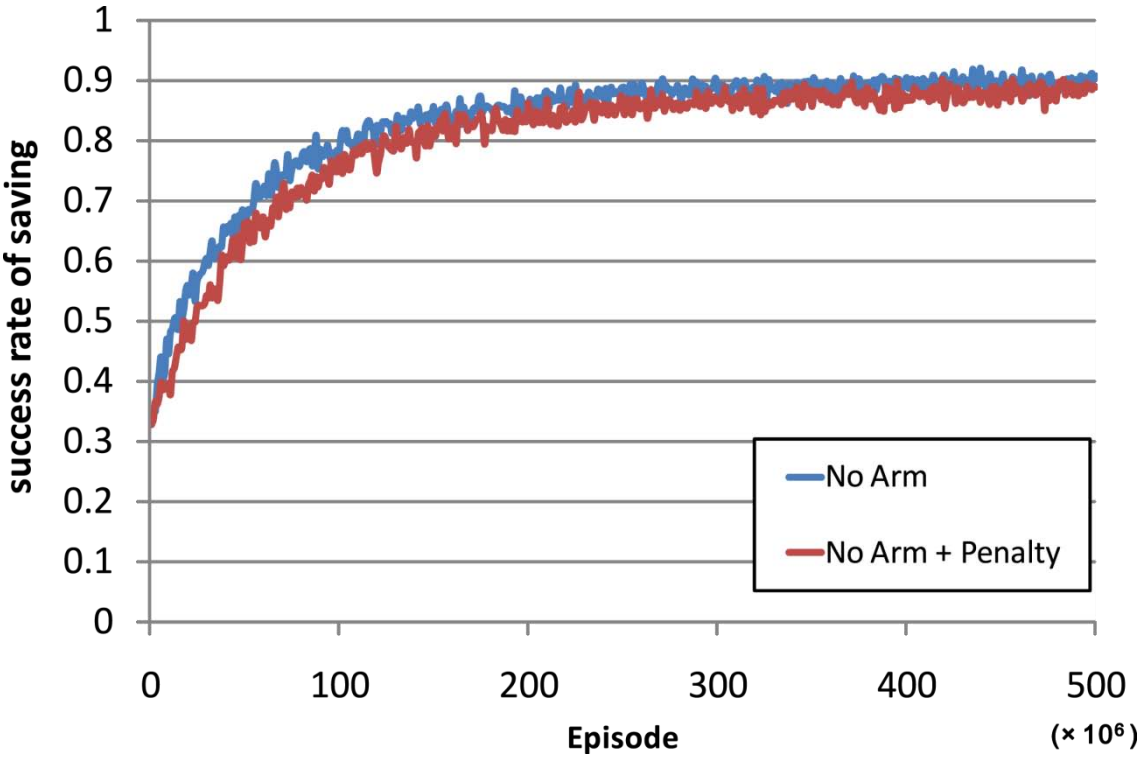


Fig. 18 Result of Saving Success Rate at No-Arm Condition

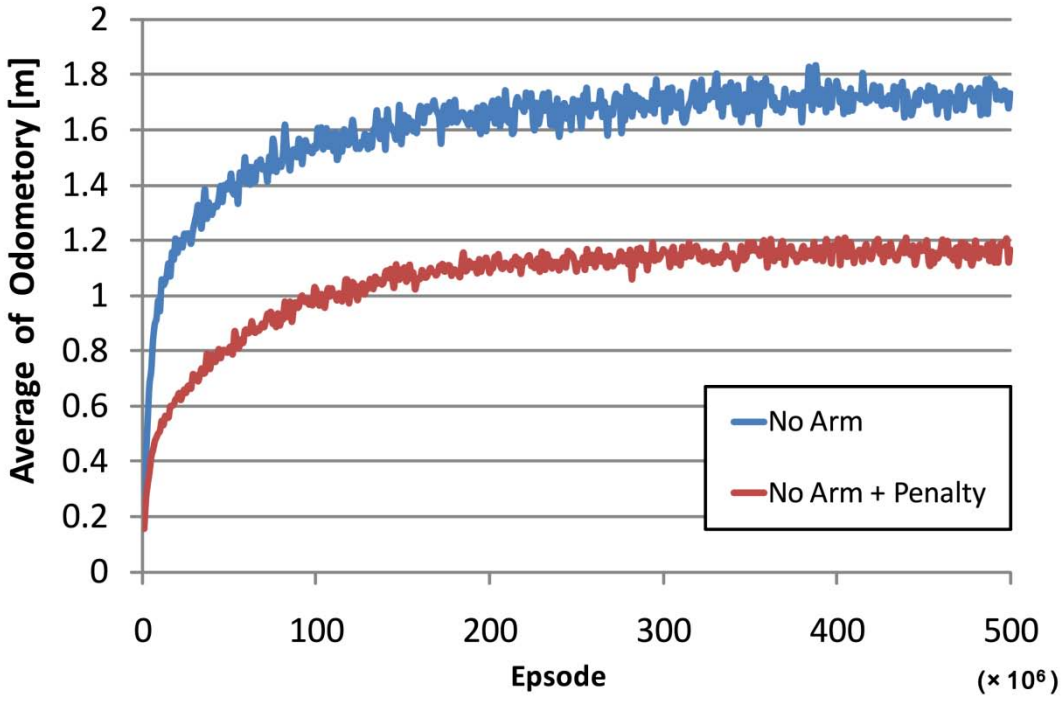


Fig. 19 Result of Odometory-Average at No-Arm Condition

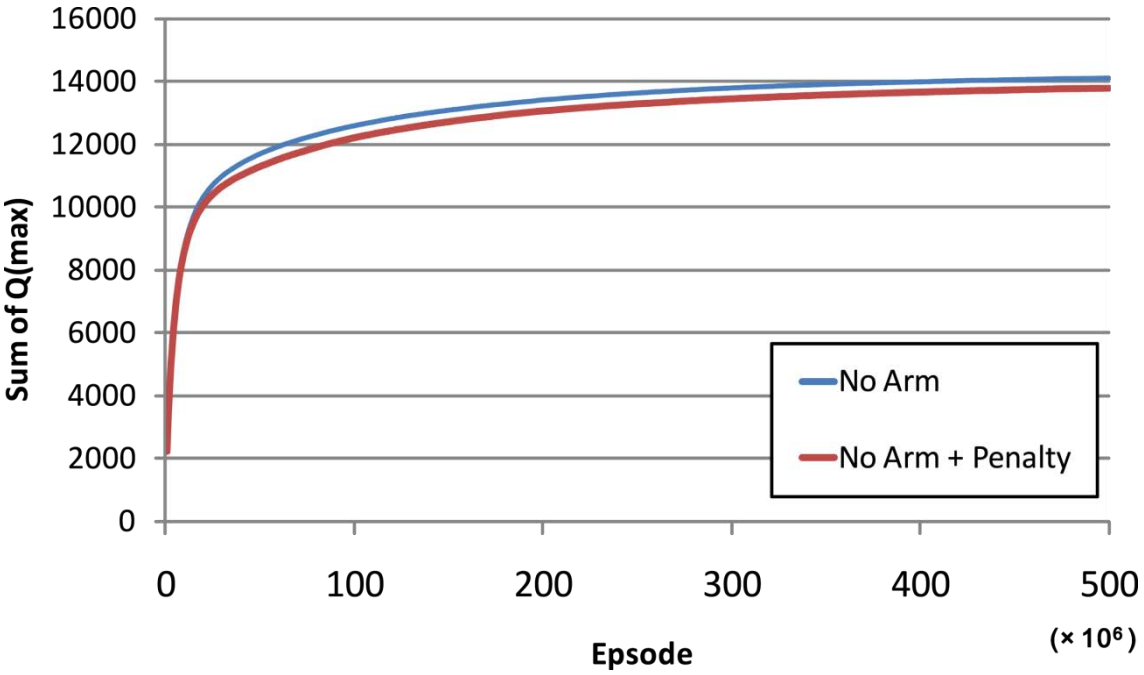


Fig. 20　Result of Sum of Q-max at No-Arm Condition

Fig. 18 に示す結果から，守備アームを用いない場合の行動獲得実験では，約 30,000,000 回（3 千万回）の学習後，行動罰の有無に関わらず約 90%の守備率を示した．Fig. 18 において，行動罰を与えた場合は，0~2 千万回までの学習において守備率の向上に遅れが見られる．これは行動罰を与えた結果, 学習初期における探索行動において各状態における負の価値が広く分布されたことに起因すると考えられる．約 3 千万回の学習後，同程度の結果を示していることから行動罰の付加によって守備率が低下しないことを示した．Fig. 19 に示す結果では，行動罰を与えない場合，守備行動に対して平均 1.7 [m] の走行距離を示したが，行動罰の付加により平均走行距離が 1.2 [m] まで減少した．Fig. 18 より守備率に大きな変化が見られないことから，ボールの守備に至るまでの行動が最短経路によって行われていることが分かる．また，Fig. 20 より，行動罰の有無に関わらず約 3 千万回の学習後，行動価値 Q の増加が定常的になったことがわかる．Fig. 18～Fig. 20 に示す結果より，守備アームを用いない状態における学習には，約 3 千万回の学習が必要であり，行動罰の導入によって行動の最適化が得られることを示した．

3.4.3　守備アーム展開を伴う守備行動の獲得実験

本節では，ゴールキーパロボットの行動に守備アームの展開行動（a = 3：左方向並進移動＋左方向への守備アーム展開，a = 4：右方向並進移動＋右方向への守備アーム展開）を加えた際の行動獲得実験について述べる．報酬は，3.4.1 節において示した実験時と同様の報酬を与えるものとした．また，守備アーム展開に関する行動数の増加を考慮し，最大エピソード数を 100,000,000 回（1 億回）とした．評価は 3.4.1 節において示した実験と同様に，学習 10 万回に対して 1000 通りのシュートパターンによるテストを行った際の守備回数から評価した．評価項目に関しても 1000 通りのシュートパターンに対する守備率と平均走行距離，行動価値 Q の最大値の総和に基づいて評価する．実験結果を Fig. 21~Fig. 23 に示す．

Fig. 21 の結果から，守備アーム展開を伴う行動獲得では，約 100,000,000 回（1 億回）の学習後，最終的な守備率は約 85[%]を示した．守備アームを用いない場合の守備率と比較すると約 5[%]の低下が見られた．また，Fig. 22 に示す総走行距離平均の結果では，平均 1.1[m]を示しており，守備アームを用いない場合と比較すると 0.1[m]の減少が見られた．Fig. 23 に示す結果から，守備アームを用いない場合に比べ，行動価値 Q の総和が 2.3 倍に増加し，1 億回の学習後も Q 値が増加傾向にあることから，学習が収束していないことがわかる．なお，Fig. 23 に見られる Q 値の増加は，守備アームを用いない場合と比較して守備アームの展開に関する状態 s_a に価値が伝播した結果である．

守備アームを用いることで，「ロボットがボールと接触する」機会が増加するため守備率の向上が得られると仮定したが，Fig. 21 の結果では，守備アームを展開する行動の追加により守備率が低下した．本実験では，行動数の増加に対して報酬関数が守備アームを用いない場合と同様であったため，守備アームを用いる行動（a=3，4）は守備アームを用いない行動（a=1，2）に対し，行動罰が高い点だけが異なる行動として学習されたと言える．従って報酬を高めることのない行動が増加した結果となり，学習時間の増加が生じた．Fig. 23 に示されるように 1 億回の学習後も行動価値 Q が増加傾向にあることからも，同様のことが言える．これらの結果から，守備アームを用いた場合の学習に関して，学習回数を増加させることで守備アームを用いない場合と同様の守備率に到達する可能性があることが考えられる．また，本実験では，守備アームを用いることに対する報酬を設定しなかったため，守備率や走行距離が示す結果から，守備アームを用いる利点が見られなかった．

実際のゴールキーパロボットにおいては, 高速で向かってくるボールに対してゴールキーパロボットが搭載しているドライブモジュールによる移動のみでは, ゴールを守備することが困難である場合が多く, 守備アームの展開によってロボットの守備面積が拡大し, 守備率の向上が期待出来る. 次節では, 守備アームの展開を考慮した報酬関数の導入について述べる.

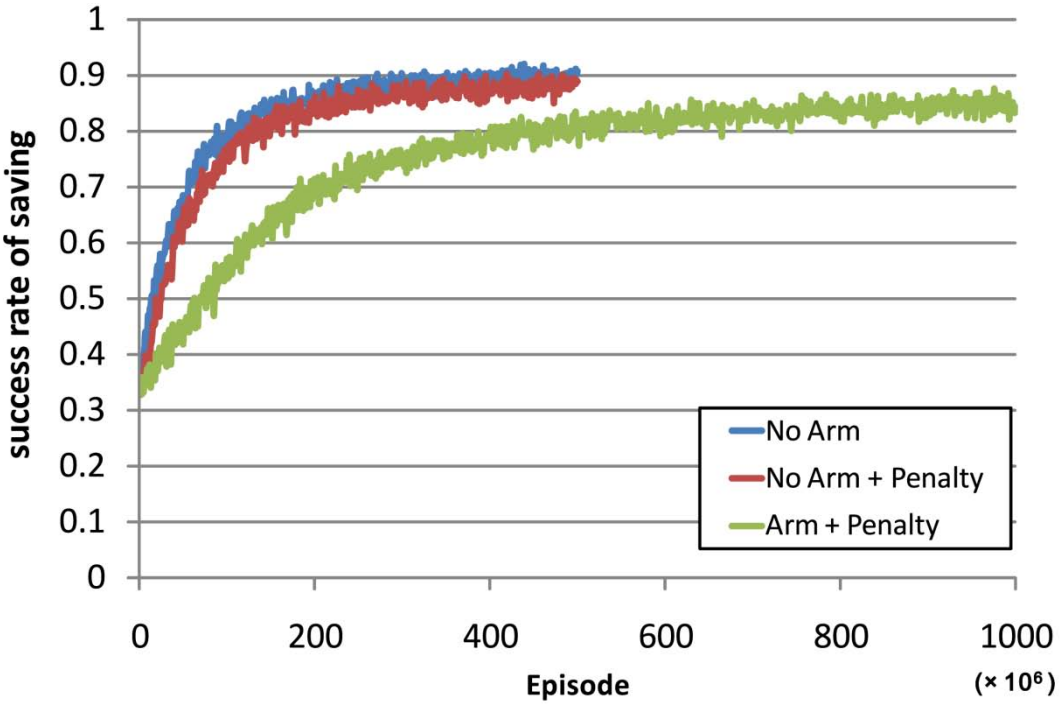


Fig. 21 Result of Saving Success Rate at With-Arm Condition

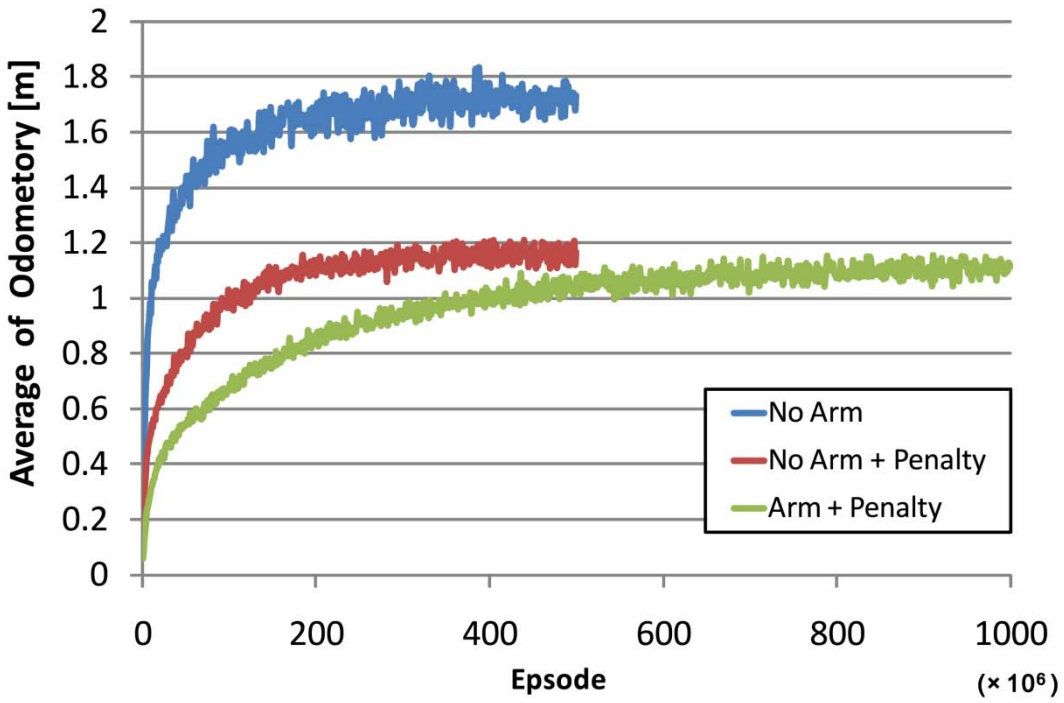


Fig. 22 Result of Odometory at With-Arm Condition

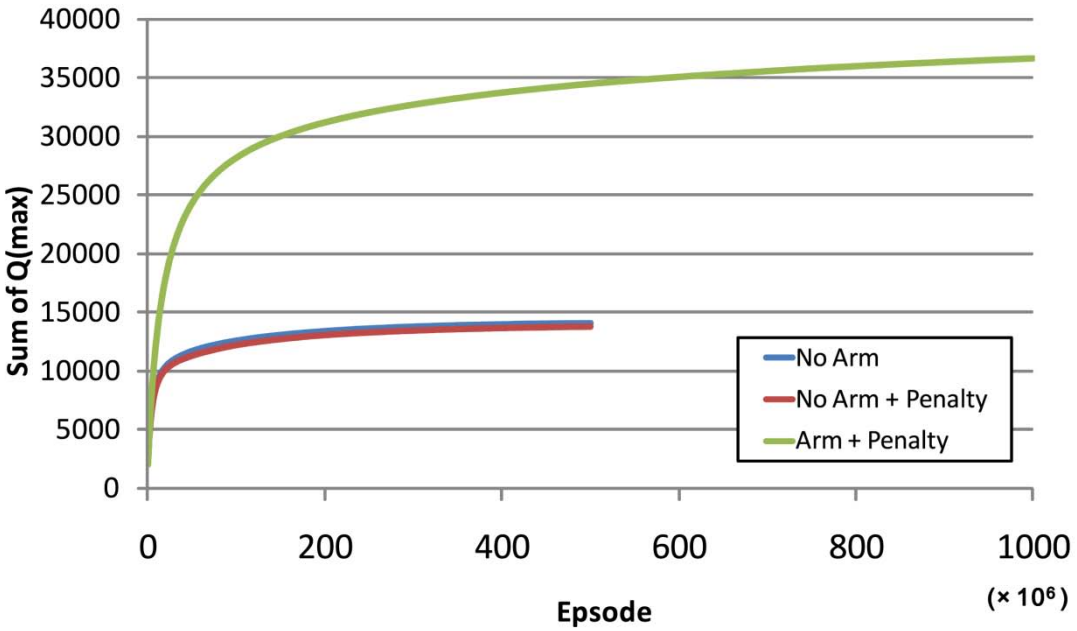


Fig. 23 Result of Sum of Q Max at With-Arm Condition

3.4.4　守備アーム展開を考慮した報酬関数の検証実験

Fig. 15 に示す報酬関数を用いた場合，ボールがエンドラインを越える際の，ロボットとボールとの相対距離|Δy|が 1.0 [m] を越えた場合，その行動価値 Q には負の報酬（罰）が与られるため，|Δy|を小さくする方向への行動を選択していた場合でも，学習が進むにつれてその行動は選択されなくなる．また，|Δy|が 1.0[m]以上の場合，0.1 [m]毎に r = -0.1 の罰が追加されるため，正の報酬の伝播よりも，罰の蓄積が大きく発生したと考えられる．Q 学習における罰の設定は，「その行動を抑制する」効果を持つ．そこで，本節では，ボールがエンドラインを越えた際に，ボールとの相対距離|Δy|が 1.0[m]以上であった状態と行動を抑制すべき行動ではなく，より高い価値を得る可能性を持つ状態と仮定し，罰を与えないものとした．

また，実際のゴールキーパロボットによる守備行動では，確実にゴールを守備するためには可能な限りロボットの中心でボールを受ける必要がある．従って本節では，ロボットの中心でボールを受けた状態に対してより高い報酬を与えることで，意図する行動が獲得できると仮定した．

これらの仮定を踏まえ，本節では式(3.10)に示すガウス関数を用いた Fig. 24 のような報酬関数を用いた．

$$r = \exp\left\{-\frac{(2/\Delta y)^2}{2\sigma^2}\right\} \tag{3.10}$$

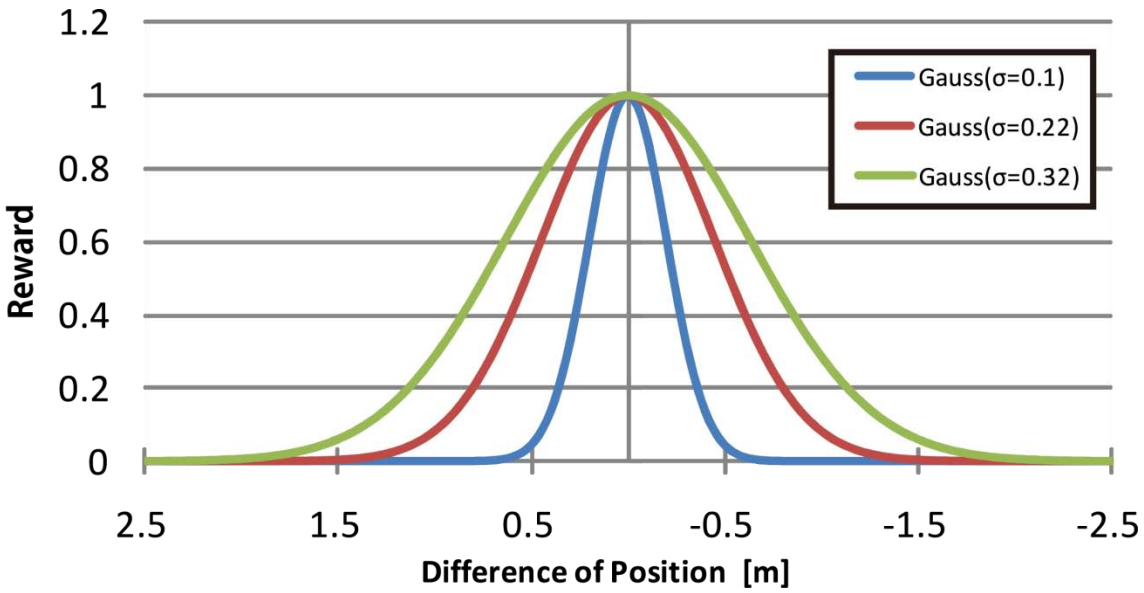


Fig. 24　Reward Function using Gaussian Function

守備アームを用いた場合の守備行動について，式(3.10)に示すガウス関数を用いて分散σの変化に対する守備率の変化について評価する．学習回数は100,000,000回（1億回）とし，分散σは，0.1，0.22，0.32とした．評価項目は線形報酬を与えた場合と同様に，1000通りのシュートパターン

に対する守備率と平均走行距離，行動価値Qの最大値の総和から評価する．実験結果をFig. 25~Fig. 27に示す．まず，Fig. 25~Fig. 27の全ての結果において，各評価項目は増加傾向にあり，設定した状態変数において1億回の学習では学習の収束に至っていないと言える．

Fig. 25に示す結果では，ガウス分布を用いた報酬関数により，守備アームを用いた場合において約90 [%] の守備率を達成した．また，分散値が増大するにつれて線形報酬の結果に近づくことから，報酬が線形である場合，ロボットの躯体にボールが接触する距離$|\Delta y|$までロボットが移動していなかったと考えられる．また他の要因として，ロボットがボールに接触することで，守備が成功したか否かを判断するための報酬を設定していなかったことも要因の一つとしてあげられる．守備アームを展開することでロボットの守備面積は幅89.0[cm]まで拡張されるが，$\sigma = 0.1$の条件では幅約1.0[m]の範囲において報酬が与えられるため，ロボットとボールが接触することで報酬が入る結果となり，守備率が向上したと考えられる．

またFig. 26に示す平均走行距離の結果では，$\sigma = 0.1$の条件において線形報酬よりも低い走行距離を示した．$\sigma = 0.22$，0.32の結果において，線形報酬を与えた場合よりも走行距離が上回った原因として，ゴールエリア外に向かうボールの軌道に対しても守備行動のために移動したためと考えられる．線形報酬の場合，$|\Delta y|$が1.0 [m]を越えると罰が与えられるため，ゴールエリア外に向かうボールに対して，ロボットは守備行動を行わないように学習することが想定されるが，$\sigma = 0.22$，0.32の場合，$|\Delta y|$がそれぞれ1.5 [m]，2.5[m]まで報酬が与えられることに加え，罰が与えられないため，$|\Delta y|$を小さくする方向へロボットが向かう行動が獲得されたと考えられる．Fig. 27に示す行動価値Qの値の増大からも同様のことが言える．一方で$\sigma = 0.1$の場合はゴールエリア外に向かうボールに対しては行動を行わず，停止行動を優先して行ったと考えられる．この結果から，σ値の変化によって意図する行動の獲得を促すことができている．

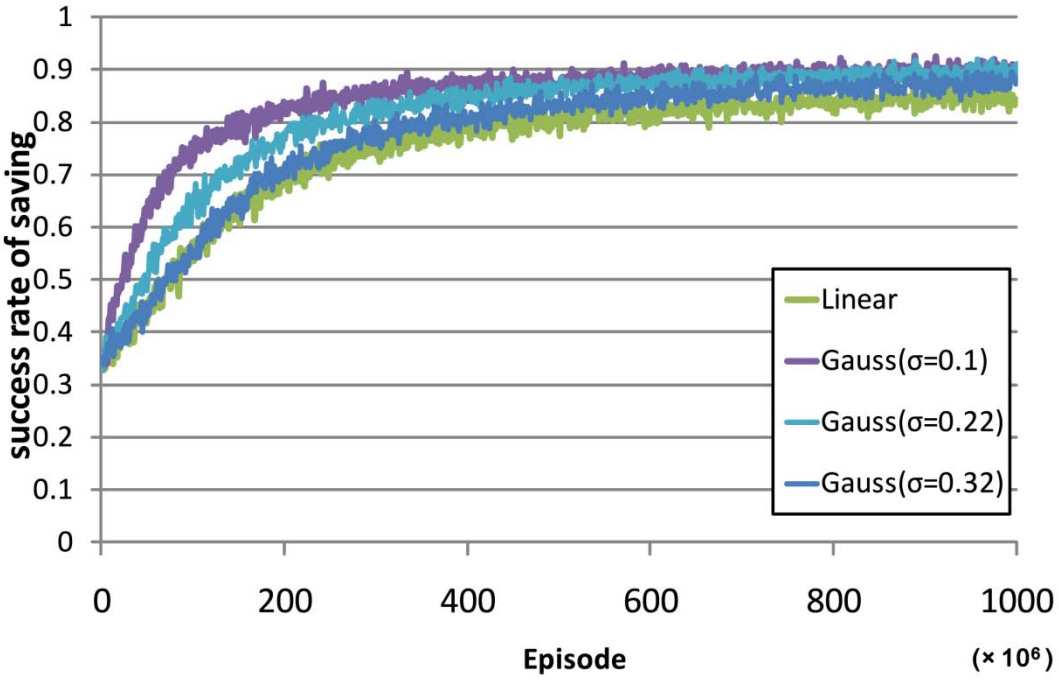


Fig. 25 Result of Saving Success Rate at With-Arm Condition using Gaussian Reward Function

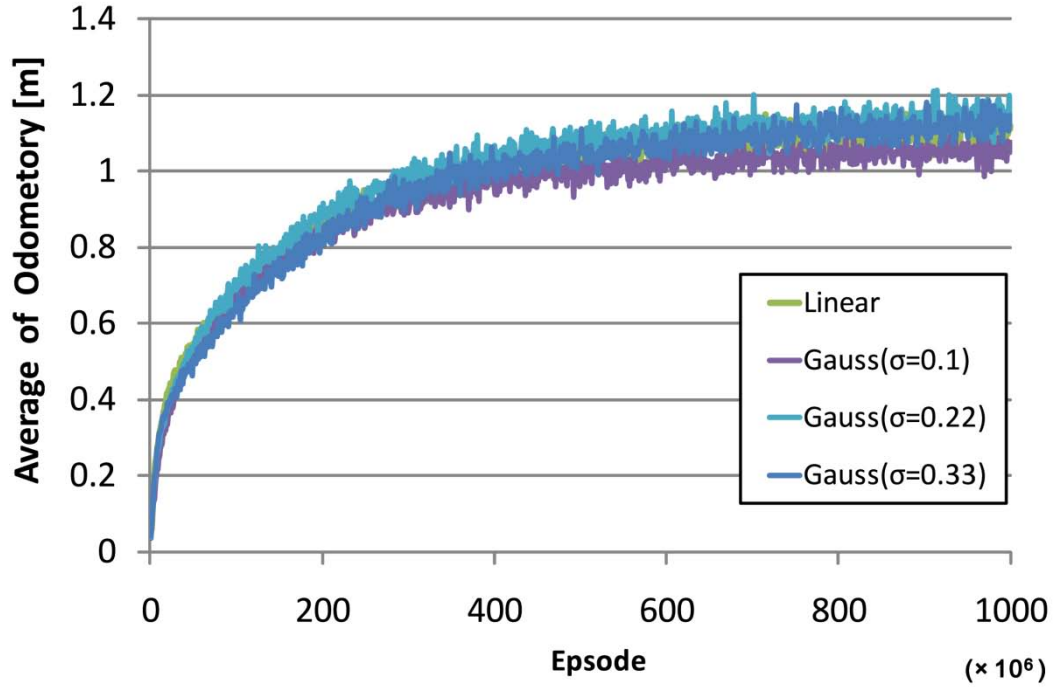


Fig. 26 Result of Odometory at With-Arm Condition using Gaussian Reward Function

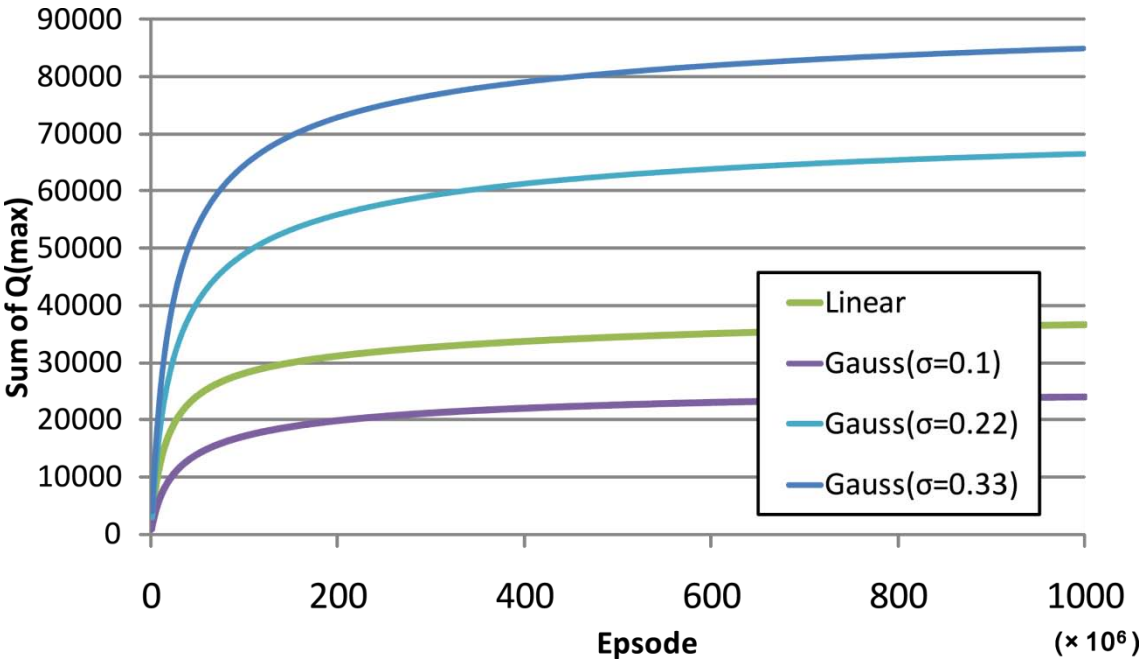


Fig. 27 Result of Sum of Q-max at With-Arm Condition using Gaussian Reward Function

3.4.5 Softmax手法による行動獲得実験

本節では，方策としてe-greedy手法とsoftmax手法を用いることで方策による行動獲得の差について評価を行った．実験環境として守備アームを用いる場合を想定し，報酬関数はσ= 0.1によるガウス分布を用いた．実験条件をTable 3にまとめる．本実験では，式(3.1)に示すギブス関数の温度τを1.0e7として式(3.11)に従って減衰するものとした．Table 3に示す条件における，ε-greedy手法εの減衰と，softmax手法の温度τの減衰をFig. 28に示す．実験結果をFig. 29~Fig. 31に示す．

$$\tau = \frac{1.0}{1.0 + (k/N)} \tag{3.11}$$

Fig. 29に示す結果では，softmax手法において学習5,000,000回（5百万回）まで守備率の向上が見られたが，その後，4千万回の学習まで守備率の低下が見られ，最終的な守備率は75[%]となった．またε-greedy手法では，学習4千万回まで徐々に守備率が向上し，最終的な守備率は90[%]を示した．Fig. 30に示す平均走行距離の結果では，softmax手法において，2千万回の学習後0.9[m]の平均走行距離を示した．ε-greedy手法では，平均走行距離は徐々に増加し最終的な平均走行距離として1.05[m]を示した．Fig. 31では，1億回の学習後も両者のQ値の総和が増加傾向にあり，学習が収束していない状態であった．また，状態変数や報酬関数が同じであるにも関わらず，両者のQ値の総和には約1.4倍の差が見られた．

Fig. 29が示すようにsoftmax手法の結果において，学習5百万回後に守備率の低下が見られた原因としてsoftmax手法の方策が学習初期に，探索行動よりも開拓行動を中心とした行動に遷移したことがあげられる．Fig. 14に示すようにsoftmax手法では温度tの減少に伴い，Q値の大きい行動が高い確率で選択されるようになるが，学習が進むにつれて各状態における行動毎のQ値の差は大きくなるため，softmax行動選択を用いた場合，開拓行動への遷移はε-greedyを用いた場合と比較するとより早く行われたと言える．一方でε-greedy手法では，Q値に関わらず一定の確率によりランダムな行動が発生するため，様々な状態と行動に対して学習を行ったと考えられる．従ってsoftmax手法では，探索行動の際に報酬を得た行動のみに対し，開拓行動により行動を獲得し，ε-greedy手法では学習全体を通して探索行動と開拓行動を適切に行ったものと考えられる．最終的な守備率結果では，softmax手法はε-greedy手法より低い守備率を示しており，Fig. 30に示す平均走行距離を見ると，ε-greedy手法比べ平均走行距離が小さいことからも，探索行動において報酬が得られた行動にのみ行動を行っていたと考えられる. 同様に, Fig. 31に示す結果からも，softmax手法では行動価値Qの総和が，ε-greedy手法と比較して小さいことから，特定の行動にのみ開拓行動を行い, 探索行動時に報酬を得ることが出来なかった状態と行動に関しては開拓行動を行わなかったと考えることが出来る.

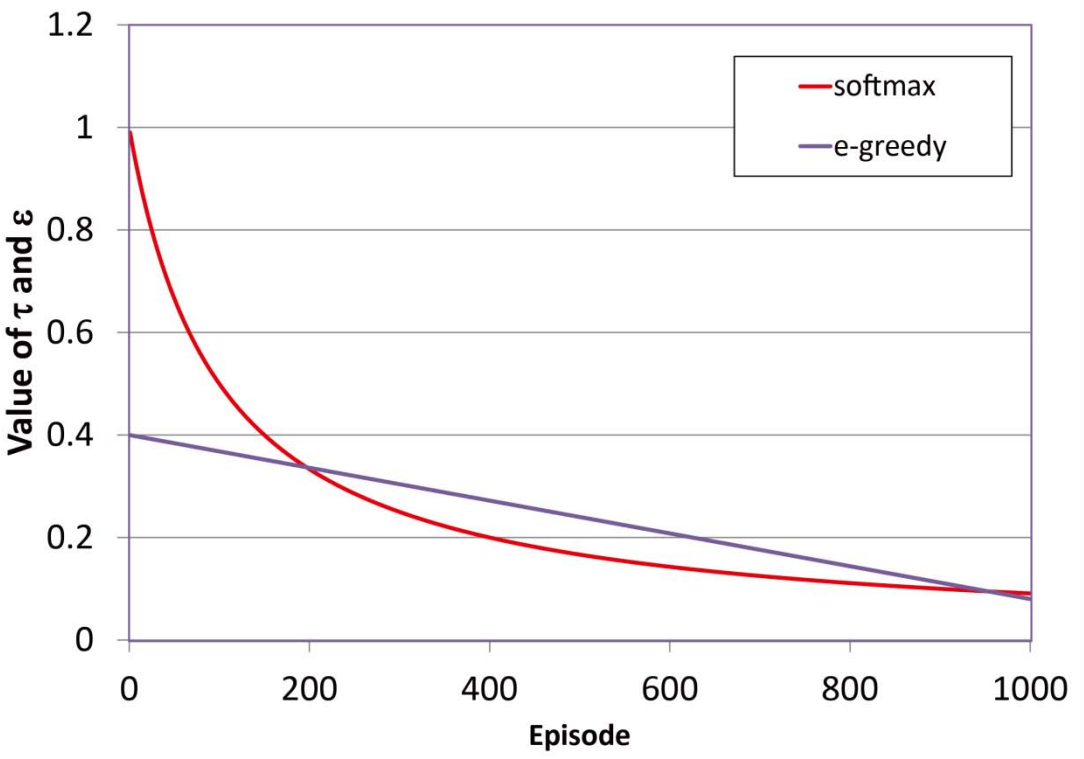


Fig. 28　Comparing value of τ with ε

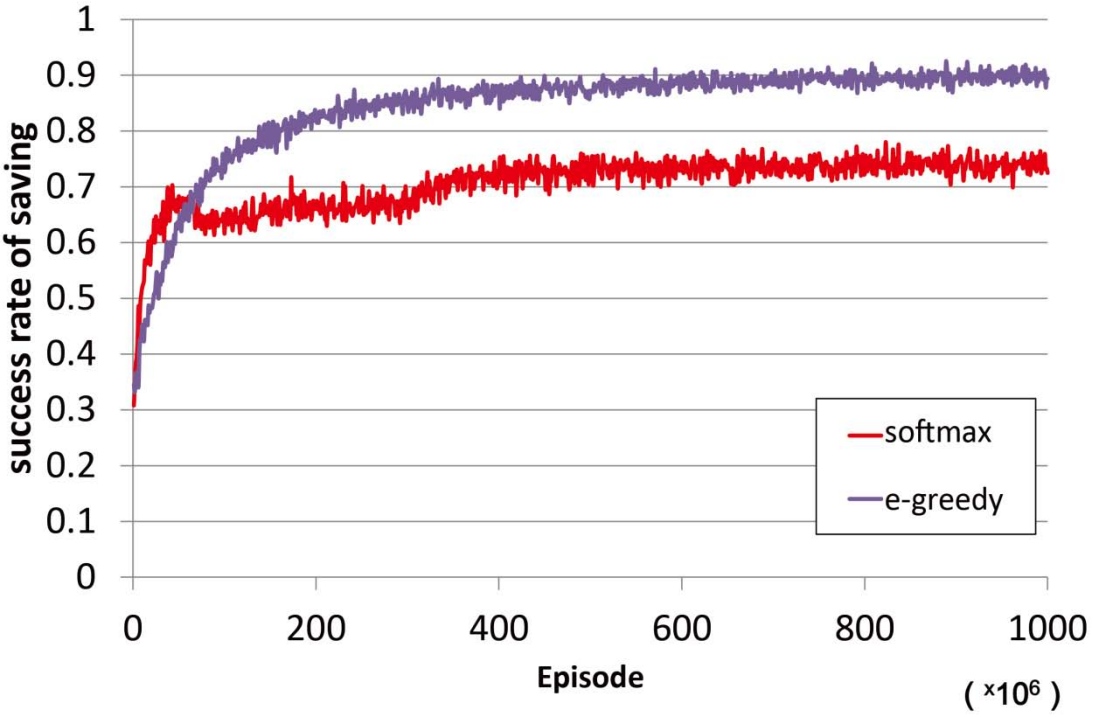


Fig. 29 Result of Saving Success Rate Comparing ε-greedy Method with Softmax Method

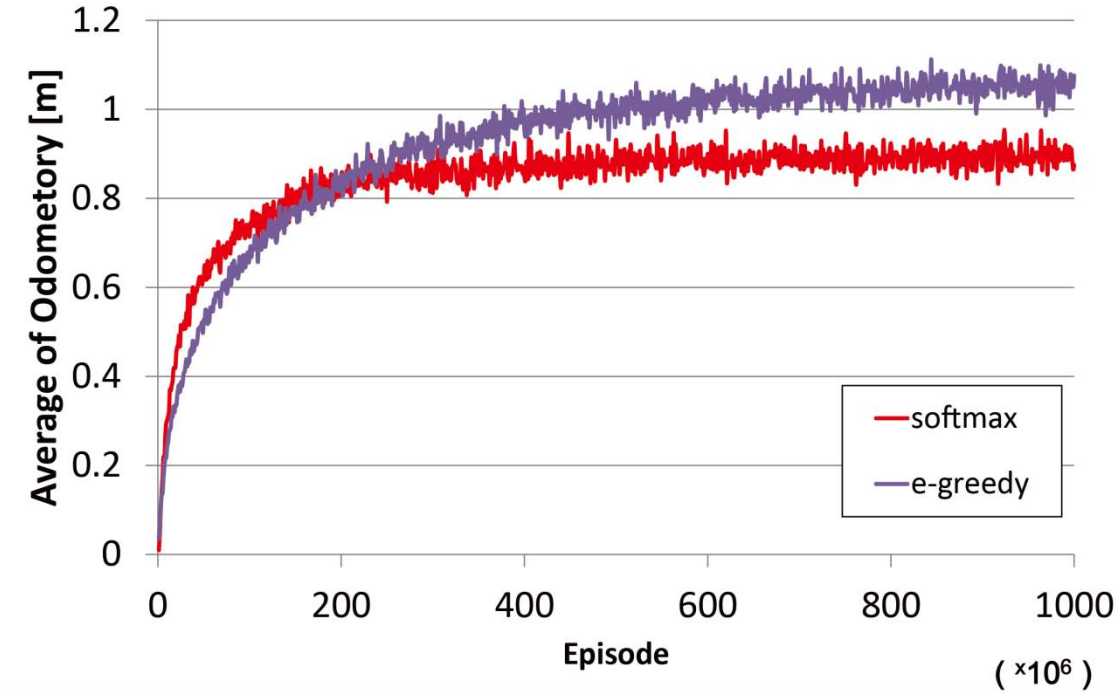


Fig. 30 Result of Odometory Comparing ε-greedy Method with Softmax Method

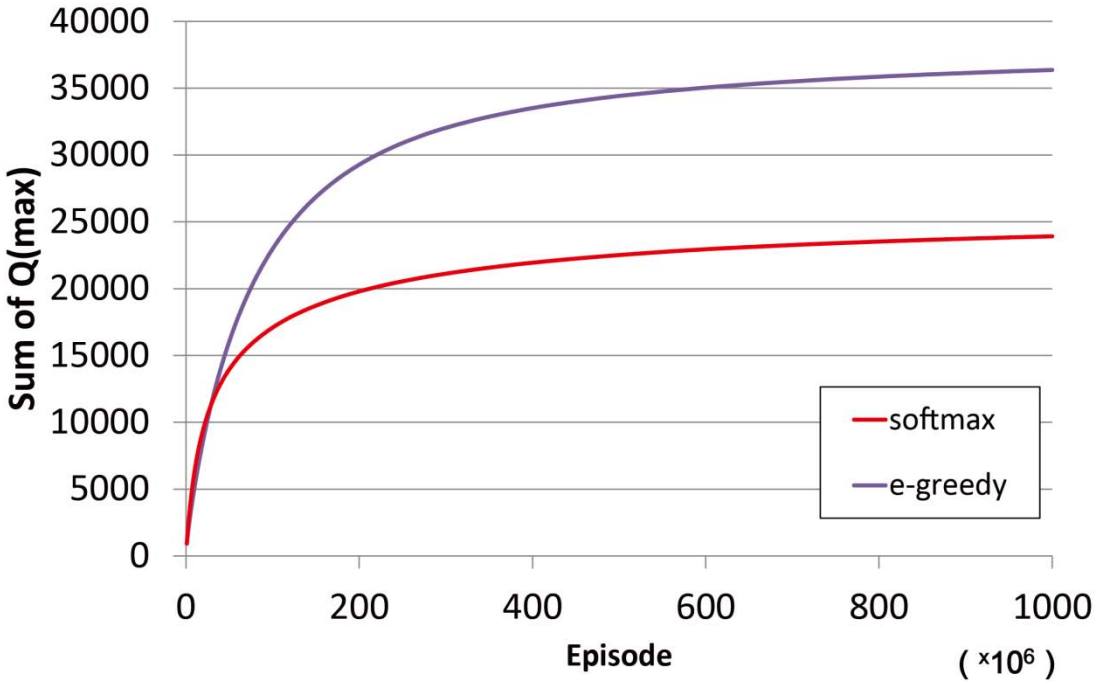


Fig. 31 Result of sum of Qmax Comparing ε-greedy Method with Softmax Method

Table 3 Condition of Experiment at Comparing ε-greedy Method with Softmax Method

Item	*Details*
Robot Position	$[x_r][y_r]$ = [1][20]
Ball Position	$[x_b][y_b]$ = [60][30]
Ball Speed	$[v_x][v_y]$ = [10][10]
Arm Condition	$[s_a]$ = [3]
Action	$[a]$ = [5]: Stop / Right-Move / Left-Move / Right-Move with Arm / Left-Move with Arm
Learning rate	α = 0.1
Discount Rate	γ = 0.5
Policy	ε-greedy / softmax
ε value	$\varepsilon = 0.4$
Damping coefficient	$\xi = 0.8$
Temperature parameter	τ = 1.0e7
Reward	$r = \exp\left\{-\frac{(2/\Delta y)^2}{2\sigma^2}\right\} \quad (\sigma = 0.1)$
Penalty	$r_{t+1} = -1.0 \quad (y_r < -1.0 \,\|\| \quad y_r > 1.0)$
Moving Penalty	$r_{t+1} = -1.0e-6 \quad (a = 1 \,\|\| a = 2)$ $r_{t+1} = -1.0e-6 \quad (a = 3 \,\|\| a = 4)$
Episode	50,000,000

3.4.6　動力学シミュレーションを用いた強化学習評価実験

離散空間内で学習した結果を連続空間で行動するロボットへ適用した場合，マルコフ決定過程が満たされない環境であることや離散化誤差等の影響により，学習結果の実行において致命的な性能低下が生じると言われている[10]．従って連続空間におけるマルコフ性の欠如や離散化誤差による影響を低減するためには，可能な限り細かく離散化する必要がある．本章において行った離散空間におけるシミュレーションでは，0.1[m]間隔の離散化を行った．本節では，動力学シミュレーション内で動作するロボットに対し，離散空間内で行った強化学習結果を適用して守備行動の評価実験を行った．なお，動力学シミュレーションには，ラッセルスミスらが開発したODE（Open Dynamics Engine）シミュレーションを用いた．ODEは他の動力学シミュレーションと比較すると，高速に計算が可能であり，安定性が高い[54]．実験にはHibikino-Musashiにおいて新福らが開発したシミュレータを用いた[55]．Fig. 32にシミュレーション環境の一例を示す．検証するボール軌道は，Fig. 33に示す2種類の軌道を対象とし，ロボットの正面3.0[m]遠方，左右方向へそれぞれ0.3[m]離れた地点からゴールへ向かうものとした．ボール速度は1.0[m/s]とした．評価項目として，各試行に対し，ロボットが観測した各状態と，実行した各行動に関する行動価値Qの時系列変化により，離散空間におけるシミュレーションとの相関について検証を行った．また学習結果にはFig. 29~Fig. 31に示す学習結果を用いた．Fig. 34Fig. 35にボール速度1.0[m/s]に対するε-greedy手法とsoftmax手法に関する行動価値Qの時系列変化，Fig. 36Fig. 37に守備動作を行った際の離散空間におけるロボットの軌道とODEシミュレーションにおけるロボットの軌道を示す．

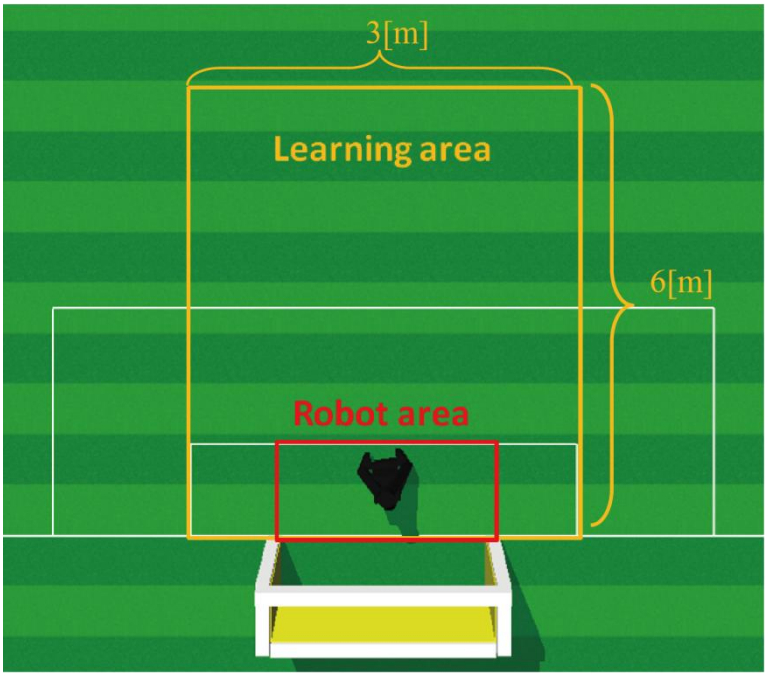


Fig. 32　Overview of ODE Simulator

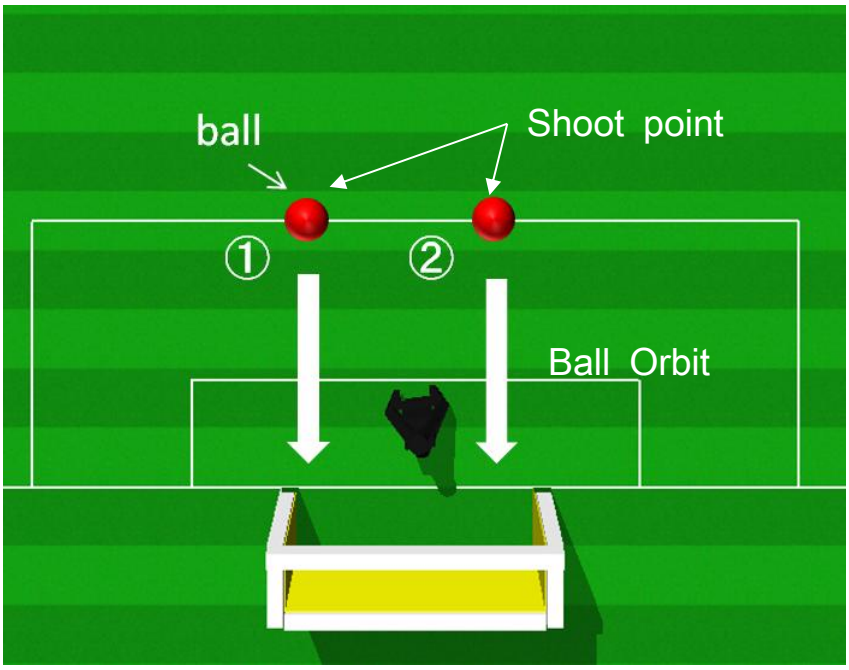


Fig. 33 Experimental Condition of ODE Simulation

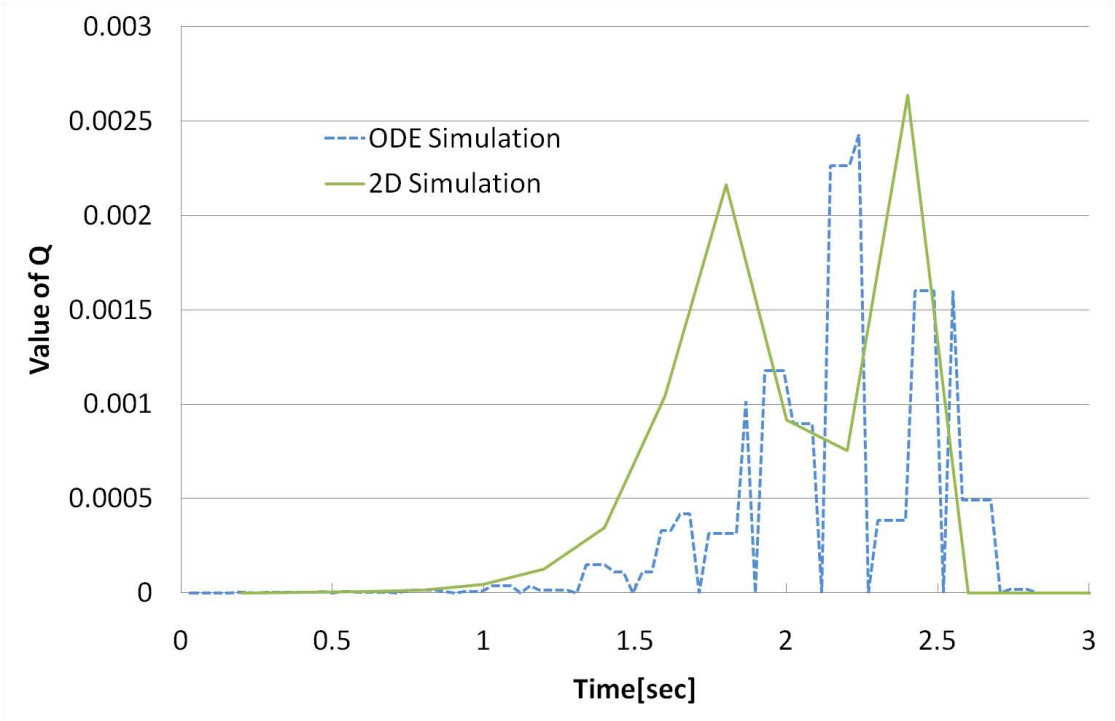


Fig. 34 Time Series Variation of Q (ε-greedy, v_x=1.0[m/s], Ball Trajectory 1)

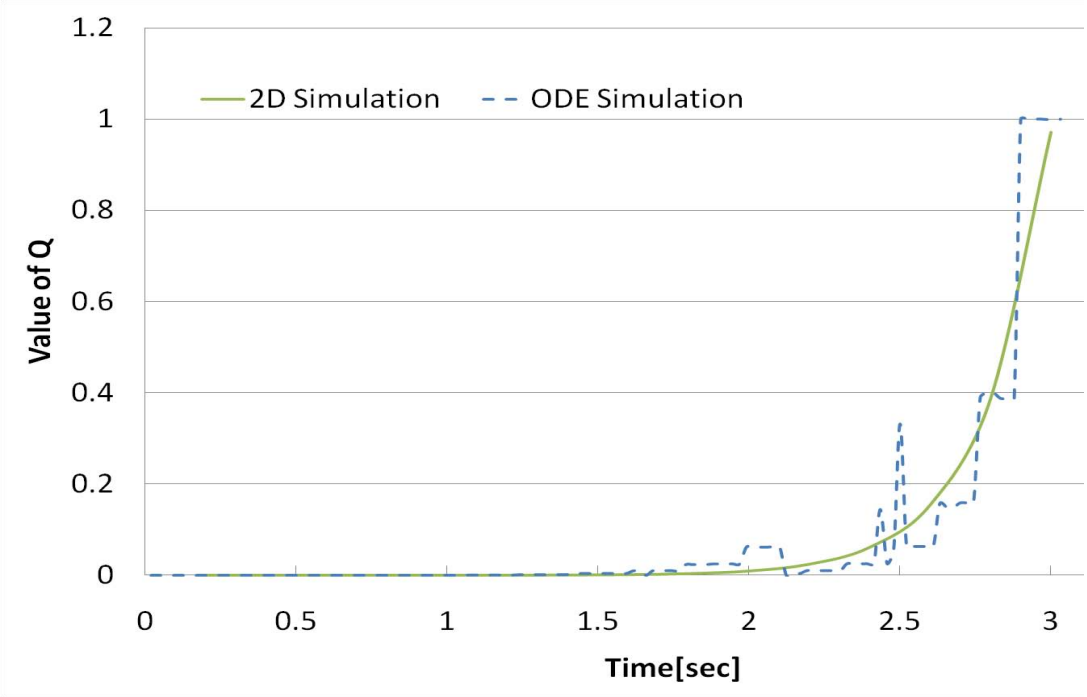


Fig. 35 Time Series Variation of Q (softmax, v_x=1.0[m/s], Ball Trajectory 1)

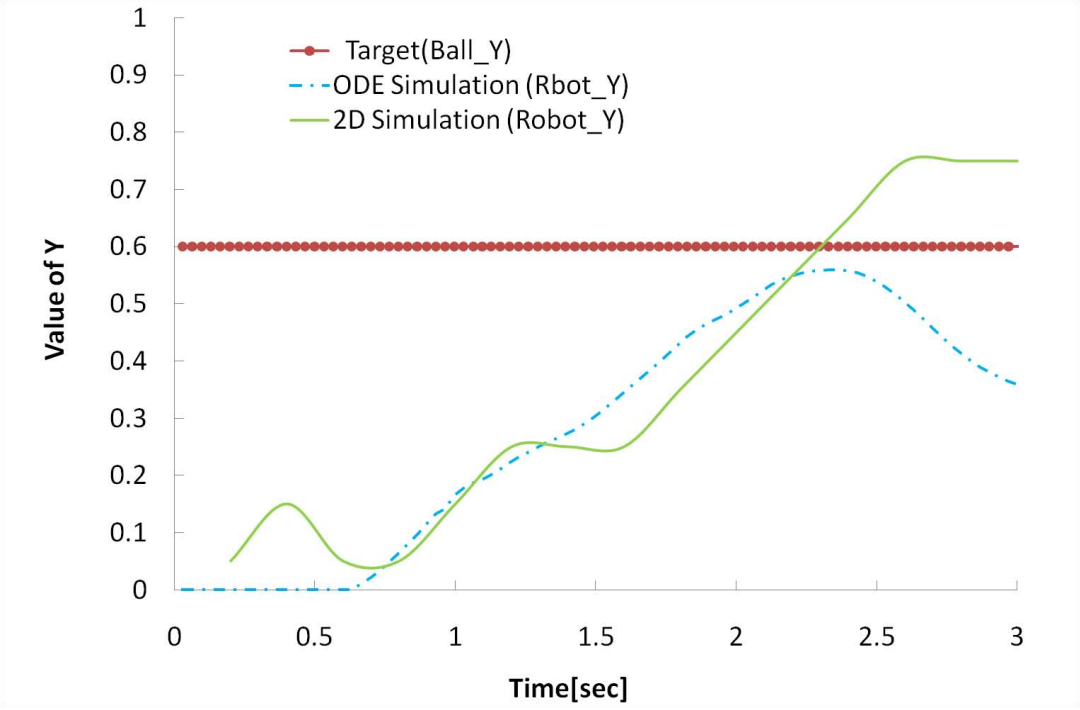


Fig. 36 Time Series Variation of Robot and Ball

(ε-greedy, v_x=1.0[m/s], Ball Trajectory 1)

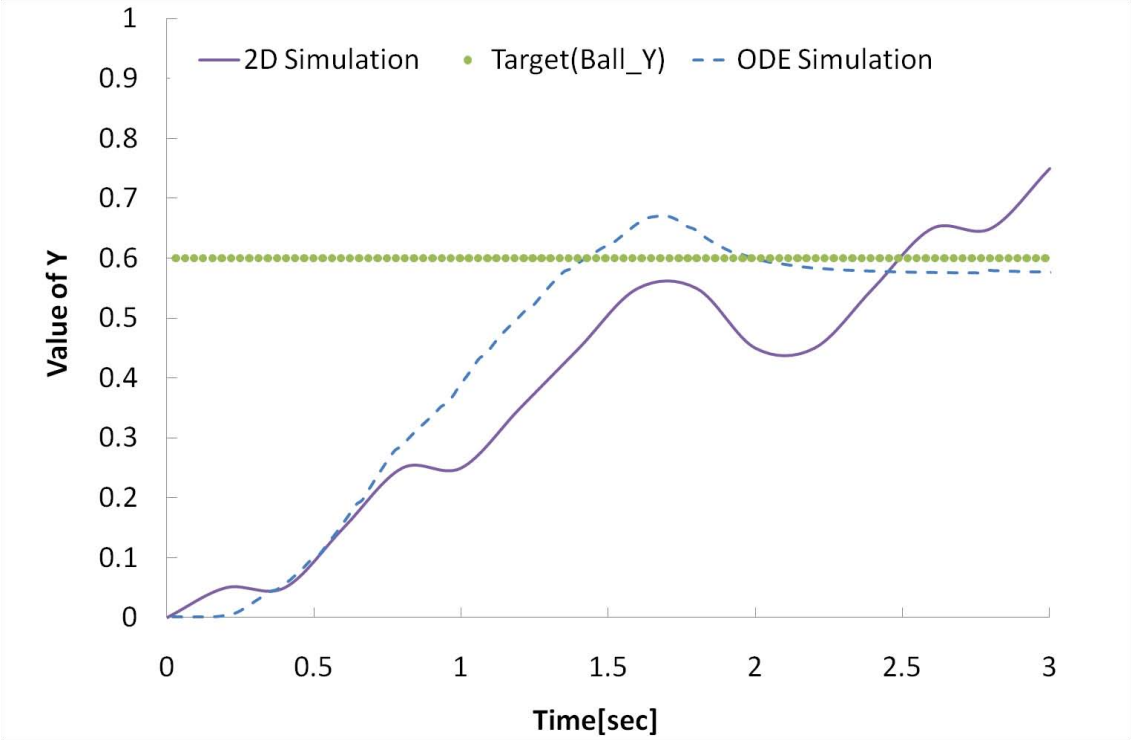


Fig. 37 Time Series Variation of Robot and Ball (softmax, v_x=1.0[m/s], Ball Trajectory 1)

Fig. 34に示す結果から，ボールとロボットの相対距離が近づくほど，行動価値Qが増加しており，離散化シミュレーションにより行動価値Qの伝搬が行われていることがわかる．離散化シミュレーションによる結果では，1.7[sec]まで行動価値Qが増加しており，2.3[sec]まで減少している．この結果は，観測情報に基づいて行動した結果，離散化シミュレーションではより行動価値Qの高い行動に繋がる行動であったが，ODEシミュレーション環境ではロボットの加速度等が考慮されるため，行動価値Qの高い行動を実行した結果，予期していない状態に陥ったためと考えられる．ODEシミュレーションが示す行動価値Qからも，行動価値Qが断続的に0を示しており，連続空間内で行動した結果，離散化シミュレーションでは行動価値Qが伝搬していなかった状態に陥っていることが分かる．しかし，動力学シミュレーションが示す結果では，行動価値Qが0になった場合においても，次状態では直ちに高い行動価値Qを持つ状態へ遷移している．これは，離散化シミュレーションでは，各ステップにおいて選択した行動が直ちに実行され状態が遷移するのに対し，動力学シミュレーションでは，ロボットが行動を選択した後もロボットの移動に伴う慣性力により，選択した行動とは異なる移動が行われていることが要因としてあげられる．Fig. 36に示す結果から，ロボットは離散化シミュレーション結果が示す行動と類似した移動軌跡を描いており，主に左方向への移動を選択していたことがわかる．この結果から，Fig. 33に示すボール軌道①の状態は，観測情報に対し左への行動に高い行動価値Qが伝搬しており，極所的

に行動価値Qが0を示す場合が存在してもロボットが慣性力により移動する際に，左方向への行動を高い頻度で選択していたことが考えられる．

Fig. 35とFig. 37に示す結果では，ボールとロボットとの相対距離が0.01[m]程の状態において行動価値Qが増大していることがわかる．これはsoftmax手法において開拓行動が多く行われた結果と考えられる．また，Fig. 35における最終的な行動価値QがFig. 34に示すε-greedy手法に対して高い値を示していることからも，Fig. 33に示す状態に対して開拓行動が行われていたと言える．また，Fig. 37の結果からε-greedyによる学習結果よりも早くボール方向へ移動しており，想定した実験条件に対して守備行動が迅速に行われたと言える．

動力学シミュレーションの結果では，ε-greedyによる結果に対してsoftmax手法によって学習した結果が1.0[sec]程早く守備行動を行ったことから，開拓行動を多く行うことにより行動の最適化が行われていると考えられる．また，Fig. 36とFig. 37に示す結果から，離散化シミュレーションによる学習結果を用いた動力学シミュレーション環境下における守備行動おいて，致命的な性能の低下は見られなかった．

3.4.7　Musashiを用いた強化学習評価実験

ODEを用いたシミュレーションでは，ロボットの位置やボールの位置に関して正確な観測値が得られるが，実機を用いた環境では，ロボットの位置やボールの位置に関して観測誤差や推定誤差が含まれるため，離散化シミュレーションによる学習結果が正しく参照されないという問題が想定される．

本節では，Fig. 29~Fig. 31に示す学習結果を用いて，実環境における学習結果の検証を行った．実験にはMusashiのゴールキーパロボットを用い，Fig. 33に示す動力学シミュレーションと同様の環境下において，各ボール軌道に対し，10回ずつの観測を行った．ボールの速度は1.0，2.0，3.0[m/s]とした．Fig. 38-(a)に各ボール速度に対するボール軌道①の守備率結果，Fig. 38-(b)に各ボール速度に対するボール軌道②の守備率結果を示す．また，Fig. 39~Fig. 41にボール軌道①の各ボール速度に対して，ε-greedy手法により行った学習結果を用いた際の，ロボットが観測した状態と選択した行動に関する行動価値Qの時系列変化を示す．Fig. 42~Fig. 44には同様の条件に対し，softmax手法による学習結果を用いた際の行動価値Qの時系列変化を示す．Fig. 45~Fig. 50には，各方策による学習結果を用いた際の，ボール軌道①における各ボール速度に対するゴールキーパロボットの観測結果と移動軌跡，さらに動力学シミュレーションにおけるボール位置とロボットの移動軌跡を示した．

Fig. 38に示す結果から，1.0~3.0[m/s]までのボール速度に対して、ロボットが学習結果に基づいて行動したことにより，守備行動を行ったと言える．また，全てのボール速度に対して10回の試行に対する守備率は，softmax手法による学習結果を用いた場合が，ε-greedy手法による学習結果を用いた場合に比べ，高い守備率を示した．この結果は，Fig. 37が示す結果のように，ボー

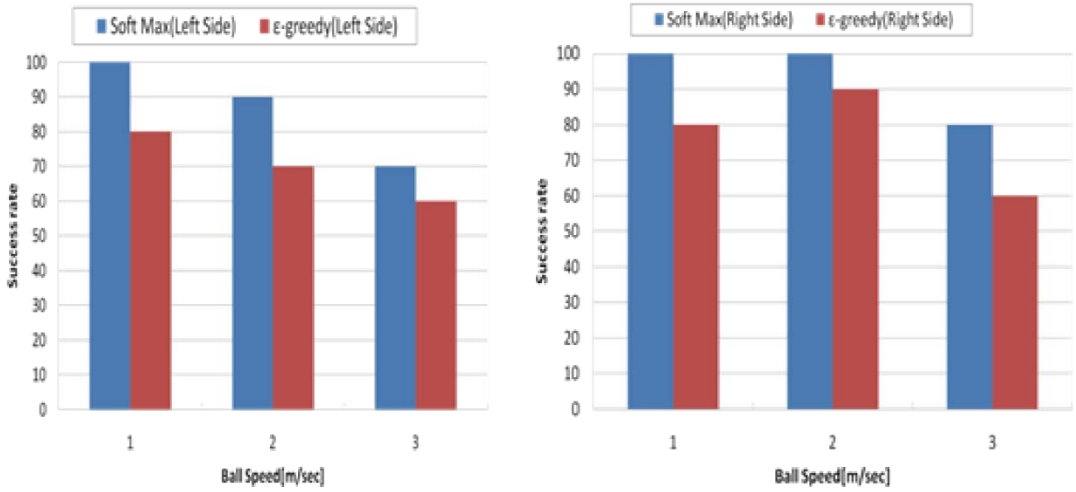
ル軌道①の状態に対する開拓行動が多く行われた結果だと考えられる．また，ボールの移動速度が増加するにつれて，両手法において守備率が低下する傾向にあると言える．

速度の増加に伴い守備率が低下する原因として，離散化誤差による影響があげられる．本実験では，離散化シミュレーションにおいてロボットの移動速度を1.0[m/s]と設定しており，1[step] = 0.1[sec]となるため，1.0[m/s]で移動するボールは離散空間において1マスずつ移動する．しかし，ボールの速度が2.0[m/s]，3.0[m/s]で移動する場合は，2マス，もしくは3マスの間隔をあけてボールが移動する形に表現される．実環境下における実験では，ロボットが自己位置とボール位置，速度を観測した際，各パラメータを0.1[m]間隔（ボール速度は1.0[m/s]間隔）で離散化し，学習結果を参照したため，ある自己位置とボール速度において，学習時には観測しなかった位置にボールを観測する可能性がある．このような場合，観測結果に対する行動価値Q値が0を示すため，ロボットは行動を選択出来ず，停止する結果となる．Fig. 39~Fig. 44に示す結果から，両手法においてボールの速度が増加するにつれて各観測状態と選択した行動に対する行動価値Qが減少していることからも同様のことが考えられる．この結果から，ボールの移動速度として10.0[m/s]を想定する場合は，1[setp]=0.01[sec]程度の離散化が必要である．

Fig. 39~Fig. 41に示すε-greedy手法による学習結果を用いた場合の行動価値Qの時系列変化から，ボール速度v_x = 1.0[m/s] では行動価値Qの最大値が0.0041，v_x = 2.0[m/s] では行動価値Qの最大値が0.0048，v_x = 3.0[m/s] では行動価値Qの最大値が0.013を示した．また，ボールの移動速度が増加する程，1試行中に高い価値が参照される頻度が低下しており，観測結果に対応する行動価値Qに対する価値が伝搬していない．しかし，Fig. 45~Fig. 47に示す結果では，ボールに対して相対距離を減少させる方向への移動を行っていることから，行動価値Qが示す値が小さい場合であっても，守備すべき方向への行動に価値Qが伝搬していれば守備行動を実行可能である．

Fig. 42~Fig. 44に示すsoftmax手法による学習結果を用いた場合の結果では，v_x = 1.0[m/s] において行動価値Qの最大値が0.35，v_x = 2.0[m/s] では行動価値Qの最大値が0.19，v_x = 3.0[m/s] では行動価値Qの最大値が0.15を示した．また，ボールが近づくにつれ行動価値Qが徐々に上昇する傾向がみられ，ε-greedy手法と比較すると離散化誤差による影響が少ない．また，Fig. 48~Fig. 50に示すロボットとボールの軌道の変化から，動力学シミュレーション上における動作に近い行動によって守備行動を行っている．Fig. 51にsoftmax手法を用いた際のゴールキーパロボットの守備行動の一例を示す．ゴールキーパロボットはゴールの右側へ向かうボールに対し，徐々に相対距離を短くするように行動した．また，10回における試行では，ボールに対して停止行動と右方向への行動を交互に繰り返す場合と，滑らかに右方向へ移動する場合が見られた．この動作から，観測情報が学習結果と連続的に一致した場合には，連続的な滑らかな動作により守備行動を実行し，観測情報が学習結果と断続的に一致した場合には，停止行動と移動を交互に行うような動作をしたと考えられる．Fig. 42~Fig. 44， Fig. 48~Fig. 50に示す結果からも，softmax手法では本実験において行った環境に対する学習が十分行われていたと考えられる．また，動力学

シミュレーションとMusashiを用いた実験から，softmax手法において開拓行動が十分に行われた結果，学習結果を連続空間内における行動選択に用いた場合，離散化誤差の影響を受けにくく，連続空間への適用に適している．



(a)Result of Ball Orbit 1　　(b)Result of Ball Orbit 2

Fig. 38　Saving Success Rate at Real Environment

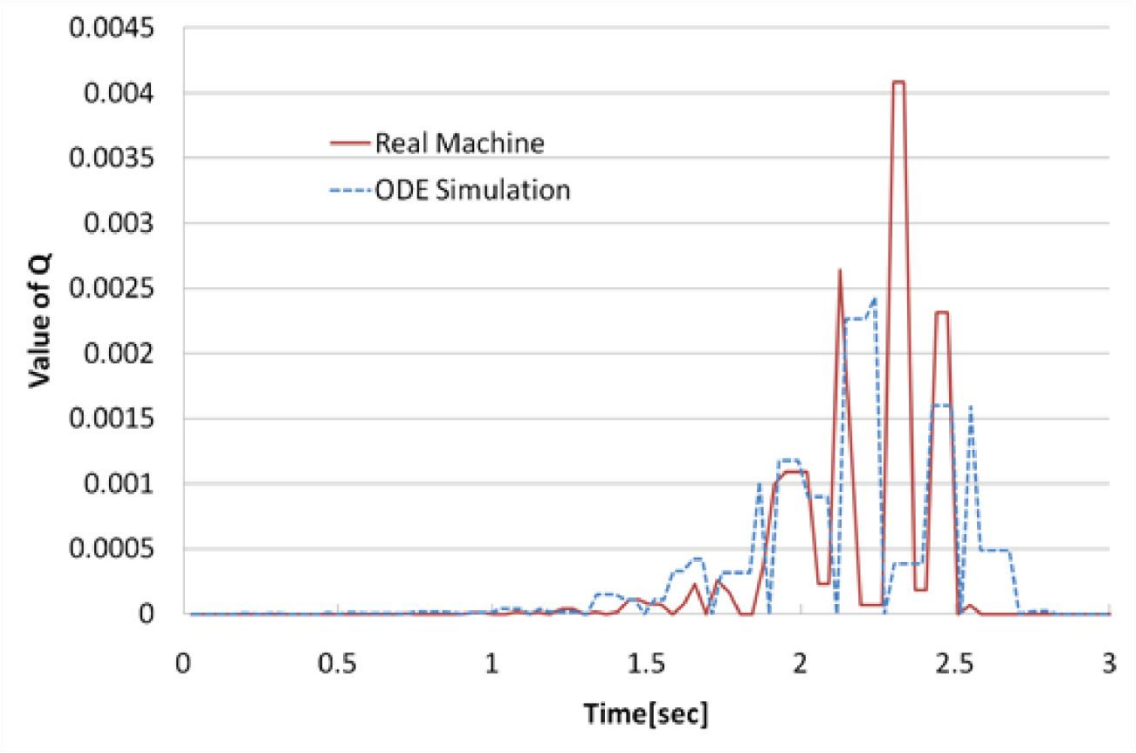


Fig. 39 Time Series Variation of Q

(Method : e-greedy, Ball Orbit : 1, v_x=1.0[m/s], Success Situation)

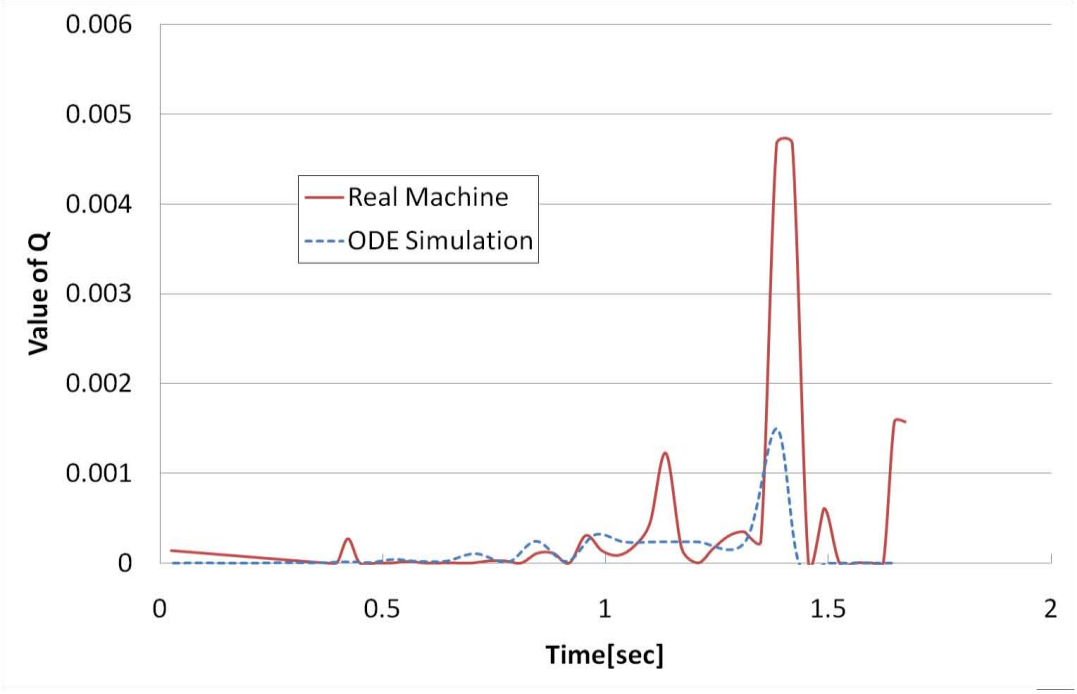


Fig. 40 Time Series Variation of Q

(Method : e-greedy, Ball Orbit : 1, v_x=2.0[m/s], Success Situation)

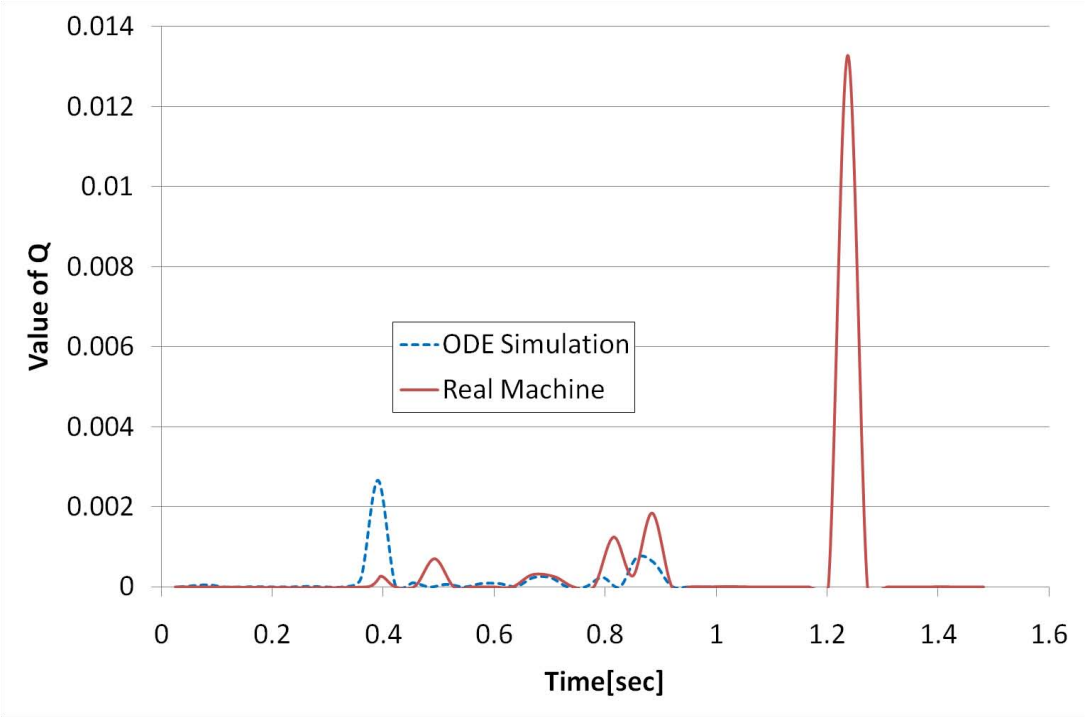


Fig. 41 Time Series Variation of Q

(Method : e-greedy, Ball Orbit : 1, v_x=3.0[m/s], Success Situation)

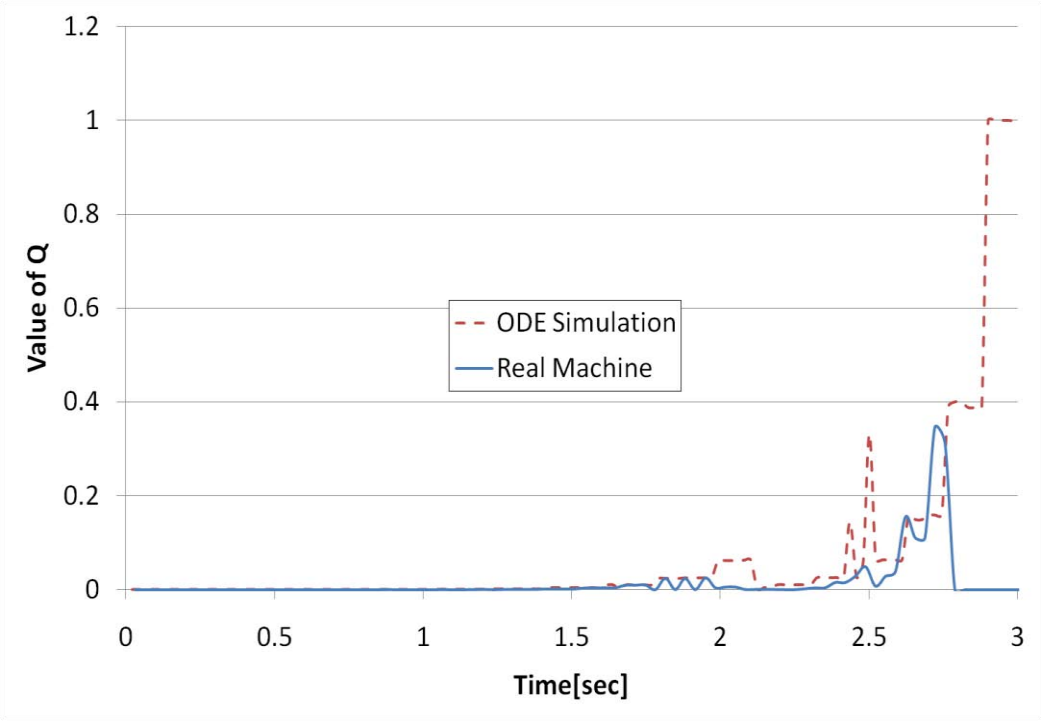


Fig. 42 Time Series Variation of Q

(Method : softmax, Ball Orbit : 1, v_x=1.0[m/s], Success Situation)

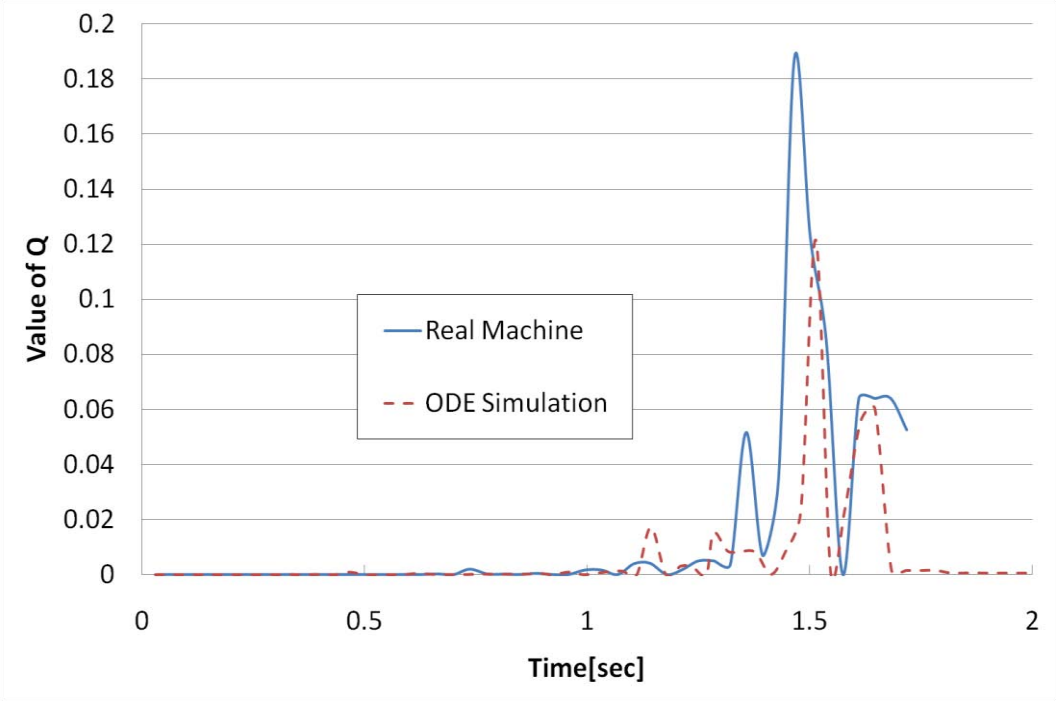


Fig. 43 Time Series Variation of Q

(Method : softmax, Ball Orbit : 1, v_x=2.0[m/s], Success Situation)

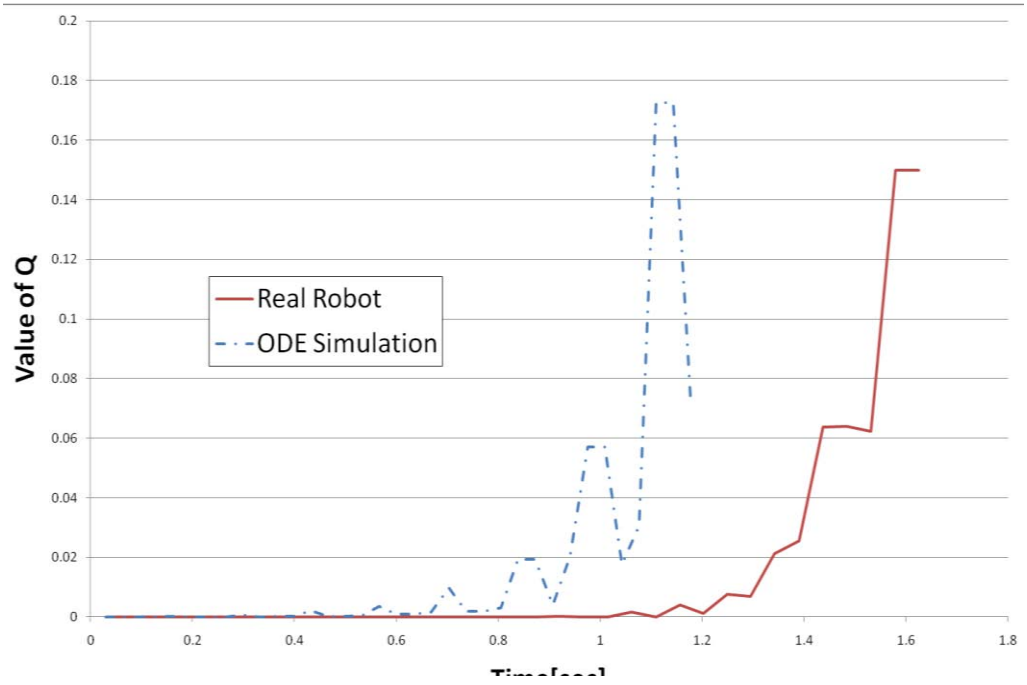


Fig. 44 Time Series Variation of Q

(Method : softmax, Ball Orbit : 1, v_x=3.0[m/s], Success Situation)

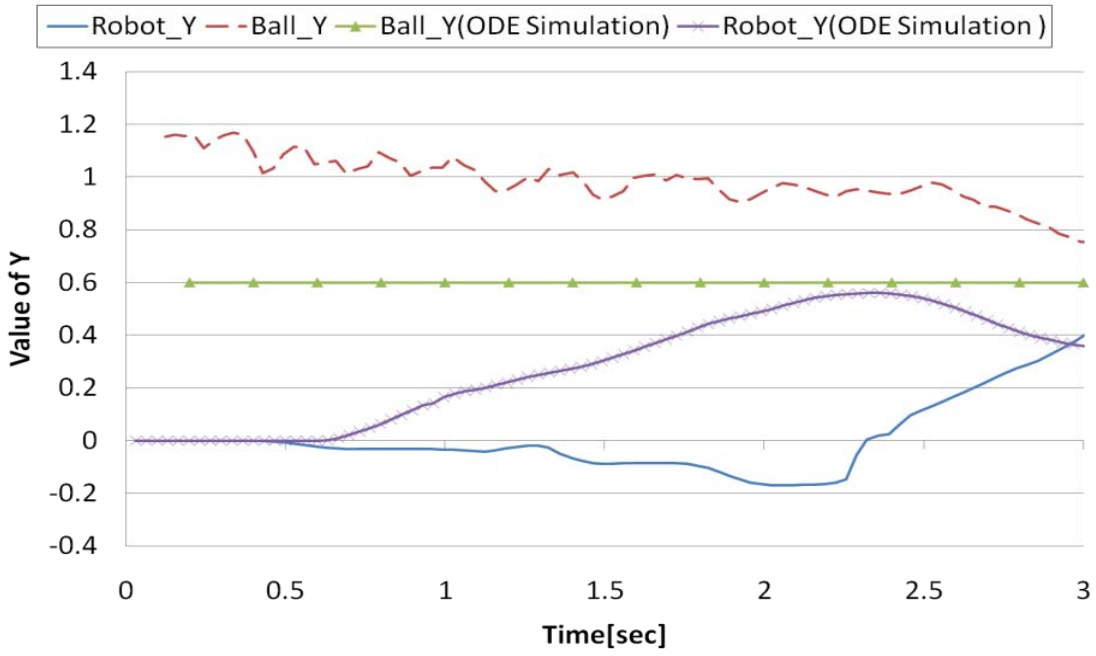


Fig. 45 Time Series Variation of Robot and Ball Orbit

(Method : e-greedy, Ball Orbit : 1, v_x=1.0[m/s], Success Situation)

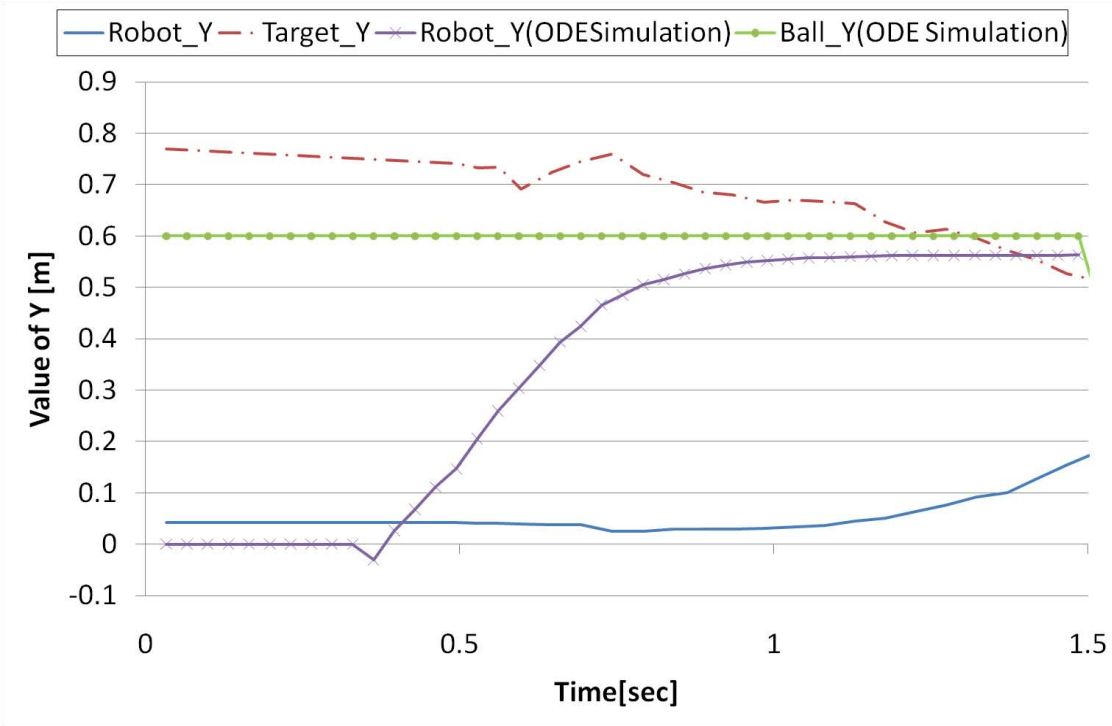


Fig. 46 Time Series Variation of Robot and Ball Orbit

(Method : e-greedy, Ball Orbit : 1, v_x=2.0[m/s], Success Situation)

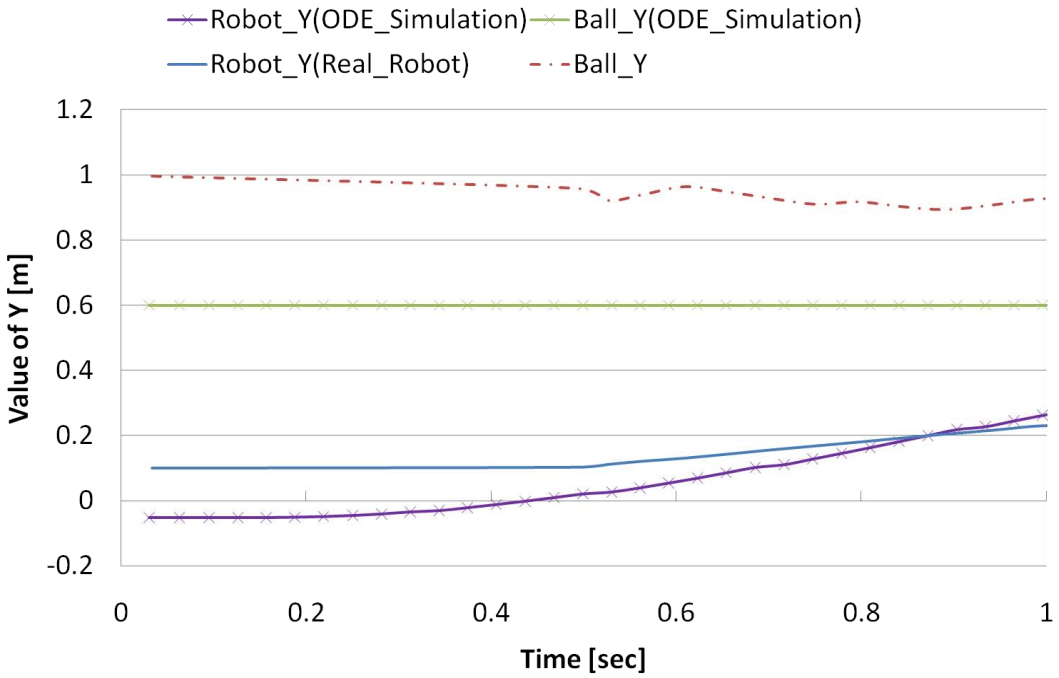


Fig. 47 Time Series Variation of Robot and Ball Orbit

(Method : e-greedy, Ball Orbit : 1, v_x=3.0[m/s], Success Situation)

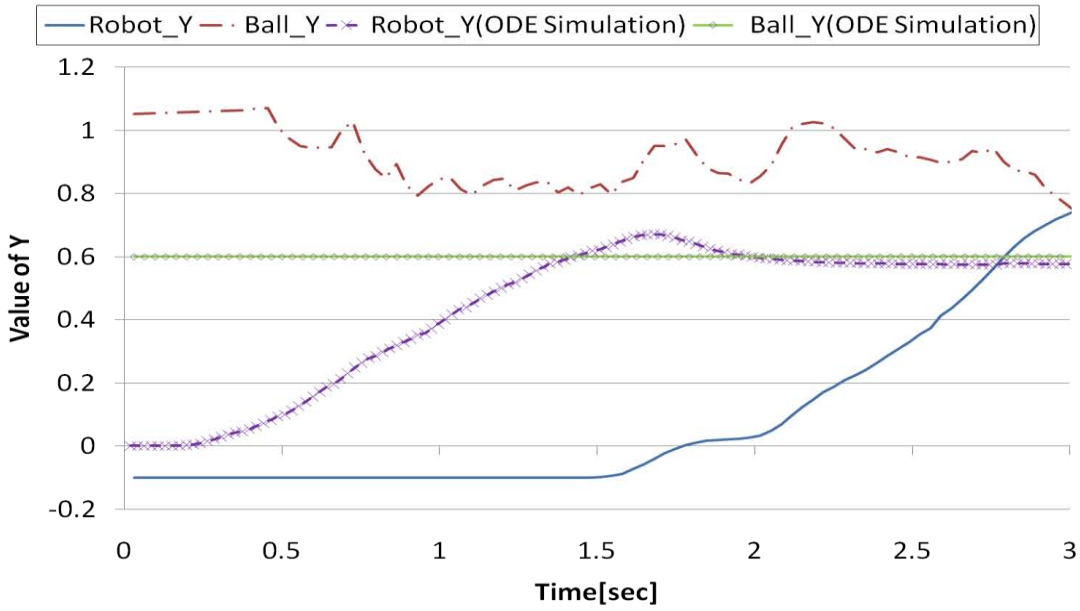


Fig. 48 Time Series Variation of Robot and Ball Orbit

(Method : softmax, Ball Orbit : 1, v_x=1.0[m/s], Success Situation)

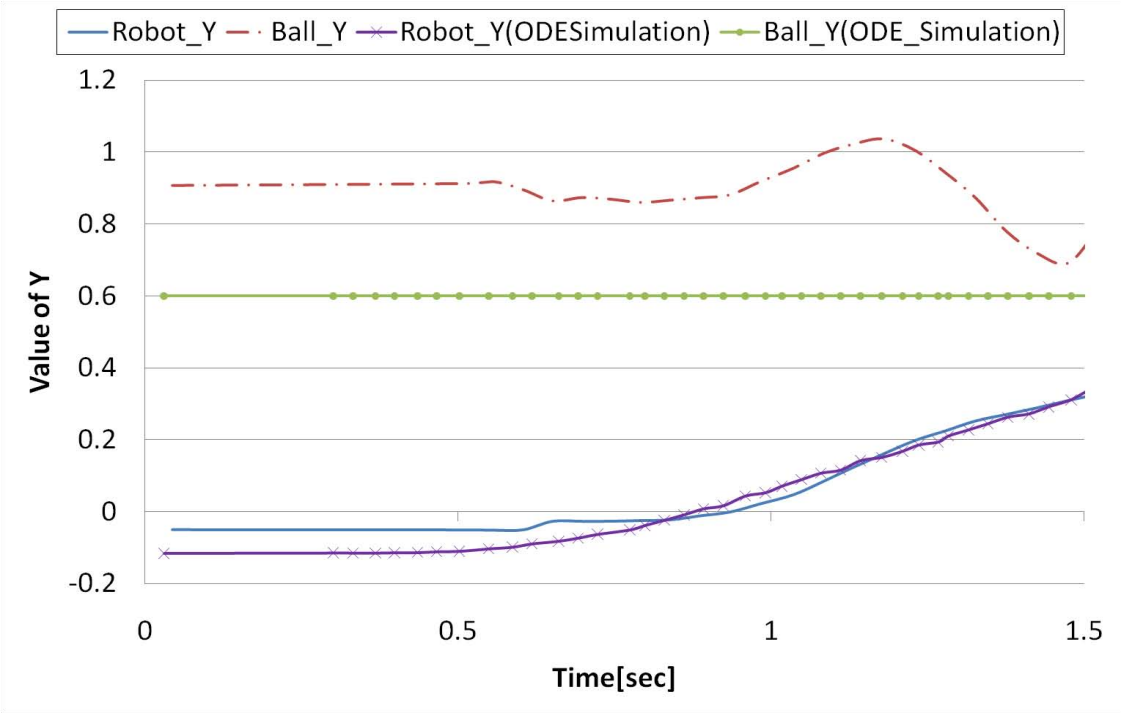


Fig. 49 Time Series Variation of Robot and Ball Orbit

(Method : softmax, Ball Orbit : 1, v_x=2.0[m/s], Success Situation)

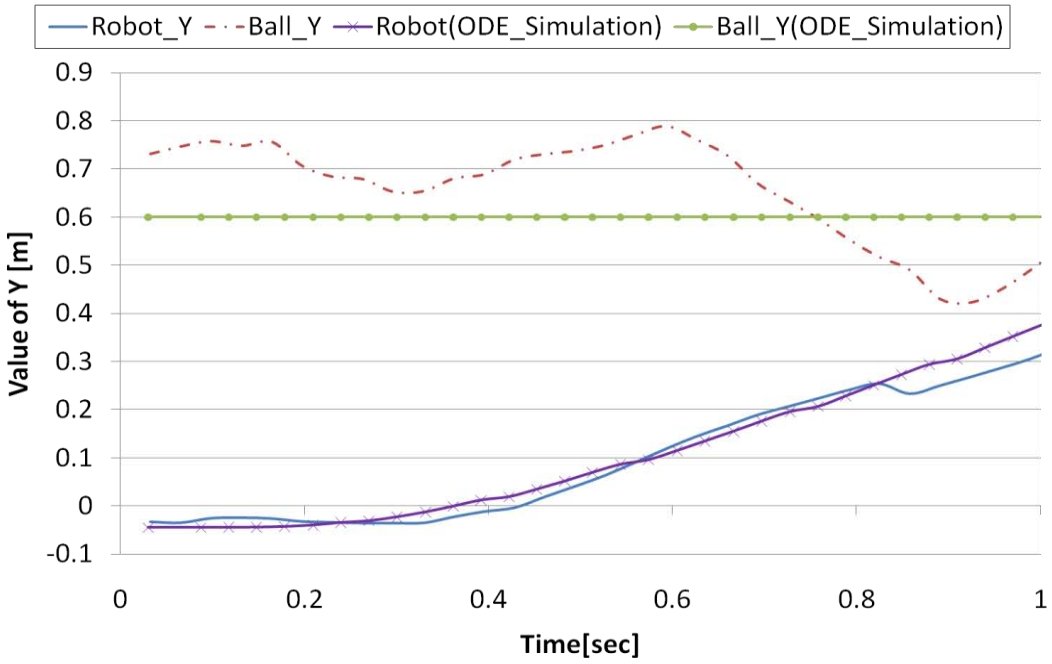


Fig. 50 Time Series Variation of Robot and Ball Orbit

(Method : softmax, Ball Orbit : 1, v_x=3.0[m/s], Success Situation)

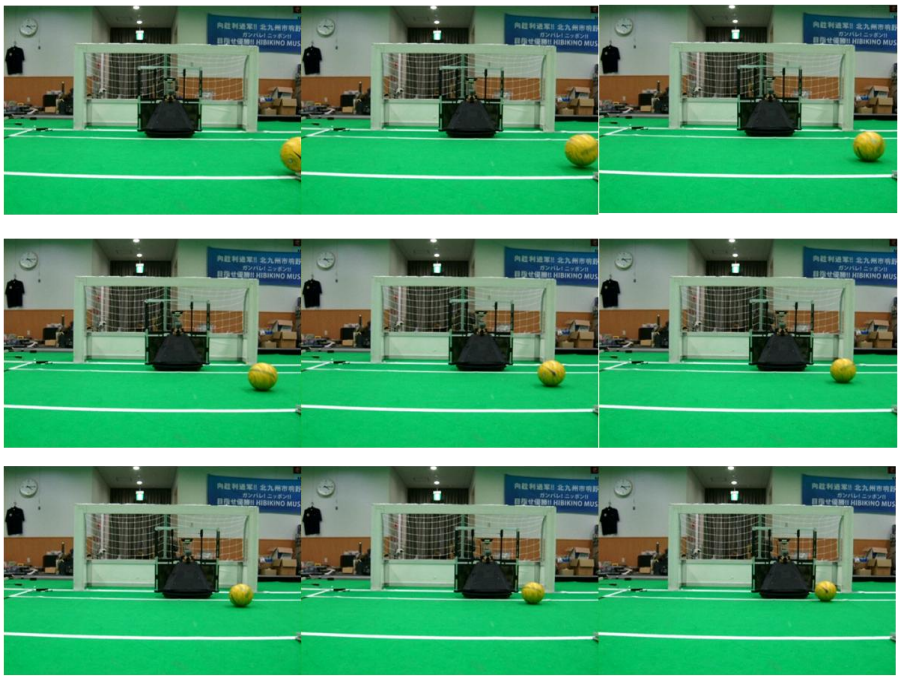


Fig. 51 Example of Saving Motion of Goalkeeper Robot (Method : softmax, Ball Orbit : 1, v_x=1.0[m/s], Success Situation)

3.5 考察

離散空間におけるシミュレーション結果では，e-greedy手法とガウス分布による報酬関数を用いることで守備アームを伴う行動に関して約90[%]の守備率を示した．また，行動罰を与えることで行動の最適化を促している．1000通のシュートパターンに対し，守備率が最大90[%]に留まった原因として，シュートパターンに含まれるボール速度10.0[m/s]に対してが守備出来なかったことが挙げられる．ボール速度10.0[m/s]のパターンに関しては，約6[setp]程度でエンドラインを越えるためボールの到着予定地点とロボットの初期位置が0.6[m]以上離れている場合は守備出来ない. 同様にロボットの移動範囲が$y_r = 20$であることから, ボール速度が3.0[m/s]であっても，ボールの到着予定地点とロボットの初期位置が2.0[m]離れている場合は守備出来ない可能性がある．これら状況を考慮すると，離散化シミュレーションにおける90[%]の守備率は，ロボットの移動速度が1.0[m/s]である場合，守備可能な全てのボールに対する守備行動を学習出来ていると考えられる．また離散化シミュレーションにおいてsoftmax手法では75[%]の守備率を示したが，動力学シミュレーションや実機を用いた実験では，e-greedy手法よりも高い守備率を示した．本実験において行った動力学シミュレーション実験と実機実験では，実験条件がボール軌道：2通り，ボール速度：3通りであったため，両手法の比較には不十分であるが，学習において開拓行動を高い頻度で行ったと考えられるsoftmax手法は，離散空間における学習結果を連続空間内に適用する上で有効な手法である．実環境下において行った実験における行動価値Qの時系列変化の結果から，学習結果の実機への適応では，離散化誤差やロボットの観測誤差により離散化シミュレーションの学習過程で学習を行わなかった状態が得られ，行動の方策が得られない場合がある．一般的に離散化誤差の影響を軽減する方法として，離散間隔を小さくすることがあげられるが次元の呪いによる計算量の増加が問題となる．さらに離散間隔を細かくすることで，学習が行われない状態と行動も増加するため，理想的な手法ではない．本研究では，softmax手法による学習結果において開拓行動が行われた結果，実環境下においてもボールの移動速度3.0[m/s]に対して70[%]の守備率を示しており，開拓行動を多く行うことで離散化誤差による影響が軽減される可能性を示唆した．本実験では，softmax手法における温度tを固定の状態で検証を行ったが，様々な条件に対する学習と離散化誤差の影響の軽減のためには，学習初期にe-greedy手法による探索行動を行い，学習中期よりsoftmax手法による開拓行動を行う手法や，温度tによる学習結果の変化についてさらに検証する必要がある．また，行動価値Qの時系列変化の結果から，Q学習では高い価値が得られた行動価値Qに対して，近い状態であっても学習時に経験しなかった状態であった場合には行動価値Qが得られておらず連続空間への適用では，そのような状態が高い頻度で参照される可能性があることがわかった．これらの結果から，学習によって生成される行動価値Qに対し，一定回数の学習後にニューラルネットワークを用いて非線形近似を用いる手法や三角分布や動径基底関数等を用いた関数近似が有効である[63]-[65].

第四章

確率的情報共有による
位置情報の信頼性向上

第四章　確率的情報共有による位置情報の信頼性向上

4.1　はじめに

MSLでは，技術促進を目的として毎年，ルール更新が行われる．例として2007年にはサッカーフィールドのサイズが9x12[m]から12x18[m]に拡張されたことがあげられる．また，2007年までは両ゴールにそれぞれ黄と青の色が塗られていたのに対し, 2008年以降は人間のサッカーゴールと同じように同じ形状・色のゴールが両チームのゴールとして用いられるようになった．このルール更新に対応するため，Hibikino-MusashiではMCLを用いた自己位置推定法を導入したが，フィールドが広大であることから，白線情報が十分に取得出来ない場所においては，自己位置推定の精度が低下することが問題となっている．さらに，照明環境の変化や他のロボットによる白線情報の遮蔽，ホイールのスリップ，局所磁場により生じる方位センサの誤差等による影響からも自己位置の誤差は発生する．

これらの影響により，試合中に自己位置が誤推定された問題を“誘拐ロボット問題[42]”，その状態を“誘拐状態[43]”と呼ぶ．MCLにおける“誘拐ロボット問題”を解決するためのアプローチとして“センサリセット法(Sensor Resetting Method; 以下，SR法)[44]”や“拡張リセット法(Expansion Resetting Method; 以下，ER法)[43]”が提案された．SR法は，パーティクルの分布と観測情報に大きな差が生じた際に，各パーティクルが持つ情報を初期化し，パーティクルを再分布することで，自己位置を再推定する手法であり，SR法によりロボットは誘拐状態からの回復を図ることが出来る．ER法では，ロボットの誘拐状態を“近距離誘拐”と“長距離誘拐”分け，パーティクルの再分布範囲をロボット周辺から徐々に拡張していくことで, パーティクルが収束するまでの処理時間を短縮させる方法である．しかし，MSLのようにロボットが白線を認識することが出来る可視範囲よりもサッカーフィールドが大きい場合, ロボットは自己位置推定のための十分な白線情報を得ることが出来ず，各パーティクルの初期化と再分布を行っても，再び誘拐状態に陥る場合がある．このような状況に起因する誘拐ロボット問題は，ロボット単体による自己位置推定では困難である．

そこで本章では，MSLのマルチエージェントシステムとしての特徴に着目し，複数のロボット間における情報共有に基づいた協調自己位置推定法を提案する．提案したシステムでは，まずサッカーフィールド上に存在するボールをチームメイト間における共通のランドマークとして設定する．ここで，各ロボットは自己位置とボールとの相対距離・角度情報に基づいて，ボールの絶対位置を推定するため, ボールの位置は自己位置の推定誤差やボールに関する観測誤差を含んでいると考えられる．共有したチームメイトロボットのIGIを基にボールの真の位置を推定する．本章では，この推定を“共有推定”と呼び，各ロボットが観測したボールを“観測ボール”，

IGIに基づいて推定したボールを“推定ボール”と定義する．また，ロボットが誘拐状態に陥った際，誘拐状態のロボットが観測したボールの絶対位置は，他のロボットが観測した位置と比べて大きな誤差を持つことが多いことから，共有推定によって推定した推定ボールの位置と，自身が観測した観測ボールの位置が一致するように自己位置を較正することで，各ロボットの自己位置を修正する．本章ではこの操作を“協調自己位置較正”と呼ぶ．

Musashi では，ボールに関する絶対位置は，ロボットとボールとの相対距離とロボットの自己位置から求められる．従って，ロボットの自己位置に誤差が含まれている場合，ボールの絶対位置にも誤差が含まれることになる．本章において述べるランドマーク位置推定の目的は，ボールの位置情報に関する信頼性を向上し，信頼性のある情報を全てのチームメイト間で共有することである．

4.2　確率的情報共有による情報信頼性の向上手法

4.2.1　ランドマーク情報に関する協調推定

観測ボールの絶対距離を算出するために用いられるロボットと観測ボールとの相対距離は全方位カメラ画像から求められる．全方位カメラ画像は距離情報の信頼性が低く，角度情報の信頼性が高いことから[41]，観測ボールの絶対位置に関しても，観測ボールとロボットとの相対距離が大きくなるほど，誤差が増加するものと仮定する．そこで，サッカーフィールド上におけるボールの真の位置に関する存在確率は，ボールとロボットとの相対距離が短いほど分散が小さくピークが高い正規分布によって表現する．逆に，ボールの存在確率は相対距離が長いほど分散が大きく，ピークが低い分布とした．

本手法では，各ロボットは観測ボールの位置に関する存在確率p^i_dを，式(4.1)に示すガウス分布を用いて表した．式(4.1)において，p_xとp_yはサッカーフィールド上における任意のx，y座標を示しており，$\overline{m^i_x}$と$\overline{m^i_y}$はi番目のロボット($\boldsymbol{m}^i=[m^i_x,\ m^i_y]^T$)が観測した観測ボールの位置に関する重み付き平均値を示している（式(4.2)）．ここで式(4.2)に示す重み付き平均計算では，n[step]間におけるデータ履歴を重みw_tとして用いることで“最新の情報ほど，大きな重みをもつ”ものとした． 分散値σ^i_dは，観測ボールとロボット間の相対距離について，式(4.3)に示す重み付き平均によって求めた値を表す．ここで$\overline{\boldsymbol{d}^i}$は，式(4.4)に示すように，$i$番目のロボットとボール間の相対距離に関する重み付き平均値である．

$$p^i_d\left(p^x, p^y\right)=\prod_{k=x,y}\frac{1}{\sqrt{2\pi\sigma^i_d}}\exp\left(-\frac{\left(p^k-\overline{m^i_k}\right)^2}{2\sigma^i_d}\right) \tag{4.1}$$

$$\overline{m^i_{x,y}}=\sum_{t=1}^{n}m^i_{x,y}(t)\cdot w_t \Big/ \sum_{t=1}^{n}w_t \tag{4.2}$$

$$\sigma_d^i = \sum_{t=1}^{n}\left(\overline{d^i} - d^i(t)\cdot w_t\right)^2 \Big/ \sum_{t=1}^{n} w_t \tag{4.3}$$

$$\overline{d^i} = \sum_{t=1}^{n} d^i(t)\cdot w_t \Big/ \sum_{t=1}^{n} w_t \tag{4.4}$$

また，ロボットの方位を定める上で用いている方位センサには，局所磁場の影響によるノイズが含まれるため，ロボットの自己位置に基づいて算出された観測ボールの絶対位置には，ロボットの方位推定に含まれる角度の誤差も含まれる．そこで，方位誤差を考慮したボールの存在確率p^i_sを式(4.5)として表す．Fig. 5に示すように，i番目のロボットと観測ボールを結ぶ直線Lと，i番目のロボットとフィールド上の任意の点を結ぶ直線との成す角をθ^iとする．またr^iはi番目のロボットと任意の点とを結んだ線分の距離を示しており，l^iはi番目のロボットの位置を中心とし，半径をr^i，角度をθ^iとした円弧の長さを表している．V^iは，式(4.6)に示されるようにロボットの正面方向と，観測ボールを観測した方向との相対角度の重み付き平均$\overline{\theta^i_{rel}}$に関する$n$[step]間の重み付き分散値を示している．重み付き平均$\overline{\theta^i_{rel}}$には，式(4.7)に示されるようにデータ履歴を重みw_tとして用いた．式(4.5)における分散値σ^i_sは，方位センサに関する許容誤差θ_{th}とV^iから，式(4.8)のように求められる．p^i_sによって表されるボールの存在確率は，円弧l^iに沿って分布され，確率密度はFig. 5に示される直線L上において最大値をとる．

$$p^i_s\left(l^i\right) = \exp\left(-\frac{{l^i}^2}{2{\sigma^i_s}^2}\right) \tag{4.5}$$

$$V^i = \sum_{t=1}^{n}\left(\overline{\theta^i_{rel}} - \theta^i_{rel}(t)\cdot w_t\right)^2 \Big/ \sum_{t=1}^{n} w_t \tag{4.6}$$

$$\overline{\theta^i_{rel}} = \sum_{t=1}^{n} \theta^i_{rel}(t)\cdot w_t \Big/ \sum_{t=1}^{n} w_t \tag{4.7}$$

$$\sigma^i_s = r^i\cdot\sin\left(\theta_{th} + V^i\right) \tag{4.8}$$

提案手法では，ボールの存在確率$p^i{}_{all}$は式(4.9)に示されるように$p^i{}_d$と$p^i{}_s$から求められる．ここでMは自チームにおけるロボットの台数を示している．最後にサッカーフィールド上において最も$p^i{}_{all}$が高い地点をボール位置の真値，つまり推定ボールと仮定した．

$$p^i_{all}\left(p^x, p^y\right) = \frac{1}{2M}\sum_{i=1}^{M}\left\{p^i_d\left(p^x, p^y\right) + p^i_s\left(l^i\right)\right\} \tag{4.9}$$

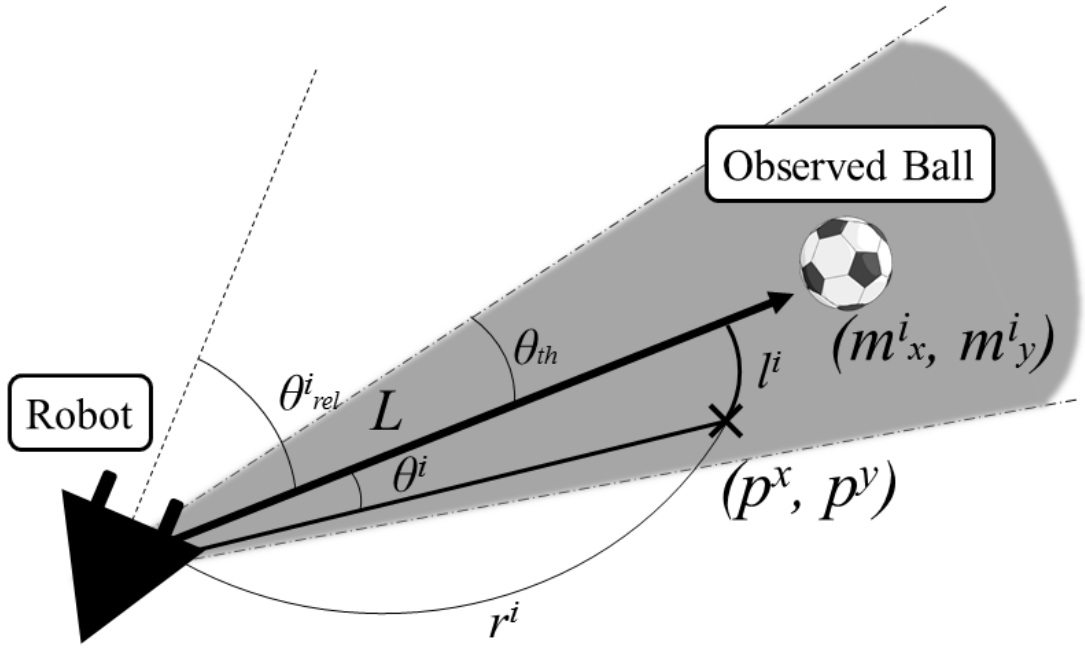


Fig. 52 Example of the probability function Ps()

4.2.2 ランドマーク情報に基づいた協調自己位置推定

ロボットが誘拐状態に陥った際，多くの場合，観測ボールは推定ボールから遠く離れた位置に観測される．従って観測ボールの位置を推定ボールの位置に重ねるようにロボットの自己位置を更新することで，誘拐状態にあるロボットの自己位置を真値（ロボットが実際に立っている位置）付近へ誘導することが出来る．提案手法では，自己位置更新のためのロボットの移動量を式(4.10)により求める．ここで$\boldsymbol{m}^i_p$とm^i_oは，それぞれ推定ボールと観測ボールの位置を示している．また式(4.11)に示されるα^i_dは，Fig. 53に示されるようにi番目のロボットにおける推定ボールの存在確率 ${}^{p}p^i_d$と観測ボールの存在確率${}^{o}p^i_d$に基づいた更新のための係数である．

$$\Delta \boldsymbol{m}^i = \alpha^i_d \cdot (\boldsymbol{m}^i_p - \boldsymbol{m}^i_o) \tag{4.10}$$

$$\alpha^i_d = \frac{{}^{o}p^i_d - {}^{p}p^i_d}{{}^{o}p^i_d} \tag{4.11}$$

ロボットの位置はボールの存在確率p^i_dに基づいた並進移動とp^i_sに基づいた回転移動によって更新される．$\boldsymbol{R(\theta)} \in \boldsymbol{R}^{3x3}$は回転行列を表しており，式(4.13)における$\beta^i_d$はFig. 54と式(4.14)に示されるように，i番目のロボットにおける推定ボールの存在確率${}^{p}p^i_s$と観測ボールの存在確率${}^{o}p^i_s$に基づいた回転運動による更新のための係数を表している．ここでθ_{pro}はロボット座標系における推定ボールの方位を示している．

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t \cdot \boldsymbol{R}(\Delta\theta^i) + \left[\Delta\boldsymbol{m}^{iT}, \Delta\theta^i \right]^T \tag{4.12}$$

$$\Delta\theta^i = \beta_d^i \cdot (\theta_{pro} - \theta_{rel}^i) \tag{4.13}$$

$$\beta_s^i = \frac{{}^o p_s^i - {}^p p_s^i}{{}^o p_s^i} \tag{4.14}$$

並進運動と回転運動による更新の例として，Fig. 53，Fig. 54 に示すような状況を考えると，まず並進運動による更新について，Fig. 53-(a)に示されるようにロボットは観測したボール位置を最大値としてボールの存在確率を分布するため，観測ボールと推定ボールの存在確率には，Fig. 53-(b)に示すような関係が得られる．同様に Fig. 54-(a)のような状況を考えると，観測ボールと推定ボールの存在確率には同様の関係が見られる．

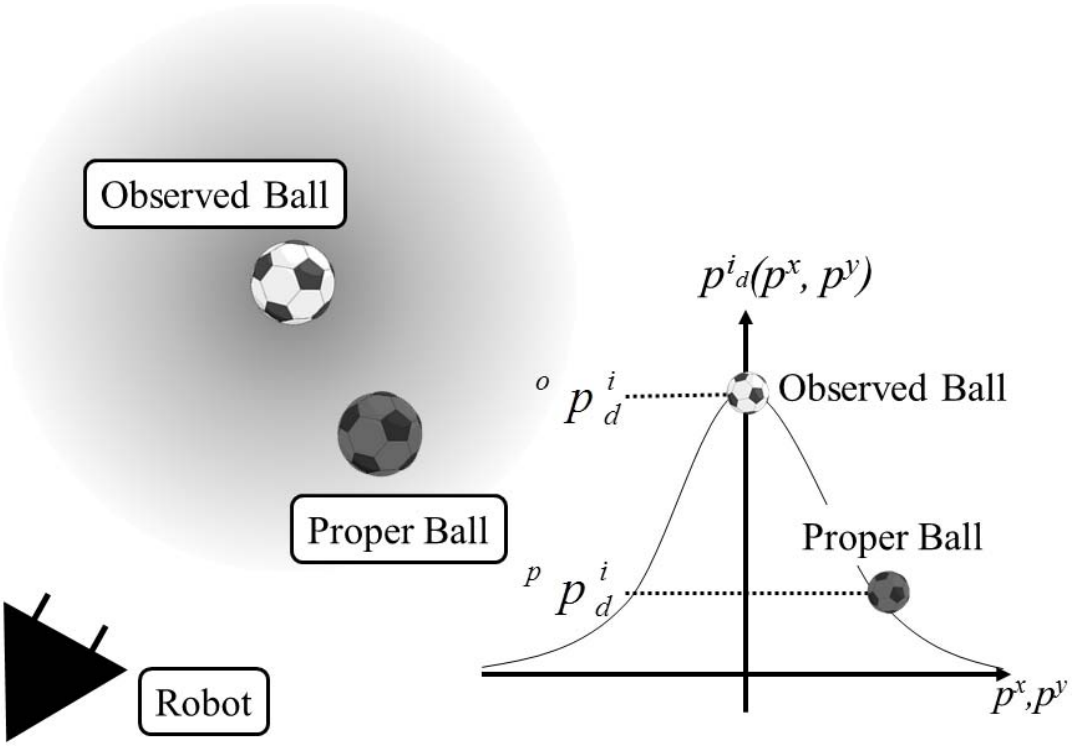


Fig. 53 Robot position calibration based on the p^i_d.

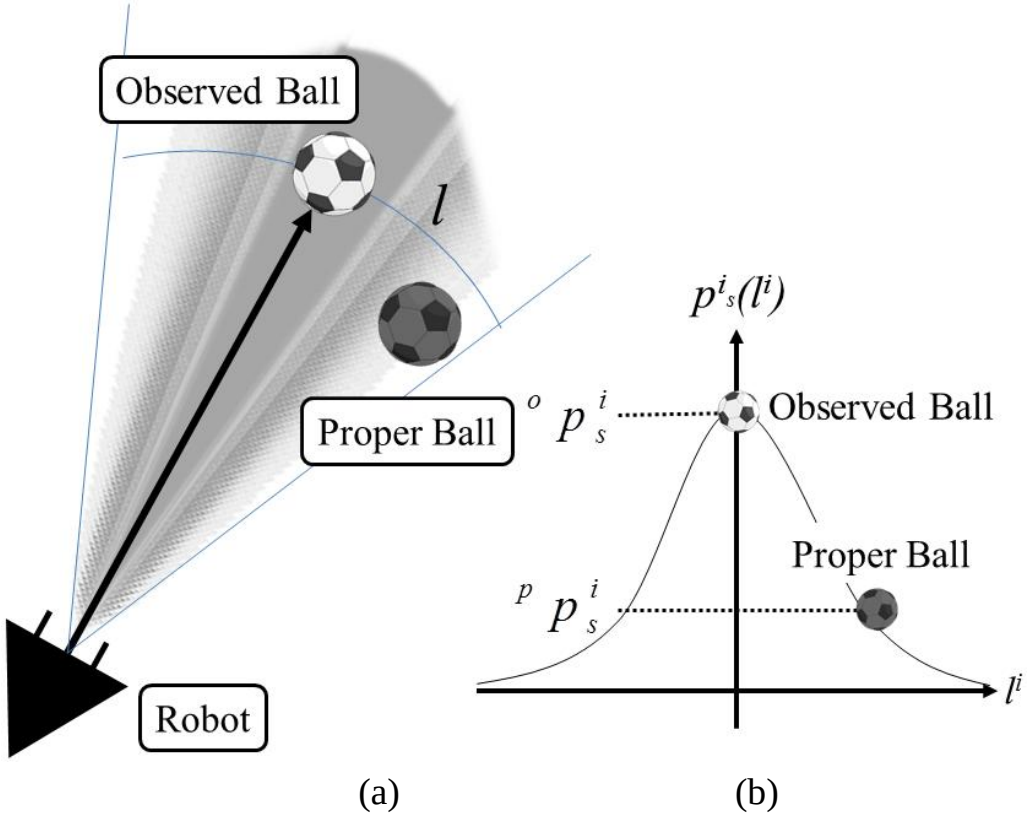


Fig. 54 Robot position calibration based on the p^i_s

4.3 実環境における情報共有実験

ロボット単体による自己位置推定とボール位置推定, 群内情報の共有によるランドマーク位置の協調推定，協調自己位置推定の精度を評価するため，Fig. 57に示す実験用サッカーフィールド(12.00x6.00[m]のハーフコート)と5台のMusashiを用いて，実環境下における本手法の評価実験を行った．実験は，Fig. 55に示すように平均照度：371.99[lx](309.40~403.60[lx])，方位センサが観測する局所磁場の平均誤差はFig. 56に示すように:-12.13[deg](-150.00~100.00[deg])の環境で行った．また，Fig. 57に示すように各ロボットの位置は，Robot 1(No.1)：(9.00, 0.00)，Robot 2(No.2)：(6.00, 3.00)，Robot 3(No.3)：(3.00, 3.00)，Robot 4(No.4)：(3.00, -3.00)，Robot 5(No.5)：(6.00, -3.00)とし，全てのロボットの方位はフィールド座標系に対し180.00[deg]とした．

364.5	367.4	362.5	356.4	349.5	338.3	339.3	347.2	349.7	337.2	340	309.4
378.8	383.7	382.4	377.7	371.7	368.9	359.8	364.3	361.2	362.2	351.3	329.3
386.7	389.3	390.8	389.1	385.9	382.3	370.7	377.4	369.5	369.8	358.8	336.1
386.1	394.1	395.2	395.8	394.8	393.8	380.7	385.8	375.8	374.7	360.8	339.3
393.4	396.9	397.9	398.9	398.9	397.8	385.9	390.1	379.2	378.1	365.1	339.3
394	397.7	398.9	400.5	401.6	403.6	389.2	392.7	379.4	377.2	365.3	339.9
385.9	390.3	393.2	395	397	399.3	386.4	392.3	380.2	377.6	364.9	337.8
376.5	380.2	382.3	381.4	383.3	383.7	374.5	381.7	371.8	370.7	358.9	332.3
346.7	350.6	353.9	354.9	356	357.4	351.2	360.1	354.7	354.7	341.7	318.2

Fig. 55 Lighting Condition of Experimental Environment

***units for all numbers are [lux].**

−35	10	−5	−30	−25	−30	0	20	−45	−25	0	−45
−20	5	−5	−10	−15	−10	5	20	10	10	−25	−35
−85	0	0	−25	−50	−55	75	40	20	0	−15	−75
−20	15	−60	−10	30	20	15	−110	0	55	40	65
−50	−45	−10	0	75	10	5	−5	−15	45	10	−40
−40	−105	95	65	5	−15	−10	−95	50	15	100	15
−115	−90	−150	0	−15	20	10	0	−70	−5	−105	−105
−55	−45	−80	−40	30	−5	25	−10	0	20	25	−30
−60	−50	−30	−30	−20	−5	15	0	20	0	−10	0

Fig. 56 Magnetic Condition of Experimental Environment

***units for all numbers are [deg].**

4.3.1 静的状態における位置推定精度評価

各ロボットとランドマークが移動しない状態(以降，静的状態[62]と記す)において，ランドマークであるボールを，A：(0.00, 0.00)，B：(4.50, 0.00)，C：(8.50, 0.00)の3ヵ所に配置した際の，各ロボットの自己位置推定とボール位置推定の測定を行った．評価は900[data](=N)を対象とした．評価関数には，式(4.15)~(4.17)に示す平均誤差e^s_{bp}，e^s_{bd}，e^s_{ba}，標準偏差(*SD*)を用いた．ここでm^{tr}_x，m^{tr}_yは，A~C地点に示すボール位置の真値を示しており，$d^{i\text{-}tr}$はA~C地点における各ロボットとボールとの相対距離の実測値，θ^{tr}_{rel}はA~C地点における各ロボットの正面方向とボール方向との相対角度の実測値を示す．また，各ロボットの自己位置については，式(4.18)，(4.19)に示すe^s_{rp}，e^s_{ra}により評価した．e^s_{rp}，e^s_{ra}はロボットの絶対位置と角度に関する平均誤差を表している．r^{tr}_x，r^{tr}_y，θ^{tr}はロボットの真の位置を示しており，${}^n r^i_x$，${}^n r^i_y$，${}^{\mathrm{n}}_{\mathrm{a}}\theta^i$は*i*番目のロボットが推定した自己位置を示している．各ロボットが観測したボール位置と推定ボールの位置の絶対誤差の平均をTable 4, 各ロボットが観測したボールとの相対距離･相対角度の平均誤差結果をTable 5, Table 6, 協調自己位置推定を行った後の各ロボットの自己位置についての平均誤差をTable 7, Table 8に示す.

$$e_{bp}^{s} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(m_x^{tr} - {}^{n}m_x^{i}\right)^2 + \left(m_y^{tr} - {}^{n}m_y^{i}\right)^2} \tag{4.15}$$

$$e_{bd}^{s} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(d^{itr} - {}^{n}d^{i}\right)^2} \tag{4.16}$$

$$e_{ba}^{s} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(\theta_{rel}^{tr} - {}^{n}\theta_{rel}^{i}\right)^2} \tag{4.17}$$

$$e_{rp}^{s} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(r_x^{tr} - {}^{n}r_x^{i}\right)^2 + \left(r_y^{tr} - {}^{n}r_y^{i}\right)^2} \tag{4.18}$$

$$e_{ra}^{s} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(\theta^{tr} - {}_{a}^{n}\theta^{i}\right)^2} \tag{4.19}$$

Table 4~Table 8において”No.1”から”No5”は，Fig. 57に示す各ロボットの番号に対応しており，”Proposal method”は協調推定の結果を示している．また，Musashiはボールとの相対距離が約4.50[m]を超えると，全方位画像の解像度低下によりボール位置に関する誤差が増大，もしくはボールが認識出来なくなるため，各観測においてボールを認識できなかったロボットの結果は，”-”と示した．Table 5，Table 6では，ロボットが単体で推定した自己位置に対する，ロボットが観測したボールとの相対距離・角度(A_{obs}，B_{obs}，C_{obs})と，協調推定した推定ボールとの相対距離・角度(A_{est}，B_{est}，C_{est})を示した．

Table 4が示すボール位置に関する平均誤差の結果から，A地点では協調推定により推定したボール位置の誤差は0.40[m]となった．各ロボットのボール位置に関する誤差は，No.3：1.36[m]，No.4：0.45[m]，No.5：1.03[m]となり，平均誤差は0.95[m]であった．B地点では，協調推定によって平均誤差0.32[m]の位置にボールを推定した. B地点において各ロボットが観測したボールの位置に関する平均誤差は0.50[m]であった．C地点では，協調推定による平均誤差は0.19[m]となり，C地点における各ロボットの観測結果は，平均1.10[m]であった．Table 4に示す結果より，5台のロボットが観測したボール位置の平均誤差よりも, 推定ボールの位置の平均誤差が小さいことから，情報を共有することにより，ボールの位置精度が向上する傾向にある．

また，Table 5に示すロボットとボールとの相対距離誤差の結果では，全方位画像の特徴として前述した“相対距離に応じて誤差が増加する傾向”は見られず, 相対距離に関わらず0.08~1.95[m]までの誤差が見られた．また表3よりロボットの正面方向とボールを観測した方向との相対角度は，0.24~11.91[deg]の誤差を含んでいたとわかる．Musashiにおいて相対距離・角度に関する情報は，①地面が局所的に水平ではない場合，②ロボット本体に対して全方位カメラが水平に取り付けられていない場合などの状況において，誤差が生じると考えられ，試合環境ではサッカーフィールドの状態やロボット同士の衝突によって上記の誤差が発生する可能性がある. Table 4に示す

協調推定結果では，共有した情報に，Table 5とTable 6に示した相対距離と角度に関する誤差が含まれる場合であってもボールの位置推定が可能である．また，Table 6に示す結果から，各位置においてロボットが観測したボールとの相対角度の平均誤差は， A地点：5.51[deg]， B地点：13.68[deg]，C地点：7.03[deg]となった．協調推定の結果では，ロボットの正面方向と推定ボール方向との相対角度の平均誤差は，A地点：9.13[deg]，B地点：16.46[deg]，C地点：18.36[deg]となった．

Musashiの自己位置推定では，パーティクルの初期化とリサンプリング時に各パーティクルに対してランダム値を与えることから，Table 7，Table 8が示すように，ロボット単体によって自己位置を推定する際，周辺環境が変化しない静的状態であっても自己位置を推定する度に，推定精度に差異が生じた． また，白線情報を取得するための色抽出過程においても，画像に対する色抽出の閾値を操作者が主観に基づいて設定するため，ロボットによって自己位置推定精度に差が生じた．本実験では，A地点におけるNo.2とNo.5，及びC地点におけるNo.2において，各パーティクルの尤度が高まらず，MCLによる更新が行われなかった．各測定における全ロボット間の位置に関する平均誤差は，A地点：0.61[m]，B地点：0.26[m]，C地点：0.37[m]となった．方位に関する平均誤差は， A地点：12.81[deg]，B地点：6.51[deg]，C地点：6.10[deg]となった．また標準偏差の平均として，0.04[m]と2.39[deg]を示した．

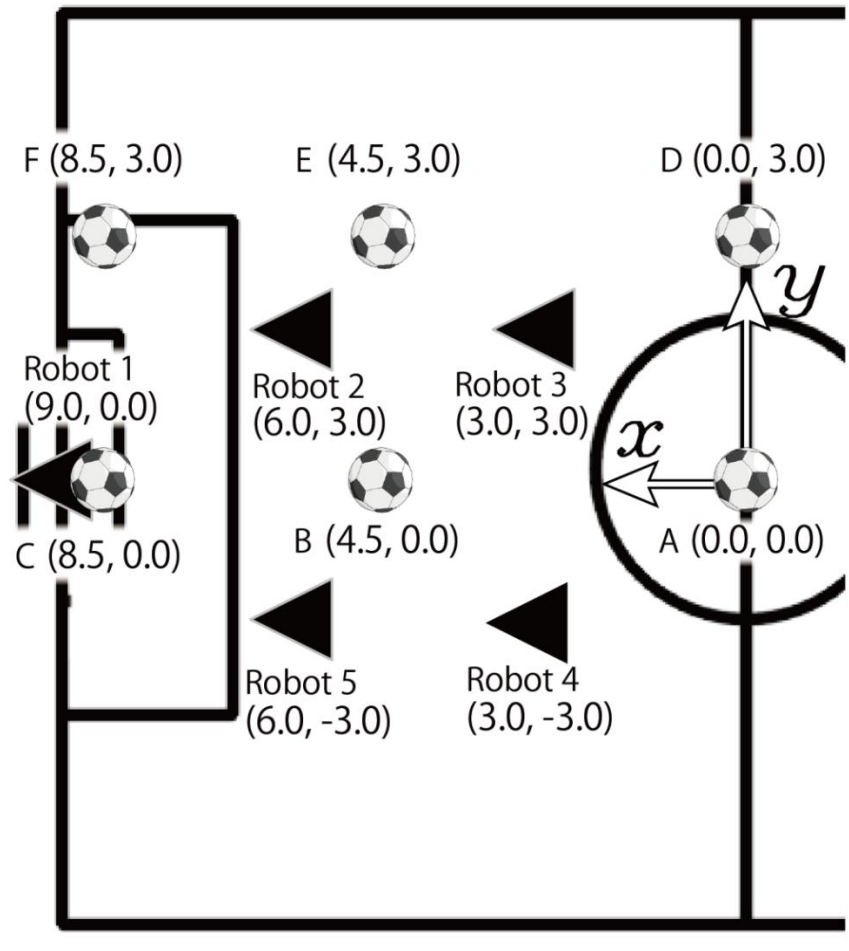
この標準偏差の原因は，MCLにおけるリサンプリング過程によるものである．また，A地点におけるNo.2とNo.5，及びC地点におけるNo.2では，各パーティクルの尤度が閾値を超えなかったため，デッドレコニングによる自己位置推定が行われた．実験は静的状態で行ったため，標準偏差は0.00[m]を示した．各測定におけるロボット単体による自己位置の誤差はサッカーフィールドの寸法(12.00x18.00[m])に対し，A地点 : 横5.08[%] / 縦3.38[%]，B地点 : 横2.16[%] / 縦1.44[%]，C地点：横3.08[%] / 縦2.05[%]であり，十分な精度が得られていると考えられる．また，ボールの直径が0.22[m]であることからもボールの絶対位置を，十分な精度で推定できる．一方で方位に関する誤差は，対象物との相対距離に応じて影響が大きくなるため，障害物に対する回避行動やボール取得行動を行う上で問題となる．

各測定地点における協調自己位置推定では，位置に関する平均誤差がA地点 : 0.61[m]，B地点 : 0.26[m]，C地点 : 0.33[m]となり，誤差の軽減に至らなかった．また，方位に関する平均誤差は，A地点 : 11.58[deg]，B地点 : 5.39[deg]，C地点10.26[deg]となり，A，B地点では誤差が軽減したが，C地点では誤差が増加した．

これら原因としてA~C地点における位置の推定では， No.2~No.5とボールとの相対距離が3.35~6.70[m]の範囲であったことがあげられる．本手法では，Fig. 53と式(4.11)に示すように，全方位画像の特性を考慮し，ボールとの相対距離が遠くなるほど，並進移動のための係数α^{i}_{d}が小さくなるため，本手法による位置推定の更新が十分に行われなかった．C地点におけるNo.1は，相対距離：0.37[m]の位置にボールを観測していたため，並進移動による更新が行われ，誤差が

0.17[m]軽減した．回転移動による更新では，C地点のNo.1において29.53[deg]の誤差が生じた．この結果からも，協調推定したボールの位置がロボット近辺であった場合(C地点におけるNo.1では，推定ボールとの相対距離が0.36[m]であった)，回転移動によって状態を更新したことで，誤差が増大することがわかる．

そこで本研究では，ロボットと観測したボールとの相対距離に応じて，並進移動による更新と回転移動による更新の割合を線形に変化させた．割合の変化は，相対距離が近いほど並進移動による更新割合が高く，相対距離が遠いほど回転移動による更新割合が高くなるものとした．Table 9，Table 10に更新割合を適応した際の結果について示す．Table 9に示す位置誤差の結果では，A地点：0.60[m]，B地点：0.26[m]，C地点：0.34[m]となり，A地点とC地点において誤差の軽減が得られた．表7に示す方位誤差の結果は，A地点：11.84[deg]，B地点：5.69[deg]，C地点4.58[deg]となり，全測定点で誤差の軽減が得られた．



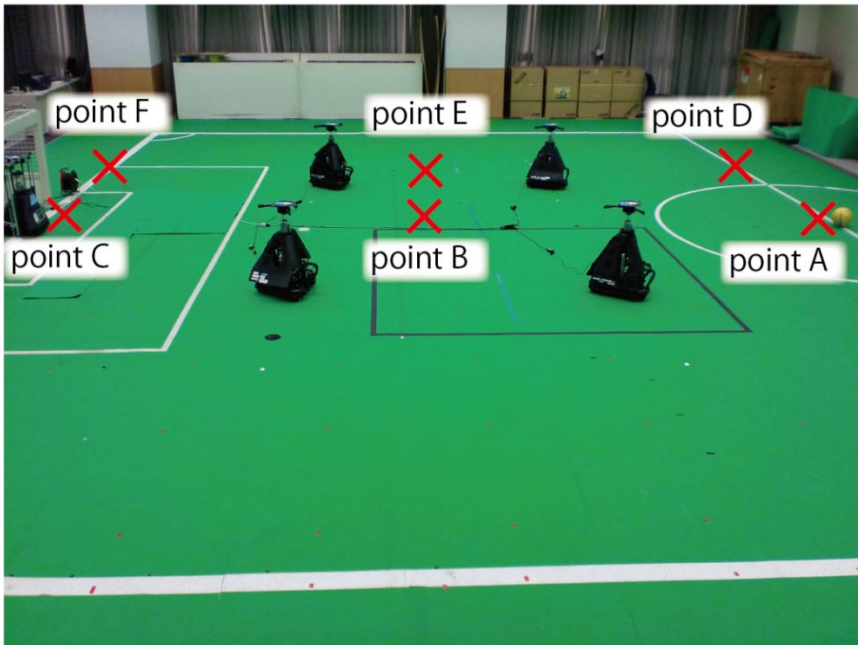


Fig. 57 Overview of Static Experimental Condition

Table 4 Result of Landmark Estimation in Static Environment (Absolute Position)

		No.1[m]	No.2[m]	No.3[m]	No.4[m]	No.5[m]	Proposed method [m]
A	e^s_{bp}	-	-	1.36	0.45	1.03	0.40
	SD	-	-	0.23	0.04	0.02	0.17
B	e^s_{bp}	0.50	1.05	0.25	0.27	0.46	0.32
	SD	0.25	0.09	0.07	0.11	0.08	0.08
C	e^s_{bp}	0.28	2.37	-	1.13	0.64	0.19
	SD	0.10	0.03	-	0.01	0.06	0.04

Table 5 Absolute Relative Distance Error of Observed Ball and Proper Ball

		No.1 [m]	No.2 [m]	No.3 [m]	No.4 [m]	No.5 [m]
A_{obs}	e^s_{bd}	-	-	0.59	-0.08	0.38
	SD	-	-	0.19	0.03	0.00
A_{est}	e^s_{bd}	0.02	-0.51	-0.69	-0.19	0.34
	SD	0.06	0.08	0.14	0.07	0.03
B_{obs}	e^s_{bd}	0.35	0.20	-0.60	-0.76	-0.58
	SD	0.06	0.05	0.02	0.01	0.02
B_{est}	e^s_{bd}	0.23	-0.61	-0.70	-0.77	-0.84
	SD	0.08	0.12	0.09	0.06	0.10
C_{obs}	e^s_{bd}	-0.13	1.95	-	-0.51	0.46
	SD	0.01	0.03	-	0.04	0.04
C_{est}	e^s_{bd}	-0.06	-0.31	-1.91	-0.23	0.21
	SD	0.05	0.05	2.58	0.03	0.06

Table 6 Absolute Relative Angle Error of Observed Ball and Proper Ball

		No.1 [deg]	No.2 [deg]	No.3 [deg]	No.4 [deg]	No.5 [deg]
A_{obs}	e^s_{ba}	-	-	0.55	8.06	7.91
	SD	-	-	0.64	0.04	0.13
A_{est}	e^s_{ba}	1.81	37.37	2.97	0.77	2.71
	SD	2.04	1.73	3.38	2.72	1.62
B_{obs}	e^s_{ba}	0.24	7.84	11.91	9.61	8.09
	SD	0.28	0.08	0.24	0.07	0.06
B_{est}	e^s_{ba}	2.88	8.59	11.33	8.82	17.27
	SD	3.74	1.89	2.06	2.16	2.36
C_{obs}	e^s_{ba}	8.08	8.19	-	1.12	10.76
	SD	0.68	0.06	-	0.07	0.49
C_{est}	e^s_{ba}	58.65	9.35	8.22	11.23	4.36
	SD	14.63	0.88	4.66	0.54	1.44

Table 7 Absolute Robot Position Error of Single Robot Localization and Cooperative Self-Localization

		Single Localization		Multi Localization	
		e_{rp} [m]	*SD* [m]	e_{rp} [m]	*SD* [m]
A	No.1	0.28	0.00	0.28	0.00
	No.2	1.37	0.00	1.37	0.00
	No.3	0.11	0.05	0.13	0.05
	No.4	0.18	0.03	0.17	0.03
	No.5	1.10	0.00	1.10	0.00
B	No.1	0.27	0.01	0.26	0.00
	No.2	0.15	0.05	0.14	0.04
	No.3	0.30	0.08	0.33	0.08
	No.4	0.28	0.07	0.28	0.07
	No.5	0.26	0.03	0.29	0.03
C	No.1	0.28	0.04	0.11	0.07
	No.2	0.39	0.00	0.31	0.01
	No.3	0.34	0.26	0.34	0.26
	No.4	0.59	0.00	0.59	0.01
	No.5	0.27	0.03	0.29	0.03

Table 8 Absolute Robot Angle Error of Single Robot Localization and Cooperative Self-Localization

		Single Localization		Multi Localization	
		e_{ra} [deg]	*SD* [deg]	e_{ra} [deg]	*SD* [deg]
A	No.1	5.87	1.62	5.87	1.62
	No.2	13.17	0.00	13.17	0.00
	No.3	21.47	3.03	18.60	2.29
	No.4	7.60	0.76	3.17	1.85
	No.5	15.96	0.00	17.10	1.13
B	No.1	5.94	3.45	3.55	1.04
	No.2	8.88	1.22	9.39	1.31
	No.3	7.69	1.30	8.05	1.41
	No.4	4.06	2.00	4.49	1.59
	No.5	5.99	1.04	1.45	1.12
C	No.1	8.65	14.43	38.18	5.15
	No.2	4.96	0.00	5.44	0.53
	No.3	2.85	5.98	2.85	5.98
	No.4	8.32	0.00	3.21	1.97
	No.5	5.72	1.04	1.64	1.22

Table 9 Absolute Robot Position Error of Single Robot Localization and Cooperative Self-Localization after applying discount rate

		Single Localization		Multi Localization	
		e_{rp} [m]	*SD* [m]	e_{rp} [m]	*SD* [m]
A	No.1	0.28	0.00	0.28	0.00
	No.2	1.37	0.00	1.37	0.00
	No.3	0.11	0.05	0.10	0.04
	No.4	0.18	0.03	0.16	0.03
	No.5	1.10	0.00	1.10	0.00
B	No.1	0.27	0.01	0.26	0.00
	No.2	0.15	0.05	0.14	0.05
	No.3	0.30	0.08	0.33	0.08
	No.4	0.28	0.07	0.28	0.07
	No.5	0.26	0.03	0.28	0.03
C	No.1	0.28	0.04	0.10	0.06
	No.2	0.39	0.00	0.38	0.00
	No.3	0.34	0.26	0.34	0.26
	No.4	0.59	0.00	0.58	0.01
	No.5	0.27	0.03	0.28	0.03

Table 10 Absolute Robot Position Error of Single Robot Localization and Cooperative Self-Localization after applying discount rate

		Single Localization		Multi Localization	
		e_{ra} [deg]	*SD* [deg]	e_{ra} [deg]	*SD* [deg]
A	No.1	5.87	1.62	5.87	1.62
	No.2	13.17	0.00	13.17	0.00
	No.3	21.47	3.03	18.98	2.13
	No.4	7.60	0.76	4.09	1.69
	No.5	15.96	0.00	17.10	1.13
B	No.1	5.94	3.45	3.85	1.07
	No.2	8.88	1.22	9.19	1.09
	No.3	7.69	1.30	7.86	1.16
	No.4	4.06	2.00	4.22	1.74
	No.5	5.99	1.04	3.32	1.05
C	No.1	8.65	14.43	8.66	14.43
	No.2	4.96	0.00	5.42	0.51
	No.3	2.85	5.98	2.85	5.98
	No.4	8.32	0.00	3.21	1.97
	No.5	5.72	1.04	2.76	1.21

4.3.2　動的状態における位置推定精度評価

動的環境(ランドマークであるボールが移動する状態)において本手法を評価するため，Fig. 58に示す環境において協調推定と協調自己位置推定に関する評価実験を行った．実験条件として，静的実験と同じ位置に5台のロボットを配置するものとし，ボールはA地点からC地点方向へ向けて初速3.0[m/s]で転がすものとした．また，ロボット間における通信の時間遅れを考慮し，実験にはNo.1が観測したボール情報とNo.1が他のロボットと共有した群内情報を用いた．

評価は，No.1の観測情報と群内情報に対して協調推定を行った際の推定ボールの位置と協調自己位置較後の自己位置を対象として行った．観測は約2.0[sec]間，60[data](=N)とした．評価関数として式(4.18)，(4.19)を用い，推定ボールの位置推定精度に関しては，式(4.20)に示すy軸方向の誤差e_{bd}のみを評価対象とした．

$$e_{bd} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\left(m_y^{tr} - {}^{n}m_y^{i}\right)^2} \tag{4.20}$$

Fig. 58に推定ボールの平均誤差e_{bd}と標準偏差の結果を示す．また，Fig. 59に各ロボットが観測したボールの移動軌跡と協調推定により推定した推定ボールの移動軌跡を示す．Table 12には，協調自己位置推定後の自己位置結果を示した．Table 11の結果から，No.1は1.26[m]の平均誤差を示し，標準偏差も0.88[m]と大きかった．Fig. 59が示すボールの移動軌跡が示すように，No.1とボールとの相対距離が短いほど，No.1が観測したボール位置の誤差と標準偏差は減少した．協調推定の結果は平均誤差0.35[m]，標準偏差0.23[m]であり，No.1の観測結果に対し誤差の軽減が得られた．動的環境下における協調推定評価実験から，No.1は群内情報に基づいた協調推定によってボールの位置を推定することで，自身が持つ情報より正確なボール位置を認識出来る．

協調自己位置推定の結果では，No.1の自己位置は0.07[m]，9.72[deg]較正された．協調自己位置推定されたロボットの方位では，標準偏差が13.85[deg]となっている．これは，Fig. 59に示すようにNo.1が単体で行った自己位置について，位置に関する標準偏差0.07[m]と方位に関する標準偏差9.63[deg]が含まれており，さらに協調推定により推定したボール位置が0.23[m]の位置に関する標準偏差を含んでいるため，観測ボール位置と推定ボール位置の標準偏差が，No.1の自己位置に影響を及ぼしたためと考えられる．

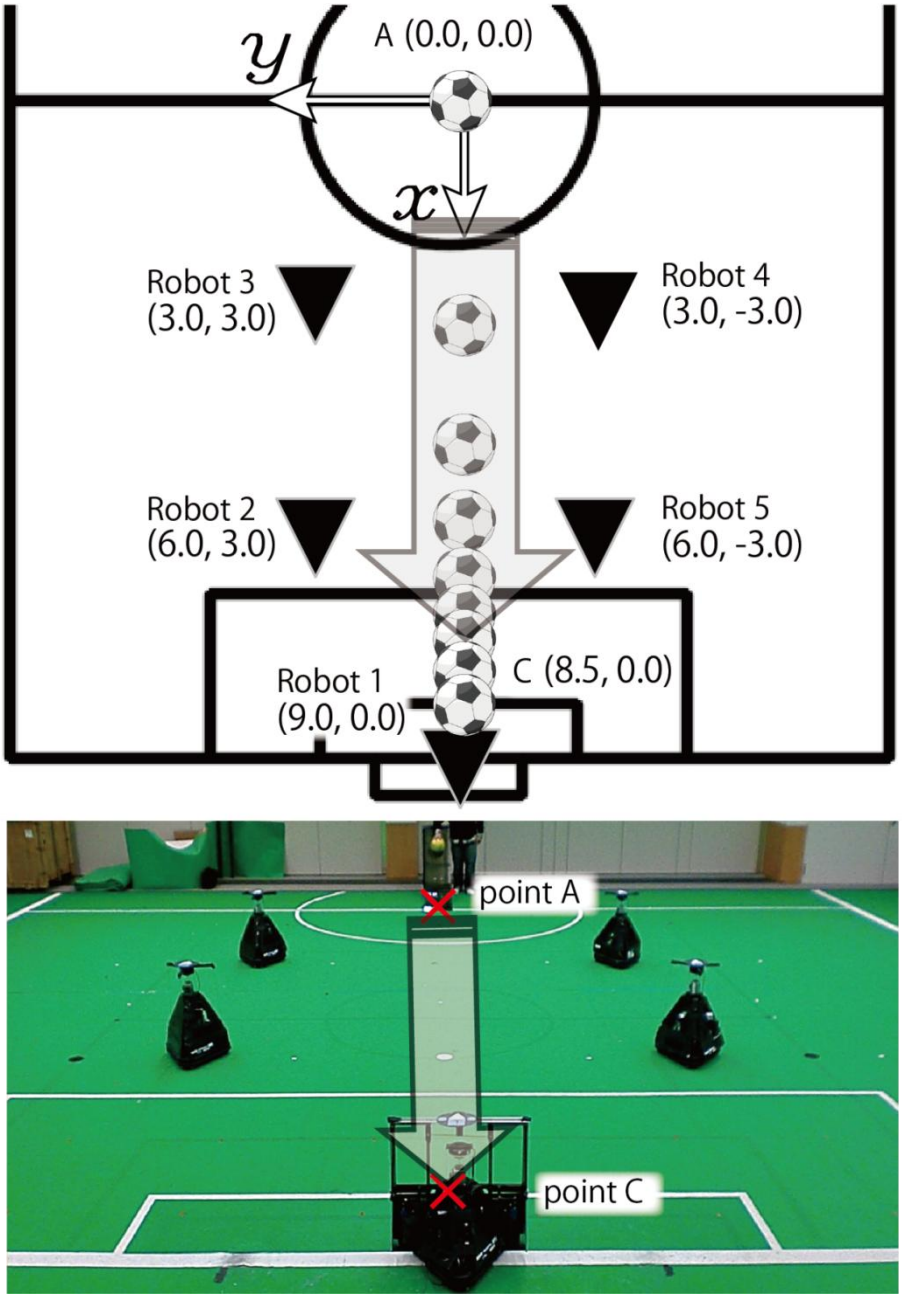


Fig. 58 Overview of Dynamic Experimental Condition

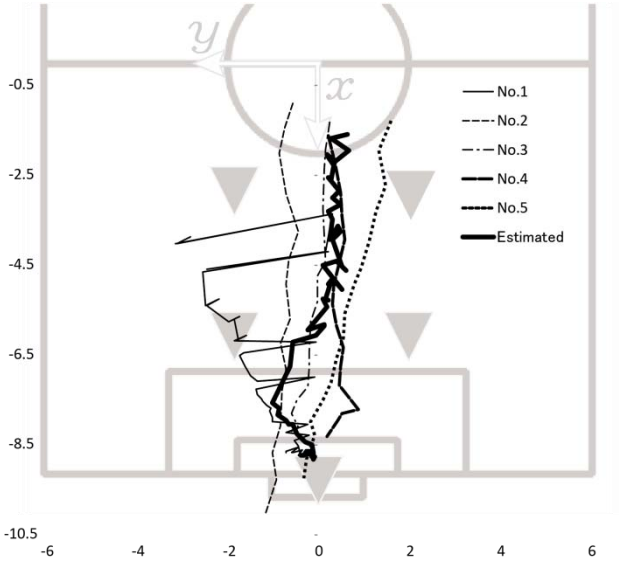


Fig. 59 Trajectory of observed and estimated ball positions

Table 11 Result of Landmark Estimation

	No.1 [m]	No.2 [m]	No.3 [m]	No.4 [m]	No.5 [m]	Proposal Method [m]
e_{bd}	1.26	0.75	0.23	0.49	0.68	0.35
SD	0.88	0.17	0.17	0.15	0.49	0.23

Table 12 Result of Cooperative Self-Localization

	Single Localization		Multi Localization	
	e_{rp}	*SD*	e_{rp}	*SD*
Position[m]	0.37	0.07	0.30	0.09
Angle[deg]	25.26	9.63	15.54	13.85

4.3.3　誘拐状態における位置推定精度評価

誘拐状態における本手法の有効性を検証するため，Fig. 60に示す実験環境によって評価実験を行った．ランドマークであるボールは，B地点：(4.50，0.00)に配置し，各ロボットの位置は，Robot 1(No.1)：(9.00，0.00)，Robot 2(No.2)：(6.00，3.00)，Robot 3(No.3)：(3.00，0.00)，Robot 4 (No.4)：(6.00，-3.00)とし，全てのロボットの方位はフィールド座標系に対し180.00[deg]とした．Fig. 61に誘拐状態における協調推定の結果を示し，Fig. 62，Fig. 63に協調自己位置推定の結果を示す．Fig. 61の結果から，各ロボットが観測したボール位置には1.83~6.37[m]の誤差が生じた．Fig. 61において，破線は4台のロボットの観測誤差の平均：3.93[m]を示している．協調推定によって推定ボールは平均誤差：1.17[m]の位置に推定された．協調自己位置推定の結果では，Fig. 62が示すとおり位置に関する平均誤差がNo.1は0.79[m]の増加，No.2：0.05[m]の減少，No.3：0.34[m]の減少，No.4：0.94[m]の減少を示した．方位に関する平均誤差は，Fig. 63よりNo.1：54.00[deg]の増加，No2：変化なし，No.3：1.37[deg]の増加，No.4：53.25[deg]の減少を示した．No.1において本手法により誤差が増加した原因として，ボール位置に関する標準偏差が大きかったことが上げられる．Fig. 60に示す実験においてNo.1とボールとの相対距離は4.5[m]であり，実験中，ボール位置は1.50~7.71[m]の範囲で変化した．協調自己位置推定では，ボールの位置を基準として自己位置を推定するため，観測したボールの位置が大きな分散を持つ場合，推定した自己位置にも分散の影響が現れた．

Fig. 60に示す実験において，No.2は位置と方位に1.79[m]と58.06[deg]の平均誤差があり，観測したボール位置には6.37[m]の平均誤差が生じた．No.2のように，誤った位置から真の位置へ復帰しない状態が継続される場合を，その状態をMCLのパーティクルが真の位置付近に存在しない“誘拐状態”と定義出来る．Fig. 60のNo.2のように，ロボットが誘拐状態に陥った場合，多くの場合においてロボットが観測したボールの位置と推定ボールの位置には大きな誤差が生じる．本手法によりNo.2の位置と方位は，0.85[m]，4.82[deg]の平均誤差まで改善された．

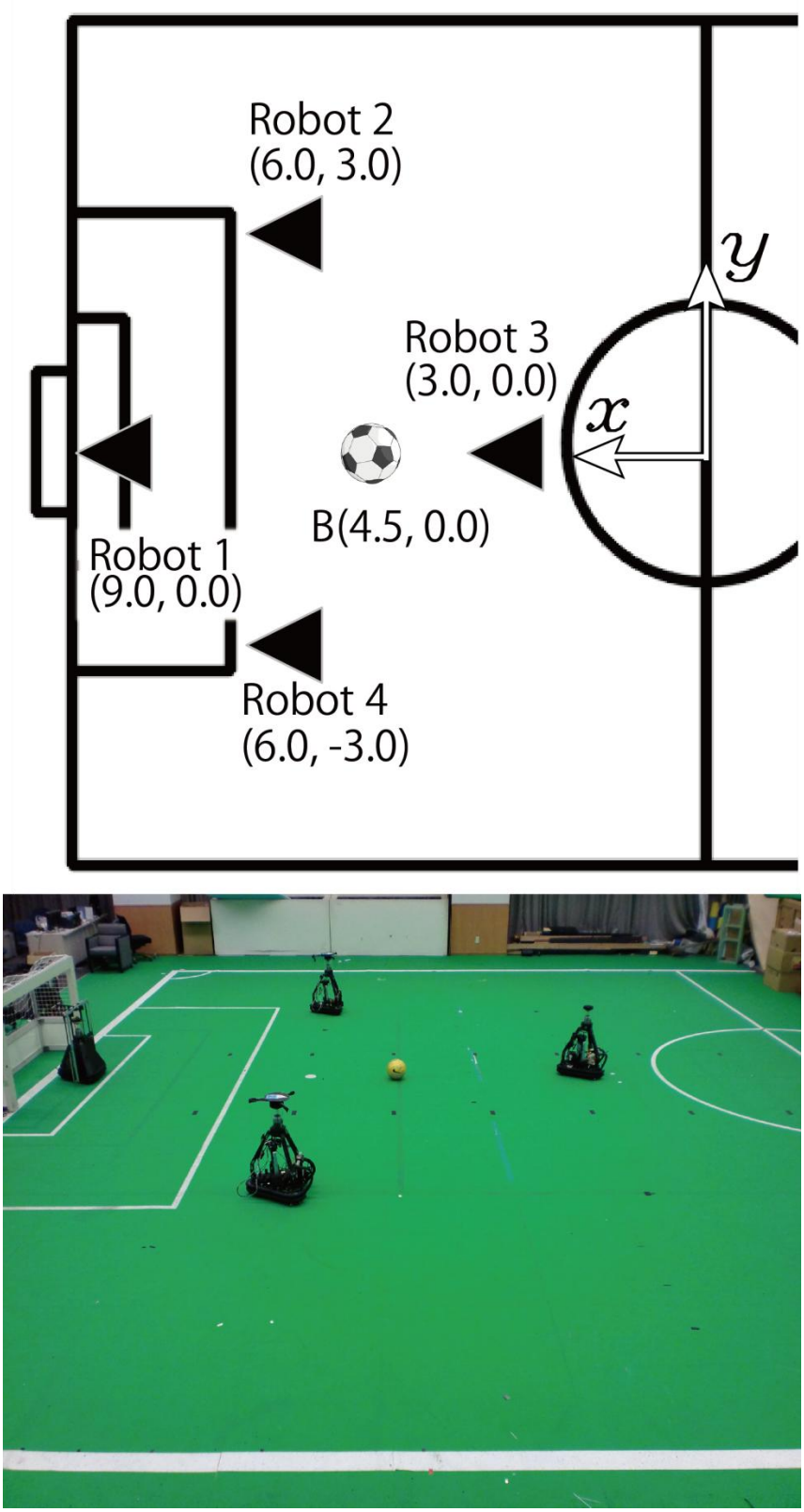


Fig. 60 Overview of Kidnapping Experimental Condition

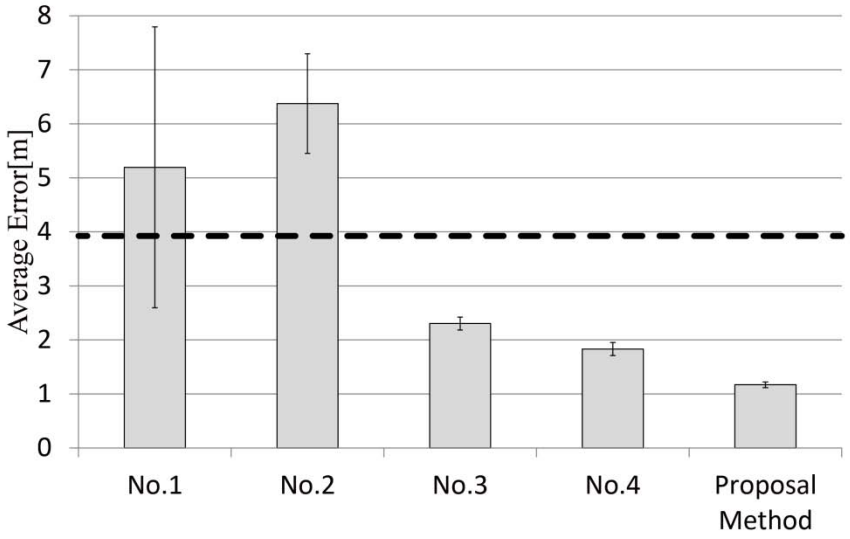


Fig. 61 Result of Landmark Estimation at Kidnapping Condition

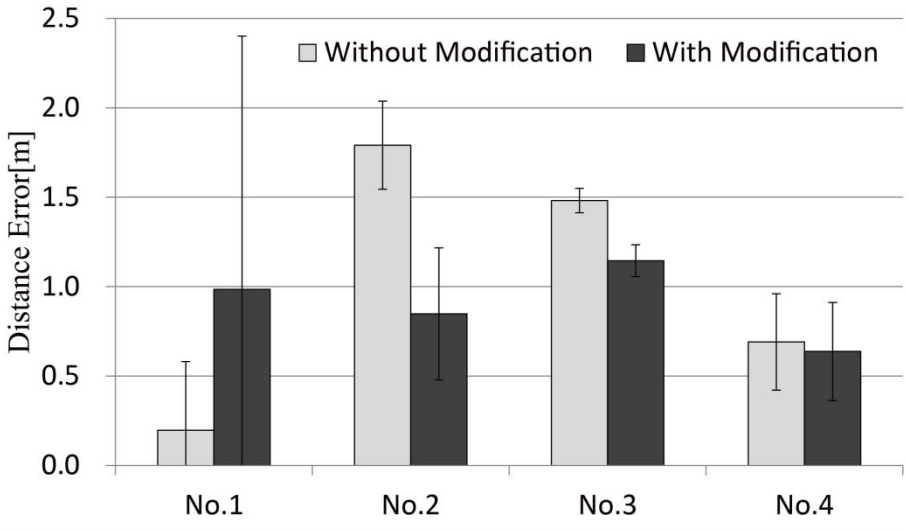


Fig. 62 Result of Cooperative Self-localization at Kidnapping Condition

(Absolute Robot Position Error)

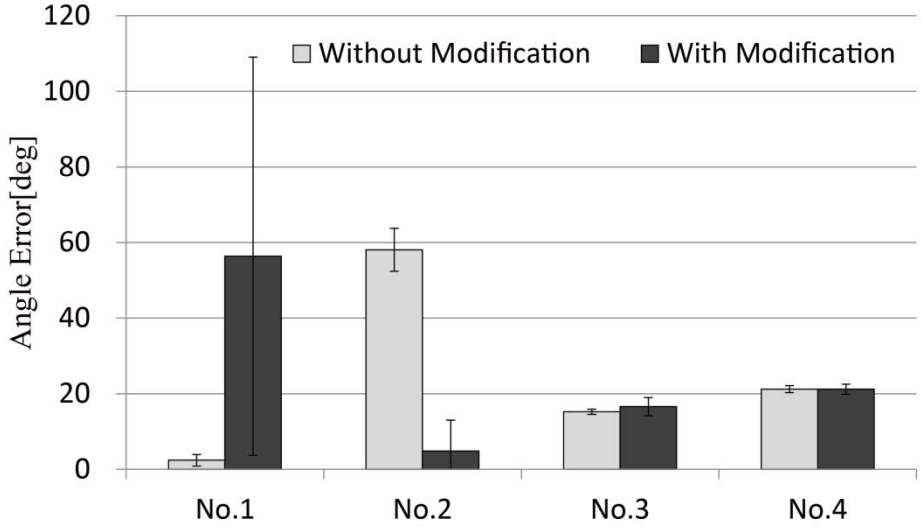


Fig. 63 Result of Cooperative Self-localization at Kidnapping Condition (Absolute Robot Angle Error)

4.4 考察

本章では筆者らが開発したサッカーロボットMusashiを用いて「複数のロボットによるランドマークの協調推定法と協調自己位置推定手法」について，実環境における推定精度の評価検証を行った．MCLは不確実性の高い環境における自己位置推定法として盛んに用いられており，自己位置推定のための十分なセンサ情報が得られる環境下では有用である．しかし，センサ情報が不十分な環境においては，局所解に陥りやすく，回復も困難である．

本章ではマルチエージェント環境を対象とし，各ロボットが自己位置や観測値に誤差を含む場合であっても，複数のロボットが持つ情報を確率的に共有することで，情報の信頼性向上が可能であることを示した．共有推定によるボール位置の推定では，全測定点において各ロボットの観測値の平均誤差よりも低い誤差でボールを推定出来ており，今後，複数のロボット間におけるパス行動や，相手ロボットの位置推定等においても，推定精度の向上が期待出来る．

また協調自己位置推定では，ロボットの方位誤差の軽減を達成した．全方位カメラによって撮影された白線情報と方位センサに基づく自己位置推定では，白線との相対距離情報に含まれる誤差や局所磁場等の影響によりロボットの方位に誤差が含まれやすい．本手法による方位修正は，上記のシステムにおいて方位情報の信頼性を向上する上で有効である．

Fig. 58に示す動的実験結果では，観測したボール位置に誤差が含まれている場合において，協調推定によりボール位置の誤差が軽減できることを示した．また，位置と方位についても精度向上が見られた．しかし，協調推定結果が連続的な軌道を描かず，協調自己位置推定を行う上でロボットの自己位置に関する標準偏差を増加させる傾向がある．また，動的状態においては他のロボットから取得した群内情報に時間遅れが含まれるため，他のロボットと共有した情報の時間遅れを考慮する必要があることがわかった．

Fig. 60に示す誘拐状態における実験では，誘拐状態に陥ったロボットの自己位置に誤差軽減が見られた．この結果から，ロボットが誘拐状態にあり，ロボットが観測したボールと推定ボールの位置に大きな差が生じている場合，協調自己位置推定によってロボットの位置を真の位置付近へ誘導可能であることを示した．本手法により，ボールとの相対距離が遠いためボールが認識出来ないロボットやボール位置に誤差を含むロボットに対しても，信頼性の高いボール位置情報を提供することが可能であり，複数のロボットによるパスなどの協調行動への活用に期待出来る．

第五章

パス行動における

確率的行動選択

第五章　パス行動における確率的行動選択

5.1　はじめに

MSL では，プロジェクトが発足した 1996 年からロボット同士によるパス行動などの協調行動に関する研究が盛んに行われてきており[16]，Table 13 に示すように 2009 年からは試合中における一部の行動においてロボット同士によるパス行動が義務付けられた．パス行動は，最低 2 台のチームメイトロボット間において，一方のロボットがもう一方のチームメイトロボットへボールを蹴り出し，相手チームのロボットにボールを奪われることなく，チームメイトロボットがボールを受け取ることで成立する．ここで，人間のサッカーにおける呼称に基づき，パスのためにボールを蹴り出すロボットを“パサー”，ボールを受け取るロボットを“レシーバ”と呼ぶ．人間のサッカーにおけるパス行動の意義は，「ドリブルよりも素早くボールを移動することが出来るため，相手チームの選手をボールから引き離せる」という点にある[70]．従って，自チームの選手の移動速度が相手チームの選手の移動速度よりも劣っている場合，パスによるボールの移動は特に有効なボールの移動手段となる．2012 年における MSL では，試合開始や再開を意味する Kick Off，Throw In，Free Kick，Goal Kick，Corner Kick の 5 つの“Set Play”と呼ばれる行動時に 2 台のロボットによるパス行動が義務付けられているため，Set Play 時においては既に多くのチームが 2 台のロボットによるパス行動を実現している．しかし，試合中における Set Play 以外の状況（各ロボットやボールが常時移動しており，ドリブルやボール取得動作，ディフェンス動作，シュート動作などが次々に行われている状況）では，多くのチームにおいてパス行動の実現には至っていない．この問題の背景には，各ロボットが常に移動している状況では，パスを行う際のチームメイトロボットや相手チームロボットの正確な位置情報を得ることが難しく，さらにパサーがボールを蹴り出した後，レシーバにボールが渡るまでのわずかな時間にもレシーバや相手チームロボットの位置が動的に変化してしまうため，レシーバが確実にボールを受け取れない状況が想定されるためと考えられる．パス行動を実現するためには，パサーがパスを出してからレシーバがボールを受け取るまでの，状況の変化を可能な限り予測する必要がある．

人間のサッカーにおいてパス行動を実現するためには，ボールを狙った位置に放つキックの技術やパスの受け取りが可能な状況にある味方選手を見つける視野の広さ，素早くパスを出す判断力，パスを出すことによる戦況の変化を予測する洞察力などがパサーに求められる[70]．MSL においてパス行動を実現するためには，パサー自身の正確な自己位置情報とレシーバの位置情報，相手ロボットの位置などを把握している必要がある．また，相手チームのロボットの移動も考慮に入れる必要がある．さらに高度なパス行動を実現するためには，「パスを繋ぐことで相手チームのゴール方向へ進行する」「相手チームロボットに囲まれた際は，後方へパスを出す」などの

チーム戦略を考慮する必要がある．しかし，MSL のように常に周辺状況が動的に変化している環境では，パサーが直面する全ての状況を設計者が想定し，事前にプログラミングすることは困難である．また，相手チームの行動アルゴリズムを事前に知ることは出来ないため，相手ロボットの行動を事前にモデリングし，予測することも困難である．これまで行われてきたパス行動に関する先行研究として，無線 LAN 通信を介して「パス行動開始の宣言とレシーバの指名」と「パス受け取りの可否」に関するメッセージをパサーとレシーバ間でやり取りする手法[71]や，強化学習を複数のロボットに適応して共進化によってパス行動の自律的獲得を目指す手法[72]などが挙げられる．メッセージのやりよりによってパス行動を実現するため，通信による情報の授受に関する実時間性が求められるが，2010 年における MSL ではロボットが 1.0[sec]あたり 5.0[m]程度移動する可能性を考慮しなければならないため，通信の遅れによってパス行動を阻害される可能性が高まるという問題点が上げられる．共進化に基づいたパス行動の獲得では，行動の獲得に強化学習を用いており，さらに相手チームロボットの行動やチームメイトロボットの行動など複数の要素に基づいてパス行動が獲得されるため，意図した行動を創発し難いという問題点がある．

そこで，本章では味方ロボットや相手チームロボットの位置や移動速度情報に基づいて，パサーにおけるパスの実行容易さ，レシーバのパスの受け易さ，敵ロボットによるパスコースの妨害可能性，戦略等を確率分布によって表現することで，確率的に最適なパス目標点を選択するアルゴリズムを提案する．自律型移動ロボットが周辺の状況から最適な行動を確率的に選択する行動は，人間の意思決定に近いと考えることが出来る．本章において提案する手法によってパス行動を実現することで，複数の自律型移動ロボットの協調行動に関する知見を得ることができる．

Table 13 History of RoboCup Middle Size League Rules

‘97	RoboCup MSL was started
‘98~’01	No change
‘02	The surround walls were removed, and put the black-white and blue-yellow poles
‘03	The black-white poles were removed Field size changed to 7x10[m] from 5x8[m]
‘04	Field size changed to 9x12[m] from 7x10[m] The referee box was tried to introduce
‘05	The referee box was introduced (Our team began to participate to the RoboCup MSL)
‘06	Direct goal on set play was prohibited
‘07	Field size changed to 12x18[m] from 9x12[m] / Six robots could be entered Set play was introduced
‘08	Deleted goal colors
‘09	Pass behaviors were forced to be introduced
‘10	Ball color became ambient color from orange

5.2 パス行動における確率的行動決定手法

サッカーにおけるパス行動では，①ドリブル行動を継続すべきか，パス行動によって他のチームメイトにボールを渡すべきかを判断，②パス行動を行う場合どの地点にボールを蹴り出すことがチームにとって有効かを判断，③パスの目標地点として複数の候補が考えられる場合，最終的にどの地点にパスを出すべきかを判断，という大別して 3 つの判断が求められる．そこで，本研究ではパス行動を以下の 3Step により実現する．

Step 1：相手チームロボットの位置情報からパス/ドリブルの行動を選択する

Step2：チームメイトロボットや相手チームロボットの位置，戦略を考慮した確率分布・条件付けに基づき，パスを出す目標地点を評価する

Step3：評価結果に基づいてパスを出す目標位置を確率的に選択する

上記した 3Step は全て，事前情報や試合中に得られる各ロボットの位置情報，速度情報に基づき，確率分布と条件付けによって以下のように表現する．

確率分布 1：相手ロボットの位置と移動範囲　⇒Step1，Step2 に使用

確率分布 2：パサーの位置と方位　⇒　Step2 に使用

確率分布 3：レシーバの位置と方位　⇒　Step2 に使用

条件 1：基本(下位)戦略　⇒　Step2 に使用

条件 2：上位戦略 ⇒　Step2 に使用

Step1では，確率分布1に基づき，相手チームロボットによってボールを奪われる確率が一定以上に達した際，パス行動を始めるものとした．Step2では，上記した3つの確率分布と2つの条件付けを掛け合わせた“パス目標位置選択確率”を形成する．Step3では，パス目標位置選択確率に基づき，パスの目標点を決定する．次節に，Fig. 64を例として各確率分布と条件の詳細を記す．

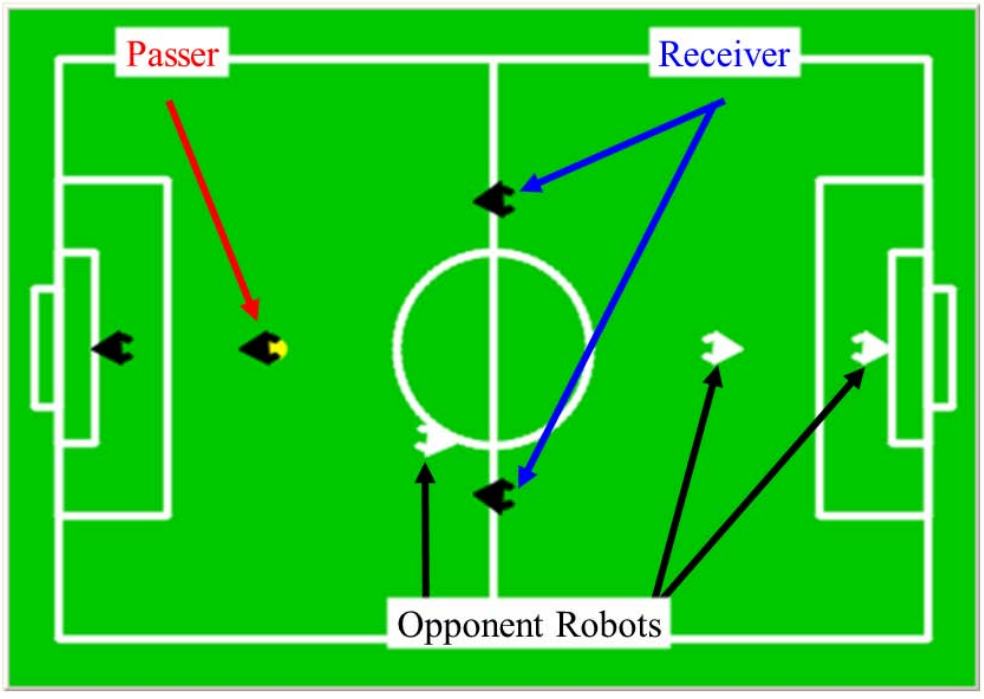


Fig. 64 Example of Pass Behavior Situation

5.2.1　行動選択確率分布と戦略条件の定義

確率分布 1：“インターセプト(Intercept)確率”

本研究では，相手ロボットの位置と移動範囲から“相手チームロボットによりボールを奪われる可能性、パスを阻害される可能性”を表現した確率 P_{opp} を“インターセプト確率”を定義する．MSL の試合では，事前に相手チームロボットの行動アルゴリズムを知ることが出来ないため，相手チームロボットの行動や次状態における移動速度を正確に推定することはできない．そこで，本研究では，各チームが公表しているロボットの最高移動速度と試合中に観測した相手チームロボットの位置から，相手ロボットが移動する範囲を確率分布によって表現した．インターセプト確率は式(5.1)によって Fig. 65 のように表現される． Fig. 65 において，白い三角形は相手チームロボットを表しており，相手チームロボットの位置を中心にインターセプト確率を分布させる．Fig. 65 は，インターセプト確率が低い位置ほど白く表されるものとし，黒いほどインターセプト確率が高い（相手チームロボットによってボールを奪われる可能性が高い）ものとした．なお Fig. 64 に示すような状況では，2 台の相手チームロボットの間が最もボールを奪われる可能性が高いため，各相手チームロボットの中心ではなく，2 台のロボットの間を最も黒く表現している．ここで，p^x，p^y はフィールド上における任意の位置，$m^j_{x,y}$ は j 番目の相手チームロボットの位置，σ_{opp} は相手ロボットの最高移動速度，N は相手ロボットの台数(ゴールキーパを除く)を示す．

$$p_{opp}\left(p^x, p^y\right) = 1 - \left(\sum_{j=1}^{N} \prod_{k=x,y} \frac{1}{\sqrt{2\pi}\sigma_{opp}} \exp\left(-\frac{\left(p^k - m_k^j\right)^2}{2{\sigma_{opp}}^2} \right) \right) \tag{5.1}$$

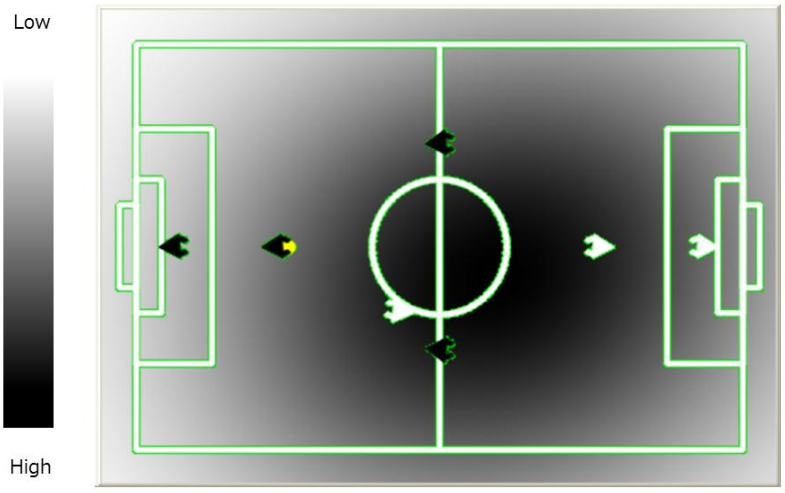


Fig. 65 Image of Intercept Probability Distribution

確率分布 2：“パサー確率”

本研究では，ある時刻におけるパサーの位置と方位に基づいて，“パサーの姿勢とパス目標地点に対するボールの蹴り易さ”を表現した確率を“パサー確率 P_{pas}”と定義する．Musashi は本体正面にキック機構を持つため，パサーがパス行動のためにボールを蹴り出す際，蹴り出す方向へ姿勢を変化させる(回転する)必要がある．回転量が大きいほどパスの行動開始から実行に至るまでの所要時間は大きくなるため周辺状況の変化（チームメイトロボットの位置や相手チームロボットの位置の変化）を伴うこととなり，パスの成功確率が低下すると仮定できる．同様に，パサーからパス目標地点までの相対距離が遠いほど，ボールが目標地点へ到達するまでの所要時間が大きくなるため，パスの成功確率は低下すると仮定できる．従って，パサー確率は式(5.2)によって Fig. 66 のように表現される．Fig. 66 では，パサー確率が高い位置ほど白く表されるものとし，黒いほどパサー確率が低いものとした．パサー確率が高い位置は，その位置に対してパサーが即座にパスを出せることを示す．ここで，θ はロボットの正面方向からフィールド上の任意の位置に対する相対角度，σ_{pas} はパサーがパスを出すことが出来る姿勢の限界角度(ここでは 70[deg]とした)を示す．

$$p_{pas}(\theta) = \frac{1}{\sqrt{2\pi}\sigma_{pas}} \exp\left\{ -\frac{\theta^2}{2\sigma_{pas}^{\ 2}} \right\} \tag{5.2}$$

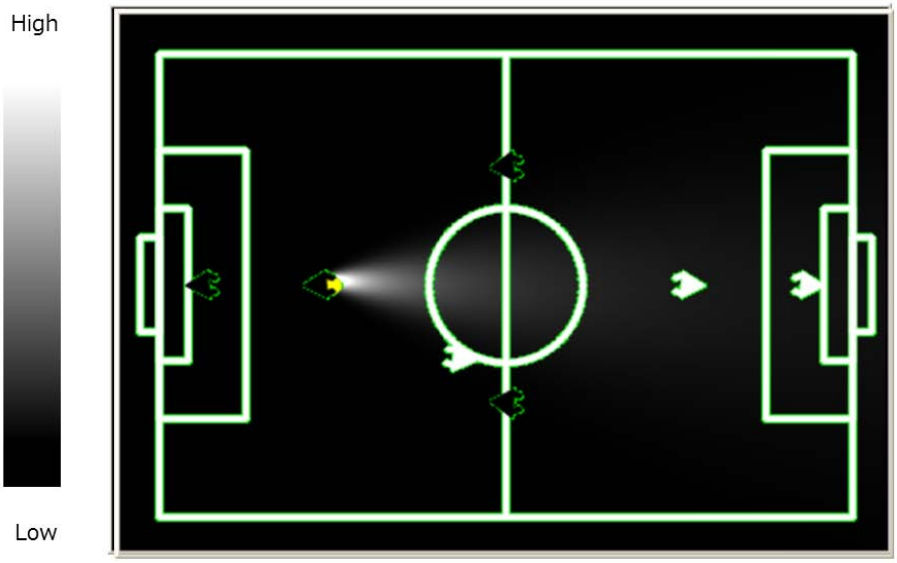


Fig. 66 Image of Passer Probability Distribution

確率分布 3：“レシーバ確率”

本研究では，ある時刻におけるレシーバの位置と方位，移動速度に基づいて，“レシーバのパスの受け取り易さ”を表現した確率を“レシーバ確率 P_{rec}”を定義する．レシーバにおいても Musashi が本体正面にキック機構を持つことから，レシーバの正面方向へボールが蹴り出される程，パスを受け取れる可能性が高まると考えられる．本研究では，レシーバの正面方向に対して 1.0[m]手前を最もパスを受け取りやすい位置を仮定し，式(5.3)によって Fig. 67 のように表した．ここで，p^x，p^y はフィールド上における任意の位置，$m^i_{x,y}$ は i 番目のチームメイトロボットの正面方向に対する 1.0[m]前方の位置，σ_{rec} はレシーバがボールを受け取れる範囲，M はレシーバの台数(ゴールキーパを除く)を示す．

$$p_{rec}\left(p^x, p^y\right) = \sum_{i=1}^{M} \prod_{k=x,y} \frac{1}{\sqrt{2\pi}\sigma_{res}} \exp\left(-\frac{\left(p^k - m^i_k\right)^2}{2\sigma_{res}^{\ 2}} \right) \tag{5.3}$$

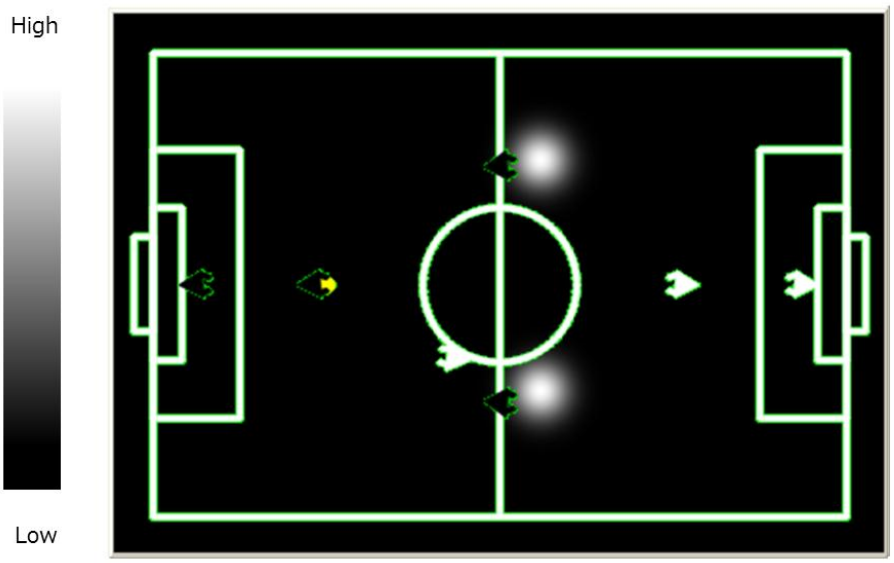


Fig. 67 Image of Receiver Probability Distribution

条件 1：“基本（下位）戦略”

本研究では，パス行動の方針を“基本戦略”を定義して決定する．ここで，パス行動における方針として，相手ゴールへ近づくほど得点を得る機会が増えることから，「より相手ゴールに近いレシーバへパスを出す」という戦略を仮定した場合，式(5.4)によって Fig. 68 のように表される．ここで，Q はパスの目標位置としての価値を表しており，p^x は任意の x 座標，d_{max} はフィールドの全長(18.0[m])を表す．

$$Q\left(p^x\right)=\frac{1}{d_{\max}}p^x+\frac{1}{2} \tag{5.4}$$

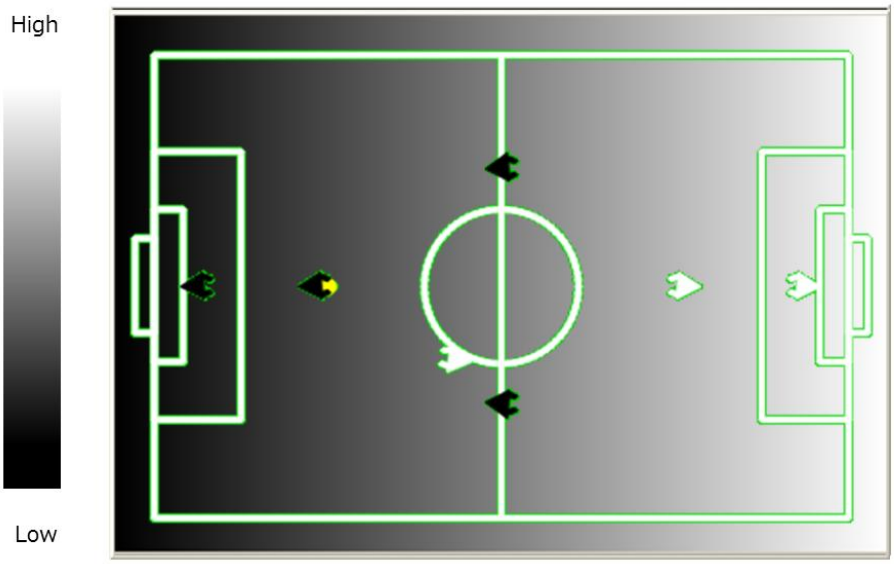


Fig. 68 Image of Base Strategy Condition

条件 2：“上位戦略”

本研究では,「どのような条件でパスを出すか」「どのような球種を選択するか」といった,「チームとしてどのようなパス行動を行うか」を決定する戦略のことを“上位戦略”と呼ぶ．本研究において設定した上位戦略は，パスの意義である「ドリブルより素早くボールを移動することが出来る」という考えに基づき，戦略①パサーを中心に，一定範囲内ではパスを行わない，という条件を定めた．

また，Musashi はループパス(ボールを空中に浮かせるようにキックするパス)とグラウンダーパス(ボールが地面を転がるようにキックするパス)などの球種の蹴り分けが出来ないことや，レシーバとなる Musashi がボールを受け取りやすいといった点から，パス行動にはグラウンダーパスを主体として行うものとし,戦略②パサー視点から相手チームロボットの後方となる位置にはパスを行わないという条件を含めた．戦略①と②は Fig. 69 のように表される．Fig. 69 において，黒い領域はパスを行わない領域を表しており，白い領域はパスを行う領域を表している．

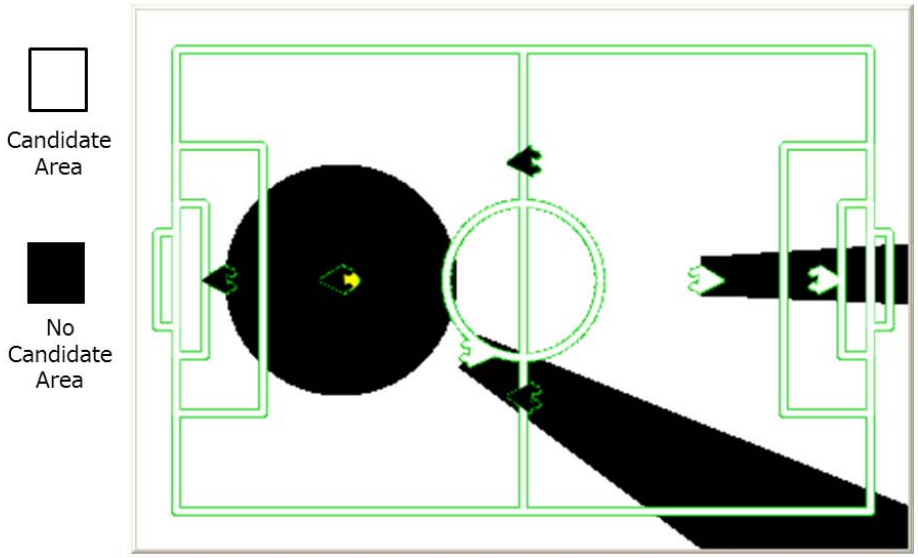


Fig. 69 Image of Strategy Condition

5.2.2　パス目標位置選択確率

パサーがボールを蹴り出すための目標位置は, 3つの行動選択確率分布と2つの戦略条件を統合することにより形成される．本研究において定義した3つの行動選択確率分布は，それぞれパサー，レシーバ，相手チームロボットの情報に基づいて定義しており，各戦略はロボットの状態に依存しないため，各確率や各条件は独立であると仮定し，各行動選択確率分布と条件は論理積によって統合する．Fig. 70に統合の概念と統合後の結果を示す．Fig. 70に示す結果 (Probability Distribution of Pass Target Position) では, パサーからほぼ同程度の距離に位置する2台のレシーバに対し，片方のレシーバが相手チームロボットの背後に位置していることから，Fig. 64における上方に位置するレシーバ周辺でパス目標点選択確率が高まっていることがわかる．

本研究では，最終的なパスの目標位置をFig. 70に示すProbability Distribution of Pass Target Positionの確率分布に従って，確率的に選択する．従ってFig. 64における上方に位置するレシーバ周辺がパス目標位置として選択されやすく，Fig. 64における下方に位置するレシーバ周辺は選択されにくくなる．最終的なパス目標位置選択の概念をに示す.

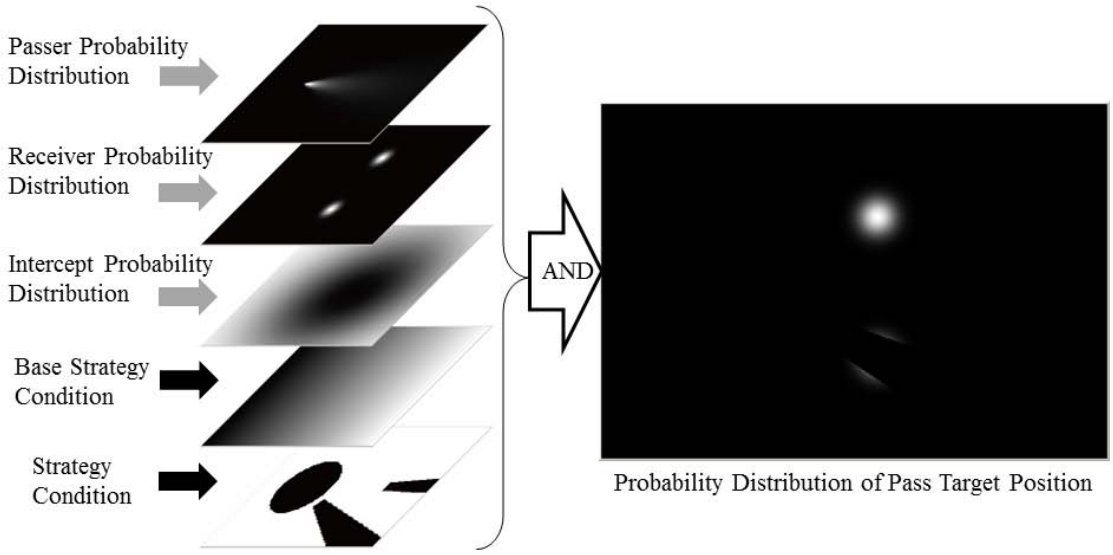


Fig. 70 Integration Image and Probability Distribution of Pass Target Position

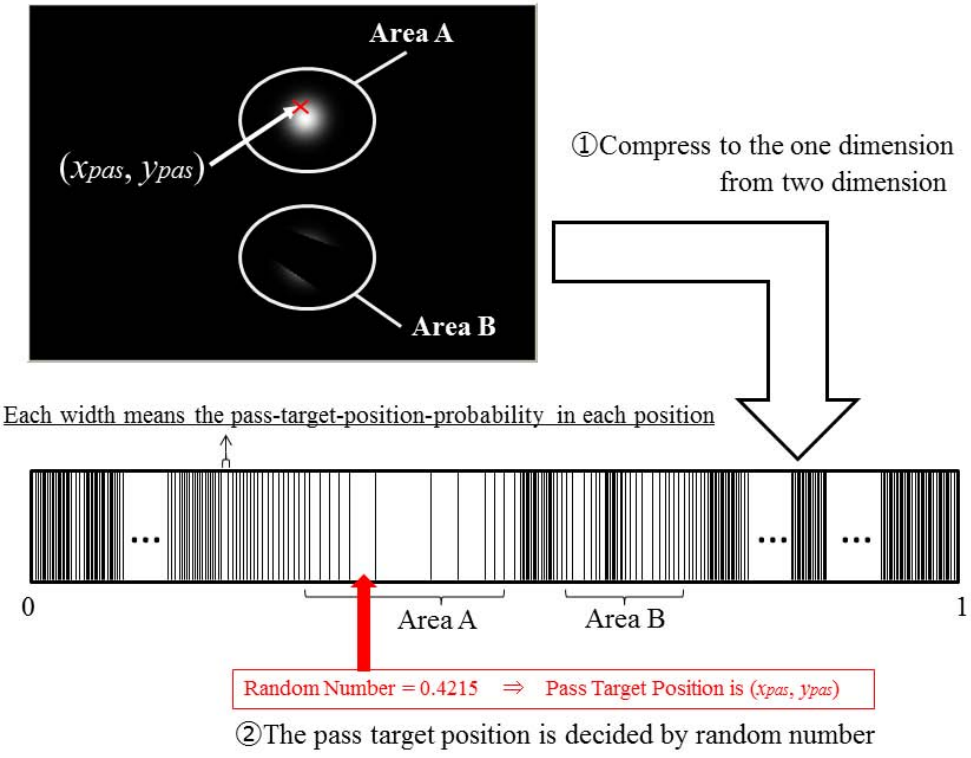


Fig. 71 Concept of Psss-Target-Position Decision Process

5.3 パス行動評価実験

ODEによる動力学シミュレーション環境を用いて，パス行動における提案手法の評価実験を行った．パス行動の評価は式(5.5)に示す評価関数を用いた．ここでd_pはパサーの走行距離を示しており，d_bはボールの総移動距離を表す．実験では，パサーとなるロボットがボールを取得した瞬間から，レシーバへボールを蹴り出し，レシーバがボールを受け取る直前までの期間，ボールを所持していたロボットをパサーとする．またt_bは自チームのロボットがボールを所持した時間を示し，t_cはパサーが相手チームロボットと接触した時間，tはロボットがペナルティエリアまで進行するまでに経過した時間を示す．ここで実験の終了条件は，①自チームのロボットがペナルティエリアを通過する，ペナルティエリアを通過しないまま60[s]が経過する，とした．式(5.5)において，右辺第1項は“距離に基づいたボールキープ率”を示しており，パサーの移動距離d_pに対するボールが移動した距離d_bを表す．距離に基づいたボールキープ率は，値が高いほど“相手ディフェンダーをより撹乱した”ことを表している．右辺第2項は，“時間に基づいたボールキープ率”を示しており，所要時間tに対するパサーがボールを所持していた時間t_bを表す．時間に基づいたボールキープ率は，自チームがより長い時間ボールを占有するほど高い値を示す．右辺第3項は“非接触率”を示しており，所要時間tに対するパサーと相手チームロボットが接触していなかった時間(t-tc)によって表される．非接触率が高いほど，相手チームロボットとの接触が無かったことを表しており，衝突によるロボットの破損といった危険性が低いことを表している．式(5.5)に示す評価関数は，これらの項目に基づいて“オフェンス行動の適切さ”を表すものとする．

$$E = \frac{d_b}{d_p} \cdot \frac{t_b}{t} \cdot \frac{(t - t_c)}{t} \tag{5.5}$$

実験では，フィールド上に自チームのロボットが4台，相手チームロボットが4台存在する状況を想定し，自チームロボットの移動速度を1.0[m/s]，相手チームロボットの移動速度を1.5[m/s]とした．また，相手ロボットの行動は「ボール方向へ直進する」のみとした．実験環境における各ロボットの初期位置をFig. 72に示す．各ロボットの位置は，フィールドの中心を原点とした座標系上に表現され，自チームロボットの初期位置を(-5.0, 0.0)，(1.0, 3.0)，(0.0, -3.0)，(-9.0, 0.0)，相手チームロボットの初期位置を(5.0, 0.0)，(-1.0, -2.0)，(0.0, 1.0)，(9.0, 0.0)とした．評価はドリブル行動のみと，パス行動を考慮した場合の2つの条件に対し，式(5.5)の評価関数を用いて行った．なお，ドリブル行動のアルゴリズムについては，**2.3.3.4節 行動判断部**を参照されたい．

5.3.1 シミュレーション実験の結果

Fig. 73とFig. 74に実験結果の様子を示す．Fig. 73はパス行動を行わずドリブル行動のみによってオフェンスを行った際の様子を示しており，Fig. 74はパス行動を考慮したオフェンスを行った際の様子を示している．Table 14に，各行動に対する評価関数結果を示す．Fig. 73からドリブル行動のみを行った場合，ロボットの移動速度よりも相手チームロボットの移動速度が0.5 [m/s] 速いため，2台の相手チームロボットによるディフェンスを受け，相手ゴール方向へ進めない状態となった (t=3.0[s]) . t=3.0[s] 以降は, 3台目の相手チームロボットのディフェンスが加わり，ロボットは壁に押し付けられる形となった．t=9.0[s] 以降，ロボットは相手チームロボットのディフェンスから抜けドリブルを再開したが, 実際の試合では壁に押し付けられた状態からドリブル行動を再開することは出来ないため (フィールドから出た時点で相手チームのThrow Inとなる) , 実際の試合ではドリブル行動だけでは相手チームロボットによるディフェンスを突破出来ないと考えられる．Table 14が示すようにドリブル行動のみの場合，ロボットは60.00[sec]を越えるまで相手チームロボットのディフェンスを突破することが出来ず, 終了条件に入った. この際, ロボットは常にボールを所持した状態で相手チームロボットによるディフェンスを受けたため, d_pとd_bはほぼ同じ走行距離を示した．またt_bについても評価実験の終了時間 (60.00[s]) までボールを所持し続けたため，tとほぼ同じ値を示した．tcは55.30[s]を示し，常に相手チームロボットによるディフェンスを受けていたことを示した．式(5.5)による評価値は0.08を示した.

Fig. 74から，パス行動を考慮した場合，相手チームロボットのディフェンスによりロボットのオフェンスが妨げられていないことがわかる．Fig. 74におけるt=3.0[s] では，相手チームロボットの接近によって，パサーが図上側のレシーバへパスを出している．また，t=6.0[s] においても，相手チームロボットの接近により，新たにパサーとなったロボットが他の図下側のレシーバへパスを出している様子がわかる．Table 14が示す結果から，パス行動を考慮した場合ではtは21.96[s] を示し，自チームのロボットがペナルティエリアを越えたことを示している．また，d_pが17.60[m]を示したのに対し，パスによってd_bは24.87[m]を示したため，距離に基づいたボールキープ率はドリブル行動よりも高い値を示した．t_bの結果では，パス行動によってボールがロボットから離れるため9.35[s]を示しドリブル行動よりも短くなった. これにより時間に基づいたボールキープ率はドリブル行動よりも低い0.43を示した．この結果から，パス行動によってボールを所持していない状態が生じており, 相手チームロボットのボール取得行動によってボールを奪われる可能性があると考えられる．t_cの結果では，相手チームロボットとの接触時間は8.57[s]であり，ドリブル行動のみの場合と比較すると短い時間であったことがわかる．また，Fig. 74から接触時間として記録されたのは，各パサーがパスを行った後であり，t_cが示す接触により自チームのオフェンスが妨げられていないことがわかる．式(5.5)による評価値は0.37であり，パス行動を考慮した場合，ドリブル行動のみを行う場合に比べ，適切なオフェンスが行われている．

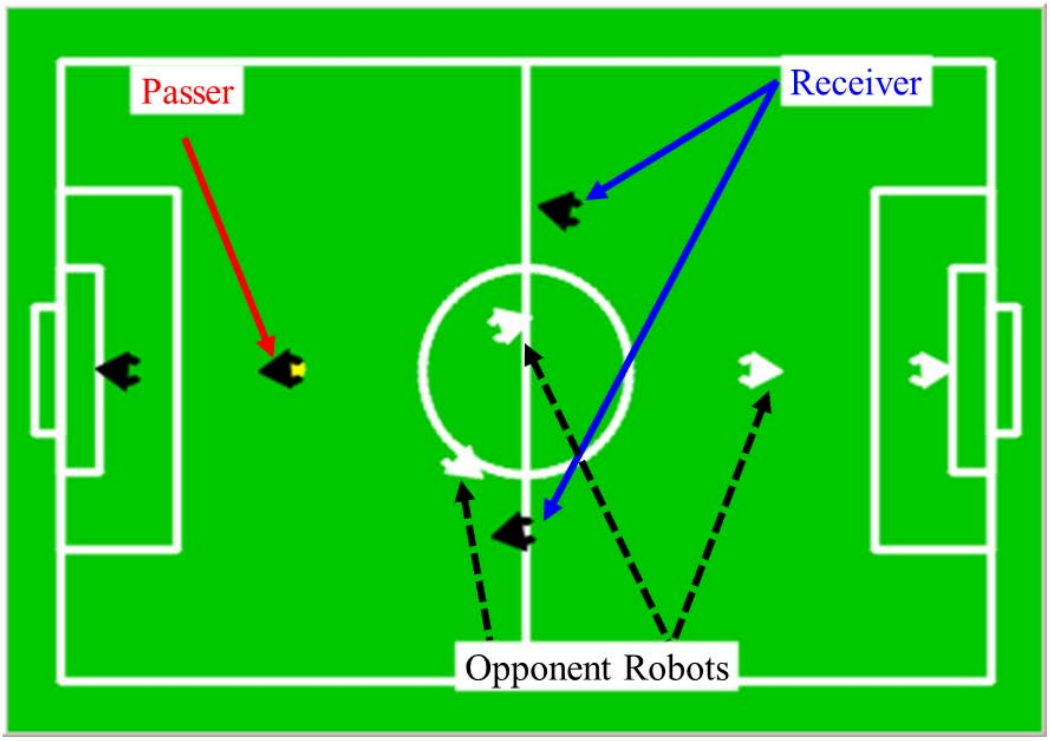


Fig. 72 Initial Position of Each Robot

Table 14 Evaluation Result of Dribble and Pass Behavior

	Dribble	Pass
d_p[m]	35.10	17.60
d_b[m]	34.94	24.87
d_b/d_p	0.99	1.41
t[s]	60.00	21.96
t_b[s]	59.03	9.35
t_b/t	0.98	0.43
t_c[s]	55.30	8.57
$(t-t_c)/t$	0.08	0.61
E	**0.08**	**0.37**

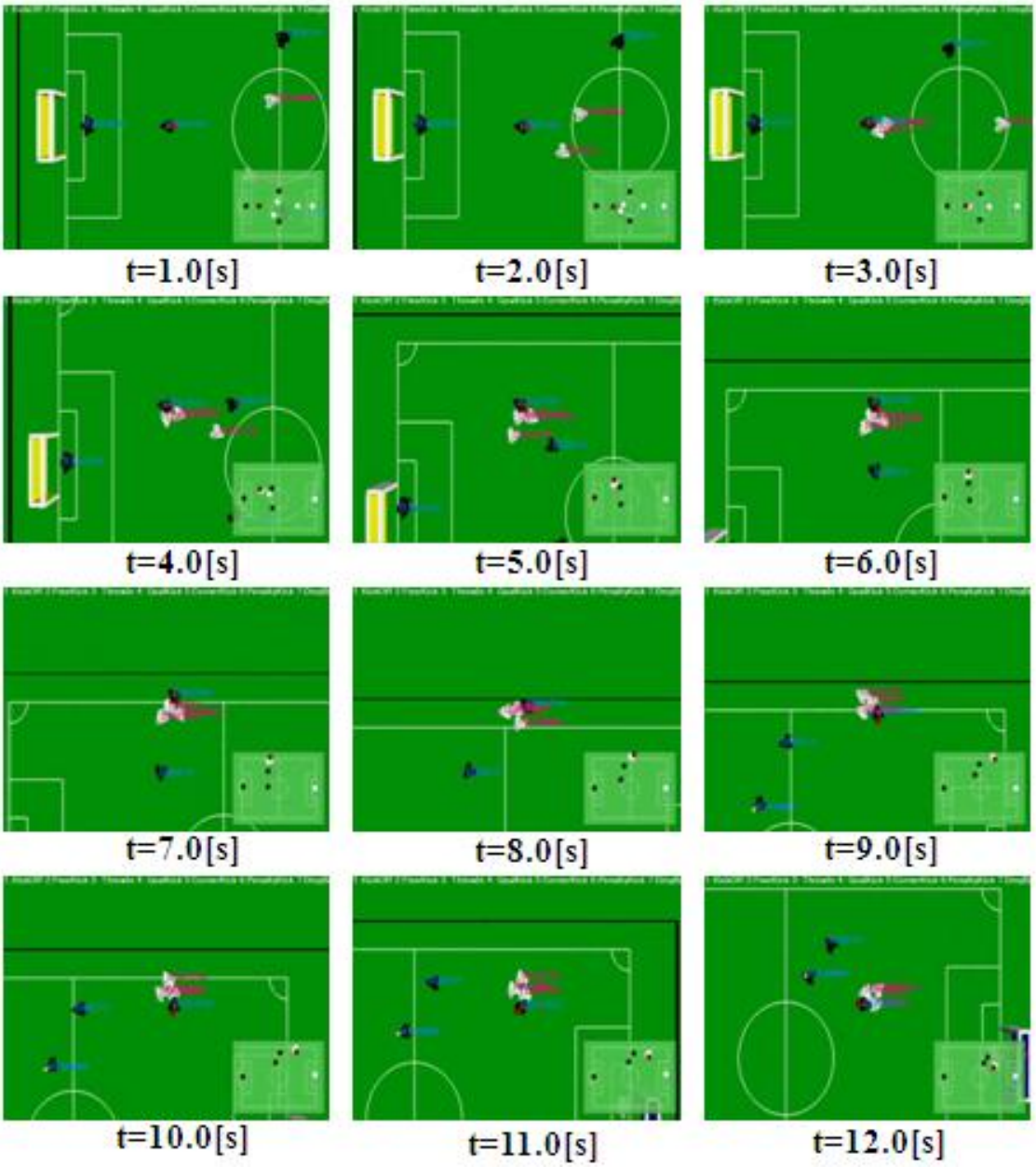


Fig. 73 Result of Dribble Behavior

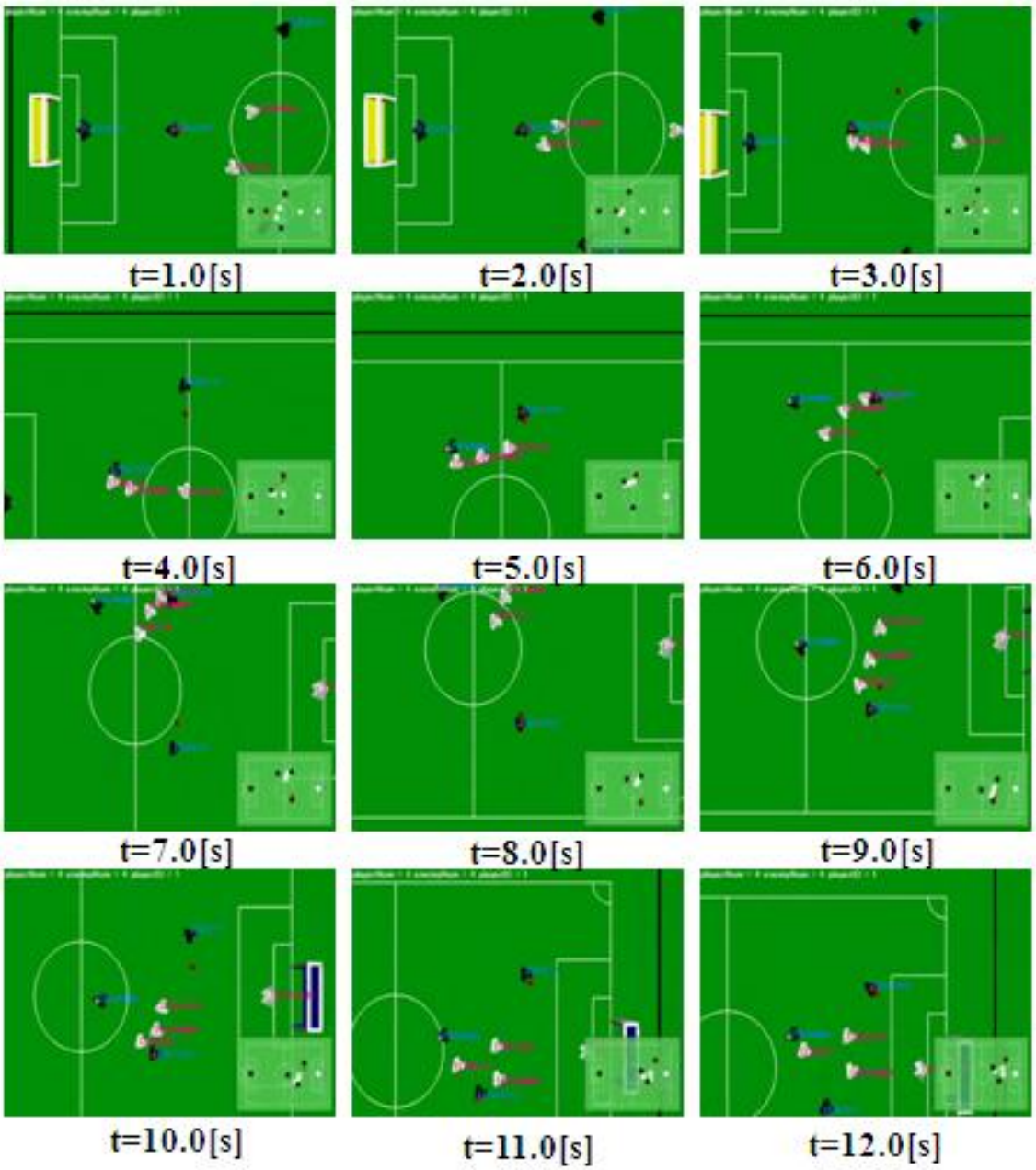


Fig. 74 Result of Pass Behavior

5.4 考察

本章では，パス行動における行動決定法について，自チームのロボットや相手チームのロボットの位置や姿勢の変化を確率分布によって表現し，確率的にパスの目標地点を選択することで自律移動型ロボットのより知的な行動決定法を提案した．MSL においてパス行動を条件分岐のみで設計する場合，どのような状況でパス行動を開始するか，各レシーバはどのような状況にあるか，相手チームロボットとの相対位置関係はどのような状況にあるか，どのような戦略に基づいてパスを行うか，レシーバへパスを出した際，ボールが到着するまでにどのような周辺状況の変化が発生するかなどを事前に考慮し，考えうる限りの状況を条件として記述する必要がある．10 台のロボットとボールが常時移動する MSL の環境では，そのような状況を全て事前に予測することは困難であり，多くの不確実性が生じることとなる．

本章において提案した手法は，周辺状況の変化を確率分布によってモデリングし，それらを確率論に基づいて統合することで，周辺状況の変化を予測する．評価実験では，パサーとなるロボットは，相手チームロボットによってボールが奪われる可能性を考慮し，相手チームロボットによるディフェンスを受けることなく，パス行動を実行した．また各レシーバの位置や相手チームロボットの位置，戦略を考慮し，より相手ゴールに近く，周辺に相手チームロボットが存在しないレシーバを選択してパス行動を行った．また，提案手法では最終的なパス目標地点を確率に基づいて選択しているため，選択確率の低い位置であってもパス目標位置として選択される可能性がある．従って，パスの目標位置を決定する各確率分布や戦略が，相手チームに知られている場合においても，パサーが選択したパスの目標位置を，相手チームのロボットが完全に予測することは困難となる．このように確率的に行動が選択されることで，相手チームにはより高度な状況判断が求められることとなり，MSL における技術発展が大きく促されることが期待出来る．本章において行った評価実験では，基本戦略や上位戦略の変化に対する行動結果の変化については評価を行なっていないが，得点状況に応じたパス戦略の変更や，ループパスを用いた相手チームロボット後方へのパスなど，より複雑な行動を条件として表現することで，意図したパス行動を実現できると考えられる．

今後の展望として，パス行動だけでなくボール取得行動やドリブル行動の経路計画，シュート行動等においても確率的な行動選択が適用されることが期待される．周辺環境の変化に応じて自身の最適な行動を選択する行為は，人間の行動選択過程を想起させるものだと考えられる．

第六章

考察と結論

第六章　考察と結論

本論文では，ロボカップサッカー中型リーグにおける自律移動型ロボットを対象とし，動作環境に内在する不確実性に対し，強化学習を用いた自律的行動獲得と確率ロボティックスを用いた情報信頼性向上アルゴリズムについて提案，及び実機による検証を行った．本章では，それぞれの提案手法についての考察を述べ、結論をまとめる．

Q学習を用いたゴールキーパロボットの守備行動学習では，離散空間における学習結果を連続空間へ適用した際の性能評価を目的として，評価実験を

①学習環境と同様の離散空間におけるシミュレーション

②厳密には離散空間であるが動力学を考慮したODEによるシミュレーション

③“Musashi”のゴールキーパを用いた実機実験

という3つの実験条件に分け，学習結果に対し，徐々に連続空間としての要素を付加した際の性能について評価を行った．実験①と実験②の実験条件を比較すると，離散空間において保障されるマルコフ決定過程が，ロボットの移動に伴う加速度や慣性力の影響により失われ，保障されなくなるという点が異なる．また，ロボットやボールの位置，速度に関する情報の離散化が実験①の実験時(0.1[m]単位で離散化)に比べさらに密になる(double型で位置を表現)ため，動作環境が限りなく連続空間に近く，量子化誤差の影響を受ける点も異なる．3.4.5節に示した実験②の実験結果から，離散空間における学習結果をODE環境へ適応した場合，観測状態に対して局所的に行動価値Qが著しく低い(0に近い)状態が生じていることが分かった．しかし，ロボットの行動軌跡では行動価値Qが得られなかった場合においてもロボットは停止することなく，ボールとの相対距離を縮める方向へ移動を行った．ここで，動力学シミュレーションの環境では，ある状態においてロボットが停止動作を選択した場合でも，それまでの移動に伴う慣性力の影響によって，ロボットが即座に停止できず状態の遷移が生じることが考えられる．従って，マルコフ決定過程が保障されない動的環境下におけるロボットの行動は，一つの行動価値Qを基に選択された行動aによって定まるのではなく，一定時間において観測した複数の行動aの頻度から決定されるものと考えることが出来る．つまり，離散空間シミュレーションでは，ロボットが遭遇したある状態に対して方策に基づいた一つの行動が定められ実行されるが，連続空間シミュレーションでは，ロボットが一定時間内に選択した複数の行動から，最も高い頻度で選択された行動が実行される．この傾向はロボットの移動速度や重量が増加する程，明確に表れると考えられ，各状況に応じた俊敏かつ最適な行動を実行する上で妨げとなると考えられる．

実験②と実験③の実験条件を比較すると，ロボットの自己位置やボールの観測位置に誤差が含まれるという点が大きく異なる．観測状態と選択した行動を関連付けて学習する強化学習において，観測状態と実際にロボットが置かれた状態に差が生じる場合，観測状態に基づいて選択した

行動は学習過程に基づいたものではないため，不適切な行動が選択される可能性が高い．観測誤差の程度については，3.4.6節の実験結果におけるODEシミュレーションが示すボールの位置と実機が観測したボール位置との誤差から判断出来る．本研究において行った実験では，ロボットの状態遷移は停止，右方向並進移動（アーム展開の有無を含む），左方向並進移動（アーム展開の有無を含む）の3パターンのみであったため，観測情報に含まれる誤差に対し不適切な行動を行う可能性は低かったと考えられるが，実機実験における守備率が示した結果と守備行動におけるロボットの軌跡から，マルコフ決定過程の欠如や量子化誤差の影響に対して，守備行動を適切に学習していると判断することが出来る．3.5節において述べた考察では，実機実験における守備率が高い点や，ε-greedy手法に比べsoftmax手法 のほうが高い守備率を示した結果についてsoftmax手法における開拓行動によるものと考察したが，ε-greedy手法とsoftmax手法における探索行動に関する評価は不十分である．探索行動に関する評価は，実験①と同様，もしくは近似出来る程度のテストパターンを連続空間において行う必要がある．

第3章において行った実験結果から，ゴールキーパロボットにおける守備行動の学習では，ガウス分布を用いた報酬関数を設定し，ε-greedy手法を方策として用いることで，約90[%]の守備率を示す守備行動を獲得可能であることを示した．また，離散空間における学習結果を連続空間に適用した場合，softmax手法を方策として用いることによって，1.0，2.0，3.0[m/s]のボール速度に対し約90[%]の守備率を示す守備行動を獲得可能であることを示した．

確率ロボティクスを用いた協調自己位置推定アルゴリズムの提案では，観測情報に誤差を含む複数のロボット間において情報を共有・統合することで情報の信頼性向上を目指した．各ロボット単体におけるランドマーク（ボール）の位置推定では，ロボットの自己位置に関する推定誤差，方位推定誤差，ボールとの相対距離と相対角度に関する観測誤差の4種類の誤差に起因してランドマークの位置に誤差が生じる．本研究では，ロボットの自己位置推定誤差とボールとロボットとの相対距離の観測誤差を2次元正規分布によって表現し，またロボットの方位推定誤差とボールとロボットとの相対角度の観測誤差を，ボールとロボットを結ぶ直線上を最大値とする正規分布によって表現した．正規分布を用いた誤差の表現は，ランドマークであるボールの存在確率を意味しているため，各ロボットと共有した各観測情報を基にボールの存在確率を評価することで，ボールの位置として最も存在確率が高い地点を特定することが出来る．本研究では，ランドマークの位置推定手法を協調推定と呼び，各観測点において各ロボットの観測結果の平均誤差よりも低い誤差を示した．4.3節に示した実験結果から，協調推定のメリットとして各ロボットの観測情報が含む平均誤差を減少出来る点と，ボールが観測出来なかったロボットに対しても信頼性の高いボール位置情報を提供出来る点があげられる．デメリットとしては，ある１台のロボットの観測情報に含まれる誤差と標準偏差が小さく正確にボールの位置を推定しており，他のロボットの観測情報に含まれる誤差と標準偏差が大きい場合は，正確にボールの位置を推定していたロボットに対し，誤差の増加や標準偏差の増加を招いてしまう可能性があることがあげられる．

このデメリットへの対策として，自身の観測情報の信頼性と，情報を共有した他のロボットの観測情報に関する信頼性に対し，ランクを与える方法が考えられる．本研究では，観測情報や推定情報が含む誤差のみを数理モデルによって表現したが， 観測したボールの位置と協調推定により推定したボールの位置との相対距離等を基準として，相対距離が小さい程，高いランクが与えられるものと設定すると，協調推定結果に近い位置に常にボールを推定したロボットが，全ロボット中最も信頼できる観測情報を所持していると仮定できる．さらに与えられたランクに応じて，そのロボットが示すボールの存在確率を高くするといったアプローチにより，協調推定によるランドマークの位置推定精度の向上が見込めると考えられる．

また，協調自己位置推定が示す結果では，各ロボットの自己位置推定誤差が小さく，ボールの観測結果に大きな誤差や標準偏差が含まれる場合，協調自己位置推定を行うことで，誤差や標準偏差の増加が見られることが分かった．また誘拐状態を想定した実験では，協調自己位置推定後もロボットの自己位置に誤差や標準偏差は含まれるが，ロボットの真の位置付近に自己位置を更新しており，自己位置に含まれる誤差を軽減出来ていることが分かった．これらの結果から，協調自己位置推定は，誘拐状態において誤差の軽減を顕著に示し，誘拐状態にないロボットに対しては若干の誤差の増加を招くことが分かった．従って，今後，協調自己位置推定によって得られた自己位置更新の移動量に基づいて，協調自己位置推定によって自己位置を更新するか否かを判断する方法が有効だと考えられる．つまり，協調自己位置推定によって得られた更新の移動量が大きい場合は，そのロボットが誘拐状態にある可能性が高いと仮定して自己位置を更新し，更新の移動量が小さい場合は，そのロボットは真の位置付近に自己位置を推定していると仮定し，協調自己位置推定による自己位置更新を行わないものとする．第４章において示した結果から，協調推定はランドマークの位置推定精度を向上させる上で有効であり，特にボールが観測出来ないロボット等へのボール情報の提供に関して，信頼性の高い情報を提供可能である．また，協調自己位置推定は，誘拐状態にあるロボットの自己位置を真の位置付近へ更新する場合に対して有効であり，特に，自己位置を誘拐状態から回復させるための十分な観測情報が得られない場合 においても自己位置の更新が可能である点で他の手法に比べ有効な手段である．

本研究では自律型移動ロボットの動作環境に内在する不確実性に対し，強化学習による自律的な行動獲得と確率ロボティクスを用いた不確実性のモデル化について提案し，実環境における評価を行った．自律エージェントの行動獲得や不確実性への対処については，様々な手法が提案され，評価結果を報告しているが，実際のロボットを用いた検証については，不十分なものが多い．しかし実時間実環境下で動作する自律型移動ロボットには，観測誤差や推定誤差等，シミュレーション環境内に十分に再現出来ない要素が多く含まれており，シミュレーションによって示された結果が，実機においてどの程度有効であるかを示すことは，自律型移動ロボットの発展において重要である．

本研究では，環境の不確実性に対するロボットの自律的適応例として，ゴールキーパロボットの守備行動に対して強化学習を適用し，シミュレーションベースの学習結果を実機に適用した際の，性能低下について実機実験に基づく評価を行った．学習対象としたゴールキーパロボットの守備行動は，行動としては「向かってくるボールの中心と守備対象であるゴールの中心を結ぶ線上を守備する」という簡単なアルゴリズムによって実現可能であるため、強化学習ではなく設計者によるプログラミングによって解決可能な問題であるが，強化学習によって行動学習を行う題材としては，高次元の状態に対する観測と学習が必要な課題である．本研究において行った実験により，本実験の実験条件のように状態変数や報酬関数を設定することにより，実機においても高い守備率を示す行動を獲得可能であることが示せた．実機を用いた検証に関しては，ε-greedy手法とsoftmax手法の探索行動について評価が不十分であるため，今後，実機を用いた評価実験として離散空間におけるテストパターンと同様の評価実験を行うことで，連続空間への適用に適した方策を明確に定めることが出来ると期待される．

また本研究では，自律型移動ロボットの観測情報に含まれる不確実性への対処として，確率ロボティクスを用いて不確実性をモデル化し，統合することにより，情報の信頼性を向上する手法を提案した．複数のロボットを用いて各ロボットの自己位置を較正する従来の手法としては，複数のロボットがお互いに相手の位置を観測し，互いに算出した相対距離の差分等から位置を較正するものがほとんどであるが，RoboCup MSLの環境下では，一つの動作環境内において所属の異なる二つの群（対戦する二つのチームを指す）が存在し，さらにそれぞれのチームに属する各ロボットは互いに同程度の寸法や似通った形状，本体の色を持つため，カメラ等のセンサにより，測定対象である同じ群に所属するロボットを特定するのは難しい．また，相手チームのロボットによって測定対象であるロボットが遮蔽され，観測出来ない場合も考えられる．提案した手法では，動作環境内に類似する形状や色が存在しないボールをランドマークとして設定したため，チームメイトロボット等をランドマークとした場合に比べ，ランドマークを誤認識し難いという利点があり，また，5台のロボットが観測した結果を共有・統合するため，数台のロボットがボールを観測出来ない場合でもランドマークの位置推定が可能である．実験結果から，協調推定によるランドマークの位置推定では，各ロボットの観測値の平均誤差よりも推定誤差が低く，信頼性の向上が達成出来た．RoboCup MSLの動作環境下では，各情報がどの程度の信頼性を示しているかについて絶対的に評価することが困難であるため，本研究が示した結果のように，各ロボットの情報が含む誤差の平均値と比較して，誤差が減少傾向にあることが明確である場合は，どの情報を共有すべきかを決定する上で有益な指標になると期待出来る．

さらに，協調自己位置推定の実験では，自己位置を較正するための指標としてランドマークを一つ設定するだけで，協調推定したランドマークの位置と自身が観測したランドマークの位置が持つ各存在確率から，自己位置をどのように修正すべきかを提示できることを示した．実験結果から提案手法は，「ボールとロボットとの相対距離が大きく，さらに推定ボールと観測ボールとの相対距離が小さい状況では，修正係数αが小さくなり誤差の減少が期待出来ない」と言えるが，以下に示す条件下では，自己位置に含まれる誤差の軽減効果が期待できる．

１．ロボットとボールとの相対距離が近い

２．推定ボールと観測ボールとの距離が遠い

３．上記の１，２の条件を共に満たす場合

誘拐状態における実験結果では，上記条件の２に該当する状況であったことから，自己位置誤差が軽減された．上記条件は全て，ロボットが持つ群内情報に基づいて表現することが可能であるため，協調自己位置推定によって自己位置を更新する条件として導入出来る．また，動的環境における実験から，観測ボールの位置に誤差が含まれている場合において，自己位置誤差を減少出来ることを示した．本実験では，他のロボットと共有した情報に含まれる時間遅れについて考慮していないため，今後，他のロボットと共有した情報が含む時間遅れに関して，時間遅れに応じた係数等により，その情報が示すボールの存在確率を低下させる方法等について検証を行う必要があると言える．確率ロボティクスを用いた不確実性のモデル化に関する研究では，提案手法が，自己位置推定を行う上で十分な周辺情報を得られないロボットにおける，誘拐状態から回復する有効な手段であることを示した．

第５章では，パス行動における行動決定法について，自チームのロボットや相手チームのロボットの位置や姿勢の変化を確率分布によって表現し，確率的にパスの目標地点を選択することで自律移動型ロボットのより知的な行動決定法を提案した．提案した手法は，周辺状況の変化を確率分布によってモデリングし，それらを確率論に基づいて統合することで，周辺状況の変化を予測したといえる．評価実験では，パサーとなるロボットは，相手チームロボットによってボールが奪われる可能性を考慮し，相手チームロボットによるディフェンスを受けることなく，パス行動を実行した．また各レシーバの位置や相手チームロボットの位置，戦略を考慮し，より相手ゴールに近く，周辺に相手チームロボットが存在しないレシーバを選択してパス行動を行った．

パスの目標位置を決定する各確率分布や戦略が相手チームに知られている場合においても，最終的なパス目標地点を確率に基づいて選択することで，パサーが選択したパス目標位置を相手チームに予測されにくくなると考えられる．このように確率的に行動が選択されることで，相手チームにはより高度な状況判断が求められることとなり，MSL における技術発展が大きく促されることが期待出来る．本章において行った評価実験では，基本戦略や上位戦略の変化に対する行動結果の変化については評価していないが，得点状況に応じたパス戦略の変更や，ループパスを

用いた相手チームロボット後方へのパスなど，より複雑な行動を条件として表現することで，意図したパス行動を実現できると考えられる．パス行動における確率的行動決定手法の今後の展望として，パス行動だけでなくボール取得行動やドリブル行動の経路計画，シュート行動等においても確率的な行動選択が適用されることが期待される．人間は，周辺環境から決定された最適な行動であっても，稀に最適とは言えない行動を選択することによって，よりよい結果を得る場合がある．本研究において提案したパス目標位置選択確率の定義と，最終的なパス目標位置の選択における過程は，上記したような人間の行動に近いと考えられ，自律移動型ロボットのさらなる知能化に繋がるのではないかと期待出来る．

謝辞

謝辞

本論文に記した研究結果をまとめるにあたり，実に多くの方々からのご支援，ご指導を頂きました．略儀ではありますが，本論文の書中により心からの感謝の意を記します．

修士入学から博士後期課程の修了に至るまで，懇切丁寧なご指導を賜りました九州工業大学大学院生命体工学研究科 脳情報専攻 神経情報処理講座運動制御機構 教授　石井 和男　先生に深く感謝致します．石井先生にはロボカップ中型リーグチーム“Hibikino - Musashi”の活動を通して学部時代から 6 年以上に渡り，ご指導を頂きました．COE マルチタレント英才教育のカリキュラムにおいては，短期間でロボット工学の基礎を学べるカリキュラムや課題を与えて下さり，2007 年度の世界大会・テクニカルチャレンジ部門において世界一位を獲得する経験を与えて下さいました．ご多忙を極めるスケジュールの中でも，毎年，ロボカップの日本大会や世界大会へ必ずご同行下さり，親身に私たち学生へのご指導等を頂きました．勉学や研究だけでなく，あらゆる物事へ挑戦し，諦めずに取り組む姿勢は石井先生に鍛えて頂きました．また，“Hibikino - Musashi”のチームリーダや研究室内の様々な加工機の運営，研究室の広報活動など，責任ある様々な機会に対し，私を信頼し，任せて下さったことは私に大きな自信を与えて下さいました．*“どんなに極限の状態でも，いかに冷静に，最大のパフォーマンスを発揮できるかが重要だ”*と経験を通して教えて下さったことは，今後，社会で生きていく上で絶大な助けとなると確信しています．修士時代には，反抗的な態度や物言いから，大変なご迷惑をお掛けしたにも関わらず，最後まで，親身にご支援とご指導を下さり，温かく接して頂いたことに，心から厚くお礼申し上げます．

また，平成 15 年から 19 年まで本校において開催されました COE マルチタレント英才教育へご推薦下さり，修士時代の 2 年間，研究者としての勉学修行を積ませて下さった九州工業大学大学院生命体工学研究科 脳情報専攻 特別推進研究 特任教授 山川 烈 先生へ心から感謝致します．学部時代の空手道や COE スチューデントカリキュラムを通して，精神的な強さや研究者としての心得など stoic に研究に取り組む姿勢をご指導頂けたことで，人間としての強さを得られたと感じております．書面にて大変恐縮ですが，改めてお礼申し上げます．COE マルチタレント英才教育では，ニューラルネットワークや強化学習など，機械学習について懇切丁寧なご指導を頂きました，九州工業大学大学院生命体工学研究科 脳情報専攻 高次脳機能講座 教授（当時）を務めておられました石川 真澄 先生，ヒトの視覚構造や心理実験手法，画像処理技術，また人生相談などにも親身にご指導を賜りました九州工業大学大学院生命体工学研究科 脳情報専攻 高次脳機能講座 准教授　花沢 明俊　先生へ深く感謝致します．

九州工業大学大学院生命体工学研究科 脳情報専攻 脳型情報処理機械講座 教授 神酒 勤先生，同じく脳情報専攻 脳型情報処理機械講座 准教授の宮本 弘之 先生より，ご多用の中，お時間を頂き貴重なご意見やアドバイスを賜りました. 博士学生としての発表のノウハウや研究内容の紐解きなど，大変貴重なご指導を頂きましたことを改めて御礼申し上げます.

また，技術を全く持たないゼロの状態から 7 年間に渡って，私を技術者として育てて下さった九州工業大学大学院生命体工学研究科 脳情報専攻 石井研究室研究員の Amir Ali Forough Nassiraei 氏，そして国際環境工学研究科情報工学専攻コンピュータシステムコース 教授 Ivan Godler 先生へ心から感謝致します．Amir 氏と Godler 先生とは，ロボカップ中型リーグでの活動を通して，公私ともにお世話になりました．熱心なご指導に対し，幾度となく激しい口論となり，大変失礼な発言や態度をとってしまったことを今も深く反省し，後悔しております．すぐに熱くなる私に対しても，いつも変わらず温かく接して下さった Amir 氏と Godler 先生は，私にとって師匠のような存在です. 数々の非礼を伏してお詫びすると供に，寛大に接して下さった御恩に深く御礼申し上げます.

石井研究室で過ごした 5 年間では，多くの先輩方や後輩に支えられ生活を送ることが出来ました．現在では日本文理大学 助教としてご活躍されている武村 泰範 先生には研究計画書や論文の書き方，後輩の指導について，武村先生がご卒業された後も親身に相談に乗って頂きました．ロボカップの活動をしながら COE スチューデントとしての勉学に取り組まれ，人一倍大変な環境で学生生活を送られた経験から，同じ COE スチューデントである私をいつも気遣って下さったことに，今もとても感謝しております. “Hibikino - Musashi”の最も過酷な時代を供に過ごし，多くの感動を経験出来たことを光栄に思っております．研究室の先輩であり，COE スチューデントの同期でもある石井研究室 研究員の西田 祐也 氏には，毎日のように励まして頂きました．西田さんは何事にも真剣に挑み，最後まで諦めない姿勢を間近で見せて頂きました．人生の先輩として，頼れる先輩としてこれからもご迷惑をおかけしますが，何とぞ宜しくお願い申し上げます．また，“Hibikino - Musashi”の活動を通して電気回路や回路設計に関するノウハウを伝授して下さった眞田 篤 氏にも心から感謝致します. いつも明るく接して下さり，楽しい話題を下さる一方で，いざ電気系の問題が発生した時は迅速かつ的確な修理と対策を施して下さった眞田さんは技術者として憧れでした．西田さんと眞田さんは，私にとって頼れる兄貴のような存在で最後までご迷惑をおかけしてしまいました．本当にありがとうございました.

また既に石井研究室をご卒業された，多くの先輩方にも言い尽くせないほどのご指導やご鞭撻を頂きました．特に西田周平さん，佐藤雅紀さん，松尾貴之さん，園田隆さんは，ご卒業後も石井研究室へご来室頂いた際や学会でお会いした際に私の進路を案じ，様々なご助言を賜りました．略儀にて大変恐縮ですが，心からの感謝の意を表します.

強化学習の研究を通してプログラムの開発や実験に熱心に取り組んでくれた山田 浩太 氏, 高木 正一 氏，安全保護方策案やリスクアセスメントについて熱心に検証を行ってくれた小川 優氏，福永 雄一郎 氏にも心から感謝致します．最後の一年間を供に過ごした藤本 和孝 氏，福田 一貴 氏にはいつも元気付けられました．研究に熱心に取り組む一方で，生活を楽しむ事にも一生懸命な二人の生き方はとても励みになりました．研究に疲れた時，楽しい話題で何度勇気付けられたか計り知れません．本当にありがとうございました．また“Hibikino – Muashi”のリーダの後任を引き受け, よきリーダとしてチームをまとめてくれた石田 秀一 氏にも心から感謝致します．頼れる先輩ではなかったと思いますが，いつも相談に来て頼ってくれたことに深く感謝致します．さらに学生生活では多くの同期達に支えられました．特に，久保 和範 君，小川 優 君，荒木 聡史 君は，いつでも相談に乗ってくれる本当に頼れる存在でした．付き合いの悪い私の我儘に今も付き合ってくれている親友たちに，心から感謝を申し上げます．

書面にお名前を述べることが叶わなかった諸先輩方や同期・友人達, そして多くの後輩達へも，略儀により大変恐縮ですが心からの感謝の意を表します．

最後になりますが 28 年間私を育て，見守ってくれた両親の北住 哲雄，北住 美穂子，姉の中田 馨，祖父母の酒井 菊次郎，酒井 静江に心からの感謝の意を表します．また，癌と闘いながら最後まで私の進学と将来を案じ, 私の大学合格を, 涙を流して喜んでくれた故 酒井 佐知子 氏にも心からの感謝を表します．

平成 23 年 12 月

北住 祐一

参考文献

参考文献

[1] NEDO 生活支援ロボット実用化プロジェクト
http://www.nedo.go.jp/activities/portal/p09009.html

[2] jon’s feed, “世界最初の知能ロボット Shakey”, http://www.johf.com/logs/20070419b.html

[3] ジョージ A. ベーキー著, 松田晃一, 細部博史訳, “自律ロボット概論”, pp.92-96, 株式会社毎日コミュニケーションズ, 2007

[4] Genghis robot, “A-history-on-robotics-Copy_1”,
http://www.dipity.com/michaelbvk/A-history-on-robotics-Copy_1/?mode=fs#!

[5] NAVLAB, Carnegie Mellon University Navigation Laboratory,
http://www.frc.ri.cmu.edu/robots/index.php

[6] P2, HONDA, http://www.honda.co.jp/robot/about/spec/

[7] Roomba, iRobot, http://www.irobot-jp.com/

[8] BigDog, engadget”蹴られても滑っても立ち直る四脚ロボ bigdog”,
http://japanese.engadget.com/2008/03/19/bigdog-robo-video/

[9] ASIMO, Tech-On!”時速 6km で走るホンダの新型 ASIMO—「ご案内とお茶出しも出来ます」”
http://techon.nikkeibp.co.jp/article/NEWS/20051213/111592/?SS=imgview&FD=-1759263373

[10] 吉本潤一郎, 銅谷賢治, 石井信, 強化学習の基礎理論と応用, 計測と制御, 第 44 巻 第 5 号, 2005, pp313-318

[11] 松原仁ら, “ロボカップの歴史と 2002 年の展望”, 日本ロボット学会誌, Vol.20, No.1, pp.2-6, 2002

[12] H. Kitano, et al., “robocup: A Challenge problem of ai”, AI magazine, Vol.18, No.1, pp.73-85, 1997

[13] 北野広明, 浅田稔, “「ワールドカップ」ロボットの挑戦”, 日経サイエンス, Vol.28, pp.74-82, 1998

[14] 松原仁, “ロボカップと地方自治体-何のためにロボカップを開くか-” The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 2E3-04, 2005

[15] 中村恭之, 高橋泰岳, NPO ロボカップ日本委員会監修, “中型ロボットの基礎技術 -対戦のための協調行動に向けて-”, 共立出版社, 2005

[16] Sérgio Monteiro, Fernando Ribeiro, Paulo Garrido, “Problems, Solutions and Trends in Middle-Size

Robot Soccer – A review," Robotica'2001 - Festival Nacional de Robótica, 25-28 April 2001, Guimarães, Portugal.

[17] K. Demura, K. Miwa, et al., "Matto: Towards a Pass – Based Tactics. ," RoboCup-99 Team Descriptions, Middle Robots League, 163-170, 1999

[18] D. Nardi, G. Adorni, et al. "Azzurra Robot Team – ART." RoboCup-99 Team Descriptions, Middle Robots League, 99-106, 1999

[19] A. Bredenfeld, T. Christaller, et al. "Behavior Engineering with "dual-dynamics" models and design tools." RoboCup-99 Team Descriptions, Middle Robots League, 135-145, 1999

[20] H. Fujii, D. Sakaki, K. Yoshida, "Cooperative Control Method Using Evaluation Information on Objective Achievement" 7th International Symposium on Distributed Autonomous Robotic Systems(DARS2004), Toulouse, pp.201-210, 2004-6

[21] "Minimax Value Iteration applied to Robotic Soccer", Gonçalo Neto, Pedro Lima, IEEE ICRA 2005 Workshop on Cooperative Robotics, Barcelona, April, 2005

[22] "1. RFC Stuttgart Overview of Hardware and Software", H. Rajaie, U.-P. Kappler, et al. RoboCup 2010 Team Descriptions, Middle Size Robot League, 2010

[23] "Tech United Eindhoven Team Description 2010", J.J.T.H. de Best, D.J.H.Bruijnen, et al. RoboCup 2010 Team Descriptions, Middle Size Robot League, 2010

[24] 湯軍，渡邊ら，"直行車輪機構を用いた全方位移動ロボット車の自律制御"，日本ロボット学会誌，Vol.17，No.1，pp.51-60, 1999

[25] Amir A. F. Nassiraei，Y. Takemura，et al.，"Concept of Mechatronics Modular Design for an Autonomous Mobile Soccer Robot"，CIRA2007，pp.178-183，Jacksonville，2007.

[26] Amir A. F. Nassiraei, "Concept of Intelligent Mechanical Design for Autonomous Mobile Robots", Journal of Bionic Engineeringm Vil.4, No.4, pp.281-289, 2007

[27] Y. Takemura, Amir A.F. Nassiraei，et al.，"Hibikino-Musashi Team Description Paper"， RoboCup 2007 in Atlanta，2007.

[28] Amir A. F. Nassiraei，K. Ishii，"How does "Intelligent Mechanical Design Concept" Help Us to Enhance Robot's Function?"，Theory and Applications，Vol.192，pp.155-178，Springer，2009.

[29] Amir A. F. Nassiraei, "Concept of Intelligent mechanical Design for Autonomous Mobile Robots", Kyushu Institute of Technology Ph.D dissertation, 2007

[30] Amir A.F. Nassiraei，Y. Kitazumi，and et al.，"Hibikino-Musashi Team Description Paper",

RoboCup 2010 in Singapore，2010.

[31] Amir A.F. Nassiraei，S. Ishida，and et al. , “Hibikino-Musashi Team Description Paper”， RoboCup 2011 in Istanbul，2011.

[32] 上岡聖, ゴドレールイヴァン, “ロボカップ中型ロボットリーグ Goalie 制御アルゴリズム”, ロボティクスメカトロニクス講演会 2008 講演論文集, 2P1-J16, ROBOMEC2008, 長野, 2008

[33] 畦浦和人, ゴドレールイヴァン, “全方位カメラを用いたロボットサッカーにおける色抽出”, 日本ロボット学会学術講演会予稿集 CD-ROM，2006 年 9 月，岡山，1B25，2006.

[34] Y. Takemura，K. Ishii，“Development of the Color Constancy Vision Algorithms Using Bio-Inspired Information Processing”，SMC2009，2009，San Antonio，pp.1746-1751，2009

[35] 新福宜侑，宮本弘之，“ロボカップ中型リーグにおける SVM を用いた色認識パラメータの自動設定”, ロボティクスメカトロニクス講演会 2011 講演論文集, 2A2-E10, ROBOMEC2011, 岡山，2011

[36] A.Merke，S. Welker，and M. Riedmiller，“Line Based robot localization under natural light condition”, In ECAI 2004, Workshop on Agents in Dynamic and Real-Time Environments, Valencia, Spain, 2004

[37] 佐藤佑介，宮本弘之ら，“ロボカップサッカー中型リーグにおける白線情報を用いた自己位置推定法の検討”，ロボティクスメカトロニクス講演会 2008 講演論文集，2P1-J14，ROBOMEC2008，長野，2008

[38] 新福宜侑，ゴドレールイヴァンら，“ロボカップ中型リーグにおけるパーティクルフィルタを用いたロボットの自己位置推定”, ロボティクスメカトロニクス講演会 2010 講演論文集, 1P1-F26，ROBOMEC2010，旭川，2010

[39] Yusuke Sato, Amir A.F. Nassiraei，et al. , “Hibikino-Musashi Team Description Paper 2008”, RoboCup 2008 in Suzhou，2008.

[40] F. Dellaert，D. Fox，W. Burgard，S. Thrun，“Monte Carlo Localization for Mobile Robots”，Proc. IEEE International Conference on Robotics and Autonomation，Vol.2，pp.1322-1328，1999.

[41] R. Tsuzaki, K. Yoshida，“Motion Control Based on Fuzzy Potential Method for Autonomous Mobile Robot with Omnidirectional Vision”，Journal of the Robotics Society of Japan，Vol.21，No.6，pp.656-662，2003.

[42] S. Thrun，D. Fox，et al.，“Robust Monte Carlo Localization for Mobile Robot”，Artificial Intelligence Journal，Vol.128，No.1-2，pp.99-141，2001.

[43] R. Ueda，T. Arai，et al.，“Recovery Methods for Fatal Estimation Errors on Monte Carlo Localization”, Journal of the Robotics Society of Japan, Vol.23, No.4, pp.466-473，2005.

[44] S. Lenser, et al.,“Sensor Resetting Localization for Poorly Modelled Robots”, Proc. of IEEE ICRA, pp.1225-1232，2000.

[45] M.Asada, et al., “Middle Size Robot League Rules and Regulations for 2011 – Version 15.0 20101207”, RoboCup 2011 in Istanbul

[46] 浅田 稔，北野 宏明，“ロボカップ戦略：研究プロジェクトとしての意義と価値”，日本ロボット学会誌，Vol.18(8)，pp.27-30，2000 年 11 月 15 日

[47] 中村恭之，“実機ロボットリーグの現状と今後の課題”，日本ロボット学会誌，Vol.20(1)，No.1，pp.11-14，2002

[48] 鈴木昭二，浅田稔，“学習によるロボットの行動獲得 -サッカーへの適用とロボカップへの取り組み-”

[49] 浅田稔，野田彰一ら，“視覚に基づく強化学習によるロボットの行動獲得”，日本ロボット学会誌，Vol.13(1)，pp.68-74，1995

[50] 内部英治，浅田稔ら，“視覚に基づく強化学習による移動ロボットの多重タスクの遂行のための協調行動の獲得”，第 21 回人工知能基礎論研究会，pp.25-32，1995

[51] 浅田稔，“強化学習の実ロボットへの応用とその課題”，人工知能学会，Vol.12，No.6，pp.831-836，1997

[52] 加藤龍憲，鈴木昭二ら，“強化学習によるゴール守備行動の獲得”，第 3 回 JSME ロボメカ，シンポジア講演論文，pp.37-40，1998

[53] 加藤龍憲，鈴木昭二ら，“複数の報酬による強化学習を用いたサッカーロボットのゴール守備行動の獲得”ロボティクスシンポジア予稿集，第 4 巻，pp.289-294，1990

[54] 出村公成，“簡単！実践！ロボットシミュレーション Open Dynamics Engine によるロボットプログラミング”，森北出版，2007 年 5 月

[55] Yuichi Kitazumi, Shuichi Ishida, et al., “Hibikino-Musashi Team Description”, RoboCup 2009 in GRAZ, 2009

[56] 三上貞芳，皆川雅章 共訳，“強化学習”，pp.159-161，森北出版，2000

[57] Y. Kitazumi，N. Shinpuku，“A Cooperative Self-Localization Method for Multi Agent System”, Proc. of ICIUS CD-ROM, MoAmC1-3，The International Conference on Intelligent Unmanned System 2011, Chiba, 2011

[58] 北住祐一，石井和男，“ロボカップロボットにおけるボール情報の共有による自己位置推定精度の向上”，第 27 回ファジィシステムシンポジウム講演論文集 CD-ROM，MD3-3，第 27 回ファジィシステムシンポジウム，福井，2011

[59] 北住祐一，石井和男，“群情報に基づいたランドマークの確率的位置推定と自己位置較正法に関する研究”，第 29 回日本ロボット学会学術講演会講演概要集，3I1-7，第 29 回日本ロボット学会学術講演会，芝浦，2011

[60] 北住祐一，石井和男，“ロボカップ中型リーグにおける協調行動アルゴリズムに関する基礎的検討”，第 28 回日本ロボット学会学術講演会概要集，AC2Q1-6，第 28 回日本ロボット学会学術講演会，名古屋，2010

[61] 北住祐一，石井和男，“郡内情報を用いた自己位置推定アルゴリズムの研究～ロボカップ中型リーグ用移動ロボットを用いた評価～”，第 23 回自律分散システムシンポジウム資料，PP.325-330，第 23 回自律分散システムシンポジウム，札幌，2011

[62] 大内 東，山本雅人，川村秀憲，“マルチエージェントシステムの基礎と応用 -複雑系工学の計算パラダイム-”，コロナ社，pp.4-17，2002

[63] 柴田聡志，神谷昭基，“強化学習の連続値への適用”，釧路工業高等専門学校紀要，2007 年

[64] 北住祐一，石川真澄，“動力学を考慮した Q 学習に関する研究”，九州工業大学平成 19 年度 COE マルチタレント英才教育レポート，2007

[65] K. Samejima, T.Omori, “Adaptive internal state space construction method for reinforcement learning of a real-world agent”, Elsevier, Neural Networks 12, pp.1143-1155, 1999

[66] 森本淳，銅谷賢治，“強化学習を用いた高次元連続状態空間における系列運動学習：起き上がり運動の獲得”，電子情報通信学会論文誌 D-II，Vol.J79-D-II，No.xx，pp.1-13，1996

[67] Sebastian Thrun, Wolfram Burgard, and Dieter Fox 著，上田隆一訳，“確率ロボティクス”，株式会社毎日コミュニケーションズ，2007 年

[68] 松本吉央，道木加絵，“「確率理論のロボティクス応用」特集について”コラム，日本ロボット学会誌，Vol.29，No.5，pp.1，2011

[69] 上田隆一，“確率ロボティクスの展望”，日本ロボット学会誌，Vol.29，No.5，pp.2-5，2011

[70] Wikipedia“パス（サッカー）”，URL:http://ja.wikipedia.org/wiki/パス_(サッカー)

[71] 川端 邦明，善林 正春ら，全方向移動ロボットによる協調サッカープレイ，ロボティクスメカトロニクス講演会1998講演論文集，1AIII4-4，ROBOMEC’98，仙台，1998

[72] 内部 英治，浅田 稔ら，共進化によるマルチ移動ロボット環境における協調行動の獲得，

北野宏明編，遺伝的アルゴリズム 4，第 7 章，pp.193-220，2000