インタラクティブ強化学習に基づく ソーシャルエージェントの協調行動の視覚化

Visualization of Cooperative Behavior of Social Agents
Based on Interactive Reinforcement Learning

○ 張　坤　　前田　陽一郎　　高橋　泰岳
○ Kun Zhang　Yoichiro Maeda　Yasutake Takahashi
福井大学大学院 工学研究科
Graduate School of Engineering, University of Fukui

Abstract: In the multi-agent reinforcement learning, multiple agents are supposed to learn cooperative behavior. One unsolved problem is how to obtain an appropriate cooperative behavior from the mutual interaction during the reinforcement learning process. In this research, we propose an interactive reinforcement learning system with the efficient cooperative ability through the social interaction among agents. It is effective to visualize the satate and agent behavior of learning process to analyze the interaction system. Therefore, we apply the visualizing tool, which can reflect the idea to the agent model immediately. The cooperative process changing dynamically is analyzed by indicating the mutual interaction among agents on the computer.

1　はじめに

強化学習は報酬を頼りに試行錯誤を繰り返し、未知の動的な環境に対応する適切な行動を獲得できる学習手法である [1]。また、マルチエージェントシステムでは、複数の自律ロボットが目標を達成することにより全体のタスクを達成する協調作業を行なう。複数のエージェントに協調動作を学習させるマルチエージェント強化学習の研究にも期待が寄せられている [2, 3, 4]。

しかしながら、マルチエージェント強化学習では、個々のエージェントがどのように相互に影響を与えれば、全体の適切な協調行動を獲得できるかという問題における決定的な解決方法は示されていない。協力作業では、センシング情報、エピソード、学習方策やアドバイスのやり取りなどを共有することで、協調行動を学ぶことが有効であることが知られている [5]。しかし、エージェントの自律性により、相手のエージェントに有効な行動が自身に有効であるとは限らない。そのため、相手の学習方策がどのような状況で、どの程度利用されるかなどを自律的に学習することが必要となる。

そこで本研究では、マルチエージェント間の強化値インタラクションを通じて、相互に学習できるインタラクティブ学習手法を提案する。インタラクションシステムを解析するため、有効性が一目でわかる視覚化ツールを利用し、エージェント同士の相互インタラクションをコンピュータ上で表示することで、ダイナミックに変化するエージェント集団の協調行動を分析する。

2　インタラクティブ強化学習システム

マルチエージェント強化学習では、同質エージェントはエピソードや学習方策を共有することで、協調行動を効率的に学ぶことができる。しかし、異質エージェントは相手に有効な方策を自身にそのまま適応できない場合が多い。そのため、他エージェントの有効な学習経験を間接的に活かせれば、マルチエージェント全体の協調性が高まるものと考えられる [6]。

本手法では、各エージェントが目標達成行動に携わりながら、それぞれのエージェントと一緒に目標を達成した時、環境から得た報酬により、相手との信頼度を構築する。その信頼度により、学習しなかった環境または学習経験がほとんどなかった環境になると、信頼度をもったエージェントの強化値を利用することができる。このように、エージェント間のインタラクションを利用して、相互協力を行う協調戦略を学習させる。

2.1　信頼度に基づく強化値インタラクション

本研究のインタラクティブ強化学習システムでは、各エージェントは自身の強化値のみではなく、知覚範囲内のエージェントの強化値を考慮し、目標となる適切な行動を選択する。そのため、試行錯誤でエージェント間には各グループへの信頼度を生成し、更新する。

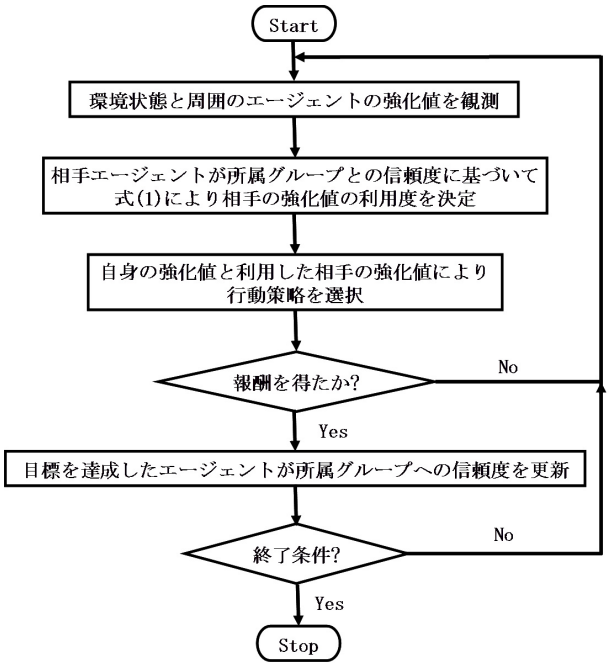

図 1: 信頼度に基づく強化値インタラクション

その信頼度に基づいた行動選択戦略獲得の流れを図 1 に示す。

ここでは能力の異なるエージェントに対し同じ能力のエージェント同士 (同質) を同じグループにする。グループ間には目標の達成度により、グループ信頼度を構築する。グループ信頼度により、知覚範囲内のエージェントの強化値をどの程度利用するかを判断する。毎回の学習結果に基づいて、グループ信頼度が更新され、このようなエージェント間のインタラクションを繰り返すことで、協調行動を自律的に学習する。

2.2 Q-learning を用いた強化値インタラクション

強化学習における実現方法は様々な手法が提案され、大きく分けて「環境同定型」と「経験強化型」の二つに分けることができる。信頼度に基づく環境同定型強化学習システムにおける Q-learning(QL) の処理手順を以下に示す。

- $Q_t(s_t, a_t)$ を初期化
- 各エピソードに対して以下を繰り返す
  - 環境状態 s_t を初期化
  - 各エピソードの各ステップに対して以下を繰り返す
    - s_t を観測
    - 式 (1) により利用する強化値 $Q_t^*(s_t, a_t)$ を計算
    - 式 (2) により方策 $\pi(s_t, a_t)$ を用いて選択行動 a_t を実行
    - 報酬 r_t と次状態 s_{t+1} を観測し、$Q_t(s_t, a_t)$ を以下の式で更新

$$Q_t(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha[r_t + \gamma \, max_{a \in A} \, Q_t(s_{t+1}, a) - Q_t(s_t, a_t)]$$

    - 式 (3) により、信頼度を更新
    - $s_t \leftarrow s_{t+1}$
- 終了条件を満たせば終了

本手法の Q 値の更新式およびボルツマン選択による行動選択式を以下に示す。

$$Q_t^*(s_t, a_t) = \sum_{o=1}^{n} (Q_t^o(s_t, a_t) \cdot \frac{C_t^o}{\sum_{i=1}^{m} C_t^i}) \qquad (1)$$

$$\pi(s_t, a_t) = \frac{\exp[(Q(s_t, a_t) + Q^*(s_t, a_t))/T]}{\sum_{b \in possibleactions} \exp[(Q(s_t, b_t) + Q^*(s_t, b_t))/T]} \qquad (2)$$

$$C_t^o = \sum_{i=1}^{t} \frac{r_i^o}{R^*} \qquad (3)$$

ここで、

s_t: 時刻 t における状態
a_t: 時刻 t における行動
$Q_t^*(s_t, a_t)$: 利用する強化値
n: 知覚範囲内のエージェント数
T: 温度定数
C_t^o: 時刻 t におけるエージェント o の所属グループの信頼度
$Q_t^o(S_t, a_t)$: 時刻 t におけるエージェント o の強化値
r_i^o: エージェント o との協調行動で得た報酬
R^*: すべてのエージェントが得た平均報酬
α: 学習率
γ: 割引率

学習前にはすべてのエージェントグループ間には信頼度が存在しない。目標を達成した報酬を得たとき、携わったエージェントの所属グループ間には式 (3) のようなグループ信頼度が生成される。グループ信頼度

を構築することで、自身に適応できる戦略のみを利用することが可能になる。

2.3 Profit Sharing を用いた強化値インタラクション

Profit Sharing 法 (PS) は Q-learning のように学習解に最適性が保障されないが、学習にかかる時間が短いという特徴がある。信頼度に基づく強化値のインタラクティブ学習システムでは、信頼度 C_t^o の計算は式 (3) と同様である。他のエージェントの行動評価値 $w_t^*(s_t, a_t)$ は式 (4) のようになる。行動を選択する際には式 (5) のようなルーレット選択で行動を決定させる。その他の処理は Q-learning の場合と同様である。

$$w_t^*(s_t, a_t) = \sum_{o=1}^{n} w_t^o(s_t, a_t) \cdot \frac{C_t^o}{\sum_{i=1}^{m} C_t^i} \tag{4}$$

$$P_{selection} = \frac{w_t(s_t, a_t) + w_t^*(s_t, a_t)}{\sum_{b \in possibleactions} (w_t(s_t, b_t) + w_t^*(s_t, b_t))} \tag{5}$$

ここで、

$w_t(s_t, a_t)$: 自身の行動評価値
$w_t^o(s_t, a_t)$: エージェント o の行動評価値
$w_t^*(s_t, a_t)$: 利用する行動評価値
$P_{selection}$: 行動 a_t が選択される確率

3 シミュレーション実験

提案したインタラクティブ学習手法を検証するために、ダイナミックに変化する現象をリアルタイムで分析できるマルチエージェント・シミュレータ (artisoc)[7] を用いて獲物追跡問題のシミュレーション実験を行なった。シミュレータの概観を図 2 に示す。このシミュレータは試行スペースでのエージェントの動きを表示しながら、各獲物エージェントが捕獲された時間、各グループ (A~F) が得た報酬量と各グループ間の信頼度関係などの実験結果を同時に出力することができる。

3.1 シミュレーション条件

このシミュレーションでの学習環境は 2 次元格子状で、100 × 100 のグリッド空間を考える。ここでは 6 種類 (A~F) のハンターエージェント (シミュレータ上では小さい四角形) グループと 3 種類 (1~3) の獲物エージェント (シミュレータ上では大きい三角形) グループが存在する。各グループのエージェントは 10 体おり、すべての設定は同じである。ハンターエージェントの目標行動はできるだけ早くすべての獲物エージェントを捕獲することである。

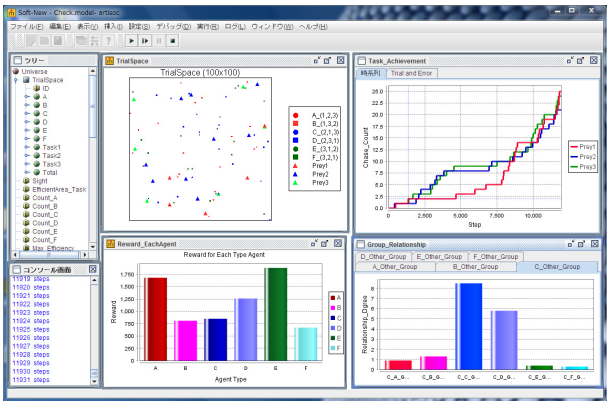

図 2: シミュレータの概観

表 1: エージェントグループの捕獲能力 (単位: N)

グループ	獲物 1	獲物 2	獲物 3
A	0.6	0.3	0.1
B	0.6	0.1	0.3
C	0.3	0.6	0.1
D	0.1	0.6	0.3
E	0.3	0.1	0.6
F	0.1	0.3	0.6

いろいろな特性をもつ異質エージェントを設定し、自身の特性に似たエージェントが多くの経験を学習できるかどうかを検証するため、各ハンターエージェントの獲物を捕獲する能力を表 1 のように異なったものに設定した。自らの能力では捕獲行動を達成できず、獲物エージェントの周囲におけるすべてのハンターエージェントの合計能力が 1N に達すると、単体獲物の目標捕獲達成となる。すべての獲物エージェントが捕獲されたとき、1 回の学習試行とする。そのため、エージェント間で協調作業を学習することが必要である。

すべてのハンターエージェントの視野は 10 × 10 のフィールド範囲とした。視野範囲内のエージェント間では強化値を利用することができるものとする。ハンターエージェント、獲物エージェントは共に上、下、左、右、左上、左下、右上、右下の 8 方向のいずれか 1 マスずつ動くことができる。また獲物エージェントはいつもランダムに動くものとする。

基本報酬 R を 10 とし、目標捕獲を達成すると、各ハンターエージェントが得られる報酬は表 1 の自らの能力に基本報酬を乗じた値とする。本実験では強化学習

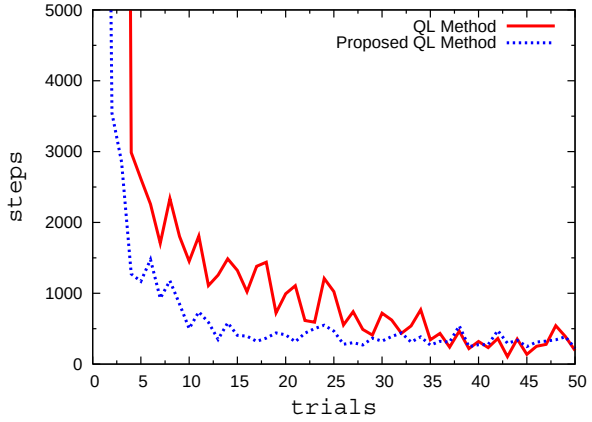

図 3: Q-learning を用いた学習結果

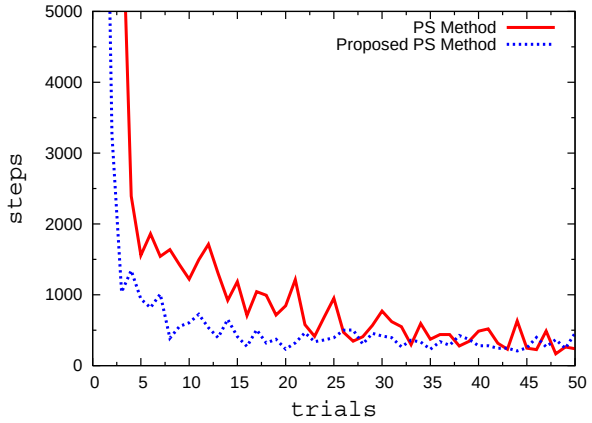

図 4: Profit Sharing を用いた学習結果

のパラメータを $\alpha = 0.8$、$\gamma = 0.8$、T=0.6 とした。提案手法の有効性を検証するため、通常の強化学習 (QL と PS) との比較実験を行なった。

3.2 シミュレーション結果および考察

すべての手法は 5 回ずつシミュレーションを行い、シミュレーション結果の平均と分散を表したグラフを図 3 と図 4 に示す。横軸は試行回数で、縦軸はすべての獲物を捕獲したステップ数を表す。比較実験結果により、提案手法は QL と PS の両方共に短い時間で収束し 50 回の試行学習を終了したことが分かった。時間的にも効率的に獲物を捕獲する協調行動を取得できたことが検証された。

また、本実験では表 1 のような 6 種類のハンターエージェントが設定され、各グループのハンターエージェントは獲物エージェントグループに対する能力が異なるため、それぞれのグループ間の信頼度は学習前にはゼロであったが、エージェント間の信頼度はリアルタイムでダイナミックに変化する。グループの学習した後の信頼度関係は、図 5(QL) と図 6(PS) のようになった。横軸は 6 種類のハンターエージェントのグループを示し、縦軸はグループ間の信頼度の大きさを表している。

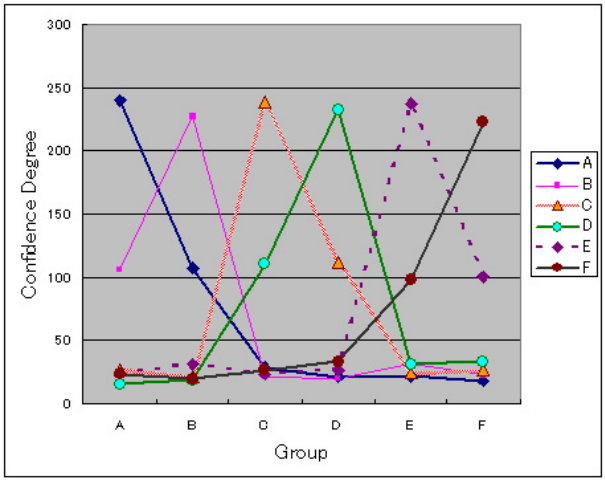

図 5: 各グループ間の信頼度関係 (Q-learning)

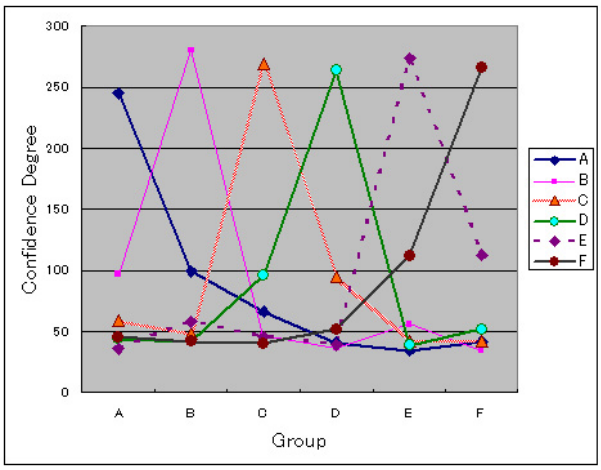

図 6: 各グループ間の信頼度関係 (Profit Sharing)

比較結果より、それぞれのグループは同質で、自身が所属するグループへの信頼度が一番高いことが分かった。表 1 によると、A と B グループ、C と D グループ、E と F グループは比較的特徴が似ており、これらはペアとなる二番目の信頼度が確立された。

インタラクションシステムを解析するためには、学習途中の状況をリアルタイムでコンピュータ上で表示することで、ダイナミックに変化するエージェント集団の協調度合いを分析できる。例として、通常手法による学習初期の協調行動を図 7 に示す。横軸は試行回

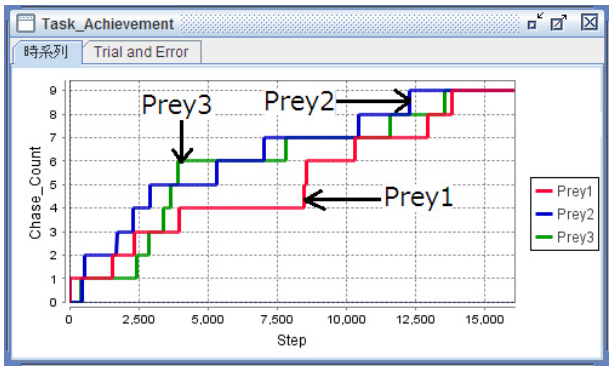

図 7: 通常手法の学習初期における協調行動の例

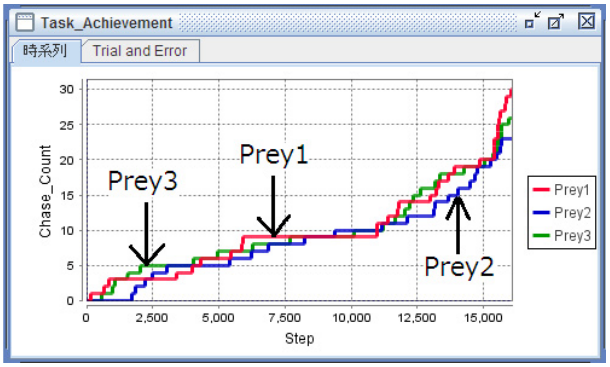

図 8: 提案手法の協調行動の例

数を示し、縦軸は各種類の獲物が捕獲される数を示す。このようなグラフは実際のステップ実行に伴い、獲物の捕獲状況がリアルタイムで表示される。10 台ずつの獲物 1~3 がすべて捕獲されると、一回の試行学習終了となる。この図によると、獲物 2 と 3 が早めに捕獲されたが、獲物 1 は遅くて捕獲されたことが分かった。つまり、多くのハンターエージェントは獲物 2 と 3 の捕獲協調行動に携わったが、獲物 1 の捕獲協調行動に時間がかかった。

提案手法の協調行動は図 8 のように、各獲物はほとんど同時に捕獲され、優れた協調行動を取得したことが分かった。また、横軸の時間とともに曲線の傾斜が徐々に高くなっており、捕獲時間が減少していることがわかる。

各ハンターエージェントの特性を分析するため、取得した報酬により、図 9 に各ハンターエージェントの協調度合いを示す。報酬量により、6 種類のハンターエージェントがグループとして、獲物を捕獲した協調行動の性能を把握することができる。

エージェント間の協調行動の様子を見るため、獲物を捕獲するまでの 30 ステップの軌跡を記録した。最

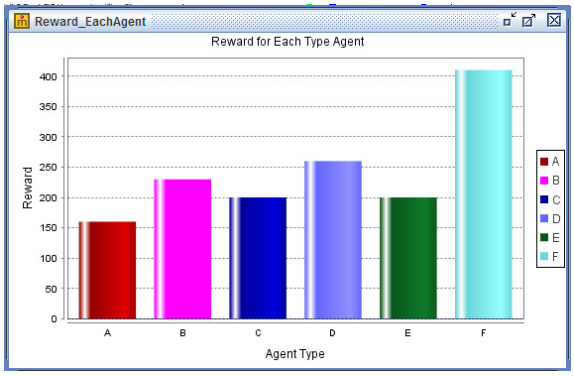

図 9: 各ハンターエージェントの取得した報酬

初の獲物を捕獲した様子を図 10 に示す。横軸と縦軸は座標で、三角、丸、四角で表されるエージェント位置は最後に捕獲された状態を示し、引かれている線はそれぞれのエージェントの軌跡を示す。最初の獲物を捕獲した時、ハンターエージェントは経験がなく、単純にランダムな行動で捕獲タスクを達成したため、エージェント間の協調の様子はほとんど見られなかった。

さらに学習後期には図 11 のような軌跡となり、ハンター A5、B4、D7 の 3 エージェントは遠いところから Prey1 という獲物エージェントに向かって効率よく協調して追跡していたことを示している。図 10 と図 11 を比較すると、ソーシャルエージェントの協調行動が獲得されたことが確認できる。

4 結 言

本研究ではエージェント間の強化値インタラクションを通じて、優れた協調能力をもつインタラクティブ強化学習システムの構築手法を提案した。学習経験がほとんどなかった環境でも、信頼度をもったエージェントの強化値を相互利用することで効率的な協調行動の獲得が可能になることがわかった。

Q-learning と Profit Sharing 強化学習を用いて、獲物追跡問題を例題にシミュレーション実験を行った。実験結果により、提案手法は通常手法より効率的に協調行動を学習により取得した。異質エージェントとして、各グループの特徴にあわせて、自身に有効な戦略のみを学習することで、良い集団戦略を取得できた。またそれぞれのエージェント間の信頼度生成の過程を視覚化して、協調関係を分かりやすく確認することもできた。

今後の課題として、相互強化値の学習のみではなく、より優れたコミュニケーションモデルの構造を構築す

ることで、相互学習機能をもったソーシャルエージェント学習システムへの発展が考えられる。

参考文献

[1] R.S.Sutton and A.G.Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.

[2] 宮崎 和光, 木村 元, 小林 重信, “ 特集「計算学習理論の進展と応用可能性」ProfitSharing に基づく強化学習の理論と応用, ”人工知能学会誌, Vol.14, No.5, pp.40-47, 1999.

[3] U.Hu and M.P.Wellman,“ Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm, ”*Proc. of International Conf. on Machine Learning (ICML-98),* pp.242-250, 1998.

[4] 荒井 幸代, “ マルチエージェント強化学習-実用化に向けての課題・理論・諸技術との融合-, ”人工知能学会誌, Vol.16, No.4, pp.476-481, 2001.

[5] M.Tan, “ Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents, ”*Proc. of the 10th International Conf. on Machine Learning,* pp.330-337, 1993.

[6] K.Zhang, Y.Maeda, Y.Takahashi, “ Interactive Learning of Social Agents Based on Confidence Degree, ”WCCI 2012 IEEE World Congress on Computational Intelligence, F-177, pp.1239-1242, 2012.

[7] MAS コミュニティ, http://mas.kke.co.jp/modules/tinyd0/index.php?id=9

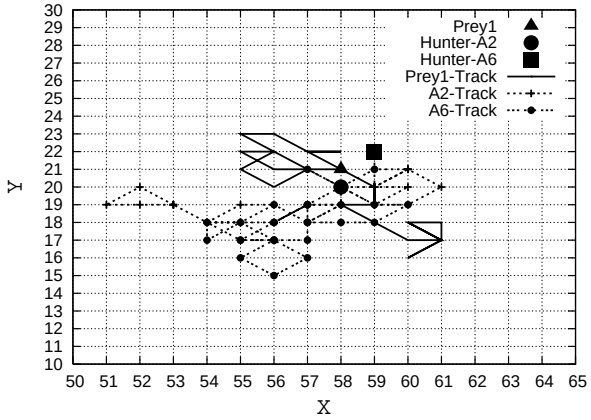

図 10: 最初の獲物を捕獲した軌跡の例

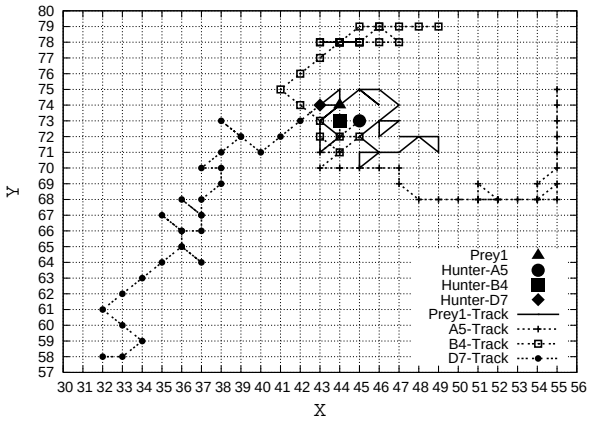

図 11: 学習後期に獲物を捕獲した軌跡の例

連絡先

〒 910-8507　福井県福井市文京 3-9-1
福井大学大学院 工学研究科 システム設計工学専攻
張 坤 (進化ロボット研究室)
Tel & Fax: 0776-27-8050
E-mail: kzhang@ir.his.u-fukui.ac.jp