ロボティクス・メカトロニクス講演会 2017 in Fukushima

谷 口 貴 一
Kiichi TANIGUCHI
機械システム工学専攻修士課程　2 年

1. はじめに

私は 2017 年 5 月 10 日から 13 日に福島県のビックパレットふくしまで行われた「ロボティクス・メカトロニクス講演会 2017」に参加した．12 日のポスター講演で「強化学習に基づく 4 脚走行ロボットの目標地点到達のための歩容生成」について研究発表した．

2. 研究内容

2.1 緒言

ロボットを制御する際に単純な制御では不測の事態や未知の環境へ完全に対応することはできないという課題がある．この問題を解決する手段の 1 つとして，機械学習法の 1 つである強化学習が挙げられる．この学習法を制御に応用することで，ロボット自らその環境における目標達成に適した行動を学習・選択可能になる．

本研究では 4 足動物をモデルとしたロボットを用い，目標地点到達というタスクに対して強化学習を適用し，学習に用いる状態や報酬の違いによる学習効果への影響を調べた．

2.2 4 脚走行ロボット

本研究では，4 足動物を模した 4 脚走行ロボット（図 1）をモデル化し，それをコンピュータ上に再現して学習実験を行う．ロボットの諸元を表 1 に示す．

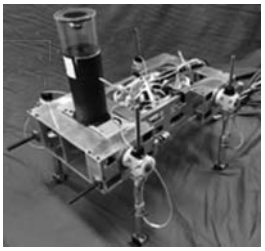

図 1　Four-Legged Robot

表 1　Details of the Robot

Length［mm］	550
Width［mm］	414
Leg Length［mm］	250
Weight［kg］	11.8

2.3 歩容

動物は脚を一定の法則に基づいて動かし，それを繰り返すことで歩行や走行などの移動を行う．この動作のことを歩容という．本研究では 4 足動物が走行に用いるバウンド歩容及びトロット歩容を，進行方向が調整できるように人工的に拡張した歩容を用いて実験した．バウンド歩容は脚毎に脚ふり角度を決定することで左右に曲がることが可能な歩容として使用する．トロット歩容はその場で跳躍しながら左及び右に旋回する歩容として使用する．

2.4 強化学習

本研究では，学習手法として Q 学習，学習方策として ε-greedy 法を使用し，学習を行う．

2.4.1 本研究における Q 学習

本研究ではロボットに 2.5 節で示す条件を達成した場合に「成功」として正の報酬を渡し，「転倒」，「制限時間（30 s）超過」，「場外（目標地点から 6.0 m 以上離れる）」のいずれかが発生した場合に「失敗」として負の報酬を渡し，動作中の行動を評価した．また，$\alpha = 0.1$, $\gamma = 0.95$ で学習を行った．

2.4.2 本研究における ε-greedy 法

本研究では $\varepsilon = 0.1$ で学習を行った．この方策を用いることで，Q 値の高い行動のみを選択する場合と比べて多くの種類の行動を選択可能となり，探査がより広く行えることである．もう 1 つはランダムに行動を選択する場合と比べて学習中も比較的高い報酬を得やすく動作の獲得・改善に必要な行動の習得が速いことである．

2.5 学習実験

4 脚走行ロボットに強化学習を適用する際に，目標地点到達というタスクに対して学習に用いる状態

や報酬の違いが学習効果へ及ぼす影響を調べた．

2.5.1　実験・評価概要

物理演算に「Open Dynamics Engine」を用いた自作シミュレータ上で学習を行った．目標地点は機体初期位置から 5 m 以内にランダム生成した．Q 値の更新は 1 周期毎に行い，用いた状態は 2.5.2 節や 2.5.3 節に記す．行動として 2.3 節で記した 2 種の歩容，0.26 s～0.28 s の 3 種の脚ふり周期，12 deg～14 deg の 3 種の脚ふり角度，35 mm⇔－14 mm またはその半分の 2 種の脚伸縮量を使用した．

2.5.2　報酬の与え方の検討実験

状態としてゴールとの距離＋進行方向と機体－ゴール直線の角度を用い，以下の①～④に示す 4 パターンの報酬を与え，それぞれ 300 万試行学習した．

①ゴール到達

②ゴール到達＋ゴールとの距離が減少

③ゴール到達＋進行方向と機体－ゴール直線の角度減少

④ゴール到達＋ゴールとの距離が減少＋進行方向と機体－ゴール直線の角度減少

学習中適宜 Q 値を保存し，学習効果の評価として 8 方向×6 種の距離＝48 パターンの目標地点に対してロボットを動作させ，到達までにかかった時間の総計をテスト結果として記録した．なお，到達不能な場合は制限時間である 30 s かかったものとした．

テスト結果を図 2 に示す．④が最も学習効果が高く，到達時間の総計が最短のものは 255 s となった．また，①は他のパターンを比べて学習効果が顕著に現れるタイミングが最も遅く，最終的なテスト時の到達時間の総計が最長であることから，ゴール到達時という複数試行後に報酬を与える長期的な報酬の与え方だけでなく，1 試行毎に報酬を与える短期的な報酬の与え方もした方が良いと考えられる．

2.5.3　状態の違いによる比較実験

2.5.2 節で最も学習効果の高かった④の報酬付与パターンで，⑤～⑦に示す 3 パターンの状態を用いて各 1000 万試行学習した．学習効果の評価は 2.5.2 節と同様のテストを行った．

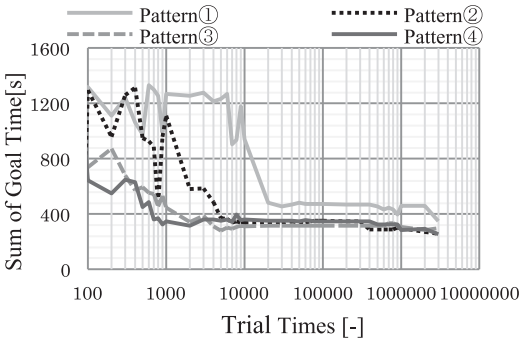

図 2　Sum of Goal Time in Test（Pattern①～④）

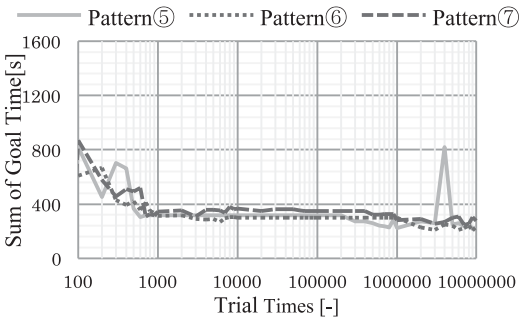

図 3　Sum of Goal Time in Test（Pattern⑤～⑦）

⑤ゴールとの距離＋進行方向と機体－ゴール直線の角度

⑥ゴールとの距離＋進行方向と機体－ゴール直線の角度＋脚毎の前周期のデューティー比

⑦ゴールとの距離＋進行方向と機体－ゴール直線の角度＋脚毎の前周期での角度のずれ

テストの結果を図 3 に示す．⑥が最も学習効果が高く，到達時間の総計は最短 205 s となった．また，2.5.2 節のような大きな差はないが，⑤のみ途中で到達時間の総計に大きなブレが出ていることを考慮すると，機体本体に関する情報も状態として与えた方が良いと考えられる．

3. まとめ

ポスター発表ではポスター以外にも動画を再生し，ロボットの動きや学習の様子をより分かりやすく伝えることができた．自分の研究に対する様々な意見や他の参加者・企業の研究を知ることができ，貴重な体験となった．