1D4-04

多自由度環状ロボットによる不整地走行の学習

Learning Motion over Rough Terrain by Circularity Robot with Multiple Degree of Freedom.

藤野智宏[*1] 佐久間淳[*1] 小林重信[*1]
Tomohiro FUJINO Jun SAKUMA Shigenobu KOBAYASHI

[*1]東京工業大学 総合理工学研究科
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

Recently, the robot that can execute various actions autonomously over the rough terrain environment is demanded. The snake robot and the module robot is known as a robot that can make good such a purpose. However, in these robots, deformation of a robot body and the movement are often decided in a specific environment. Therefore, the control can be used only in a specific environment. In this paper, it aims at adaptive learning under an unknown environment that has rough terrain and the obstacle for the circularity robot composed of the multi link.

1. はじめに

近年，ロボットには様々な分野や環境での行動が求められるようになってきている．これらのロボットが活躍する環境は凹凸や障害物が存在するような不整地環境であり，そのような環境下では，ロボットの形状を柔軟に変形させ地面の形状に合わせるなどといったように，状況に合わせて行動できることが望ましいと考えられる．

モジュール型ロボット [1] は形状を柔軟に変形するロボットの一つであり，注目されている．しかし，パーツ同士を切断・結合して形状変形する構造であるため，目的の形状となるまでに一定の時間を必要とし，地面を捉えるような滑らかな形状変形ができないといった問題がある．また制御においては，環境とのインタラクションを行わずに制御する方法であるため，各環境に適した形状へ変形し動作が行われているとはいえない．構造は，短い時間で環境にフィットでき地形を捉えることができるような柔軟性をもつ必要がある．制御は，環境とのインタラクションを行うことで環境を判断し，状況に適した動作を行うような方法が望ましい．

多自由度環状ロボットは，各剛体リンクをジョイントで結合し環状とした多リンクロボットである．そのため，地形にフィットするような滑らかな形状変形をすることができる構造であり，不整地環境内の，様々な状況に合わせて行動できる．しかし，このモデルの制御は従来の人によって作り込まれた古典的制御方法で制御することは困難である．

強化学習はエージェントと環境のインタラクションによって，環境に適応した政策 (状態から行動への写像) を学習できる枠組みである．よって各環境に対して最適な形状変形および動作を獲得することが期待される．特に，強化学習手法の中で確率的傾斜法 [3] をはじめとする政策勾配法は連続状態行動空間を取り扱うことができる．このため，ロボット等の実問題に対する適応的な制御手法として注目されており，様々な実問題に対してその有用性が確認されつつある．そのため強化学習でのアプローチが有効である．しかしながら，未知の環境に対して初めて適用する際には各種センサや報酬設計を行う必要がある．

本研究では，多自由度環状ロボットを対象に，未知の不整地環境下において，ロボット自身が環境の変化を判断し，環境に適した形状変形・走行動作を行う適応的学習の構築を最終目的とする．その実現の為には，未知環境に対してユニバーサルな報酬設計が必要となる．そのために多様な環境において実験を行ない提案する報酬設計が有効であるかどうかを検証する．

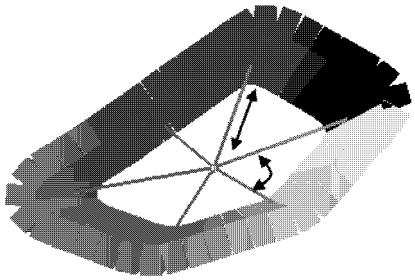

Fig. 1: 多自由度環状ロボット

以下，2 章では問題設定として対象とするロボットについて説明し，適用する各環境を示す．3 章では目指す動作を得るための接近法と報酬設計について述べる．4 章において各環境に対して移動動作獲得実験を行い，報酬設計の有効性を示す．5 章においてまとめと考察を述べる．

2. 問題設定

本章ではモデルについて述べ，次にそのモデルを適用する不整地環境を示す．その後，強化学習の定式化を行なう．

2.1 多自由度環状ロボット

対象ロボットは，40 個の各剛体リンクをフリージョイントで環状に結合した 11 自由度ロボットである．このようなロボットモデルとした理由は，滑らかに形状を変形でき，地面や障害物にフィットした形状にもなれるようにするためである．

学習環境は動力学シミュレータ Vortex 上に実現されている．Fig.1 に開発した多自由度環状ロボットを示す．各剛体リンクは隣同士でジョイントによって結合しているが，ゴム素材を想定して各リンク同士の衝突は柔らかい設定にしている．

剛体リンク一つのサイズは $1.0[cm] \times (30/40)[cm] \times 5.0[cm]$ であり，ロボットの総質量は $500[g]$，初期形状は円周 $30[cm]$ の円である．剛体リンク間結合ジョイントの可動範囲は $-180[deg] \sim 180[deg]$ である．

連絡先: 藤野 智宏，東京工業大学 総合理工学研究科 知能システム科学専攻，tomohiro@fe.dis.titech.ac.jp

2.2 地形環境

複数の環境を用意することにより，本研究の最終目的である「未知の不整地環境下において，ロボット自身が環境の変化を判断し，環境に適した形状変形・走行動作を行う適応的学習」を実現するために必要な情報を得る．以下に実験に用いる 4 つの地形環境について述べる．

1. 平らな環境

   最も簡単な環境として用意した環境であり，この環境で獲得された動作が基本となる．滑らかな走行が目標となる．

2. 傾斜角度一定の坂環境

   傾斜角度 $10[deg]$, $15[deg]$, $20[deg]$ と $-15[deg]$ の 4 つの坂環境とする．前者の上り坂環境では，形状を縦に細長く伸ばした場合，後方へ回転しやすく前進が困難である．後者の下り坂環境は，逆に前進がしやすい環境となっており，特に前進移動速度を速くするには形状を縦に細長く伸ばす必要が出てくるが，バタバタとした動きとなり滑らかに走行することが困難となる．このような困難を解決し走行することが目標となる．

3. 傾斜角度が変化する環境

   下式のような高さ関数 $height(x)$ を用いて作った直線上の道環境とする．この環境は傾斜角度が $-11[deg] \sim 11[deg]$ となる坂環境が混在するため，状況に合わせて走行動作を切り替えていく必要がある．よって現在の状況が下り坂なのか上り坂なのかを判断し，それぞれに適した走行動作を行うことが目標となる．

   $$height(x) = 5.0 \cdot \cos(0.009\pi \cdot x) \cdot \cos(0.03\pi \cdot x)$$

4. 月重力環境

   重力加速度を地球上の $1/6$ に設定した環境である．この環境では，重力から得られる前進方向への回転モーメントが小さくなるため効率よく走行するにはより縦に細長くなる必要がある．しかしながら，縦に細長くなりすぎると動作がバタバタとしたものになってしまう．このような問題をうまく解決し走行することが目標となる．

2.3 強化学習の定式化

ロボットの制御は中心から剛体リンクへ伸びている伸縮自在な 6 本の棒 $(i = 0, 1, 2, \ldots, 5)$ によって行われる．各棒は $2[cm] \sim 8[cm]$ の範囲で伸縮し，$i = 0$ 以外の制御棒は $-30[deg] \sim 30[deg]$ の範囲でピッチ方向への自由度を持つ．制御棒の長さと角度は与えられた目標値に追従するように制御される．ここで使用する各制御棒の制御器は，設計者が適当にパラメータを与えた簡易な PD コントローラである．時刻 t における制御出力 $a(t)$ は各制御棒の目標長さ $length_{m_i}^{obj}(t)$ と目標角度 $angle_{m_i}^{obj}(t)$ のベクトルで与えられる．

$$a(t) = \left\{ length_{m_0}^{obj}(t), length_{m_i}^{obj}(t), angle_{m_i}^{obj}(t) \right\}, \quad i = 1, 2, \ldots, 5 \qquad (1)$$

また，ロボットは中心の回転角度を計測するセンサ $s_g(t)$, ロボットの高さを計測するセンサ $s_h(t)$, 傾斜角度を計測するセンサ $s_a(t)$ および各制御棒の長さと角度を測定する各センサ $length_{m_i}^{obs}(t)$, $angle_{m_i}^{obs}(t)$ を有する．よって，時刻 t における

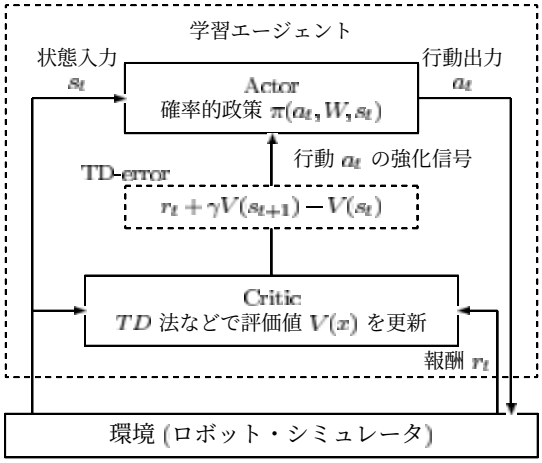

Fig. 2: Actor-Critic の一般的枠組み

入力 $s(t)$ は次式となる．

$$s(t) = \left\{ s_g(t), s_h(t), s_a(t), length_{m_0}^{obs}(t), length_{m_i}^{obs}(t), angle_{m_i}^{obs}(t) \right\}, \quad i = 1, 2, \ldots, 5 \qquad (2)$$

以上から，政策 π は入力 $s(t)$ から出力 $a(t)$ へのマッピングとして表される．

$$\pi : s(t) \rightarrow a(t) \qquad (3)$$

各時刻 t において環境から返される報酬 r については，3 章で述べる．

3. 接近法

3.1 Actor-Critic

本研究では学習アルゴリズムに，連続空間における強化学習手法として実績がある Actor および Critic に適正度の履歴を用いた Actor-Critic アルゴリズムを用いる [4]．適正度の履歴を用いる事で，MDP の状態観測にノイズなどの不完全性や，状態変数の一部しか観測できないという部分観測性が加わるなどによって発生する隠れ状態問題に対してロバストとなる．本研究においては，状態観測数を十分に増やすことが出来ないため隠れ状態問題となってしまう可能性が考えられるのでこのアルゴリズムを用いている．Actor-Critic の一般的枠組みを Fig.2 に示す．

Actor は状態入力 $s(t)$ から行動出力 $a(t)$ への確率分布である確率的政策 π に従って行動を選択する．Critic は離散化された状態 $s(t)$ について，Actor の政策に基づいてその $s(t)$ を訪問した後どれだけ高い報酬が得られるかを表す評価値 $\hat{V}(s_t)$ の保持・更新を行う．

3.2 滑らかな動作を実現するための報酬設計

対象ロボットの目的は滑らかに前進する動作を獲得することである．つまり前進移動を行いつつ倒れ込むような動きは避けるべきである．このように二つのタスクを同時に達成する必要があるような学習においては発達段階を考慮した報酬設計 ($\mathrm{R}_{\mathrm{dev}}$) が有効であると考えられる [5].

具体的には，倒れ込むような動きは縦に細長くなることによって生じると考えられるため，倒れ込みが生じる状態を高さが一定値以上になる場合であるとし，そのような状態を $S_{penalty}$

とする．そうではない滑らかな前進動作をするのに適した状態を S_{normal} とする．以上のように二つの状態を考え時刻 t での状態 $s(t)$ にて得られる報酬 $R(t)$ を次式のように設計する．ただし，Δx は 1 ステップあたりの前進距離とする．

$$R(t) = \begin{cases} \mathrm{SCALE} \cdot \Delta \mathrm{x} & (S_{normal}) \\ \mathrm{PENALTY} & (S_{penalty}) \end{cases} \quad (4)$$

ここで，SCALE はサンプリングタイムを考慮して設計者が与える定数，PENALTY は避けるべき高さになっていることに対する罰として与える負の定数である．

上記の報酬設計により，学習初期段階においては主に状態 $S_{penalty}$ へ遷移してはいけないことを学ぶ．これにより全状態領域から探索すべき状態領域の切り出しが行われる．学習中期段階においては，切り出された状態領域内において前進動作発見が行われる．さらに学習が進むにつれその動作が洗練化されていき，最終的に二つのタスクを同時に達成することが期待される．このようにまず探索領域を選び，次にその領域内で行動を発見し出てきた行動を洗練化するという発達に応じた報酬設計となっている．

一方，自然な考えとして状態を区別せずに移動距離のみに応じた報酬設計 ($\mathrm{R_{move}}$) も考えられ，この場合つねに $R(t) = \mathrm{SCALE} \cdot \Delta \mathrm{x}$ とする．実験では，２つの報酬設計の比較を行う．

4. 実験と考察

4.1 実験の目的と設定

実験の目的は，3 章で述べた報酬設計 ($\mathrm{R_{dev}}$) が様々な環境において有効であるかどうかを検証することである．

3 章に述べた設定の下，2 章に示した各環境に対して前進動作獲得実験を行なった．その際，報酬設計の有効性を検証するために移動距離に応じた報酬設計 ($\mathrm{R_{move}}$) での実験と滑らかな動作を実現するための報酬設計 ($\mathrm{R_{dev}}$) での実験を各環境に対して行い比較を行なった．学習における意思決定時間は 100ms とし，Actor-Critic のパラメータは Table1 に示す値を用いた．

Table 1: 学習アルゴリズムの実験設定

Actor	学習率 $\alpha_\pi = 0.002$ 適性度の履歴の割引率 $\lambda_\pi = 1.0$
Critic	学習率 $\alpha_v = 0.1$ 適性度の履歴の割引率 $\lambda_v = 1.0$
TDerror	割引率 $\gamma = 0.95$

4.2 実験結果

各環境における学習曲線を Fig.3~Fig.6 に示す．また，下り坂環境における獲得動作を報酬設計別に Fig.7 に示す．ここでは，4 つあった傾斜角度一定の坂環境の結果を上り坂環境 (傾斜角度 20[deg]) と下り坂環境 (15[deg]) の二つに分けて示す．学習曲線から平らな環境と下り坂環境においては報酬設計に関係なく動作の獲得は達成されているが，上り坂環境では報酬設計 $\mathrm{R_{dev}}$ のみで前進動作が獲得されたことが分かる．最大平均獲得報酬の値をみると平らな環境においては報酬設計 $\mathrm{R_{dev}}$ のほうが多く得ており最終的に約 1.5 という値を得ている．対して下り坂環境においては逆であり報酬設計 $\mathrm{R_{move}}$ において最大の報酬を得ておりその値は約 7.0 と非常に大きい値となった．また，学習速度も同様の結果であった．

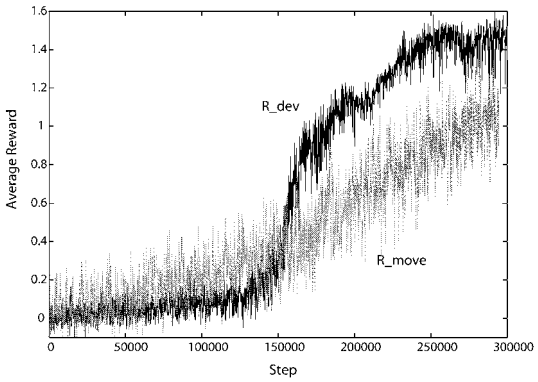

Fig. 3: 平らな環境における学習曲線

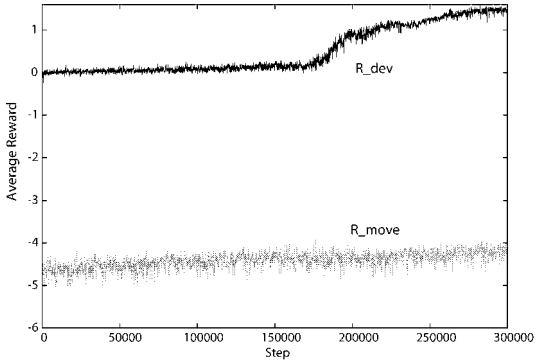

Fig. 4: 上り坂環境における学習曲線

4.3 考察

平らな環境での挙動:

平らな環境では，報酬設計 $\mathrm{R_{dev}}$ により高さ方向の伸長が抑制され，その結果高さを大きく変化させない安定した動作を獲得したと考えられる．すなわち，低い高さを保つことが形状を横に長くすることにつながり，結果として接地面積を大きくするような不安定な状態になりにくい形状を保つためであると考えられる．また，移動速度も R_{move} に対して速い動作であった．このような事から，平らな環境において報酬設計 $\mathrm{R_{dev}}$ は滑らかで速い前進動作を得るのに有効であるといえる．

上り坂環境での挙動:

上り坂の環境では，報酬設計 $\mathrm{R_{move}}$ によって前進動作を獲得することができなかった．この理由として上り坂環境では，縦に細長くなるような形状になってしまうと確実に後方へ回転してしまい前進できない状況となるためである考えられる．よって，上り坂環境においては高さ制限を加えた報酬が無くてはならない設計であるといえる．

下り坂環境での挙動:

下り坂環境では，報酬設計による獲得動作の違いがよく現れている．移動速度および獲得報酬の面からみると報酬設計 $\mathrm{R_{move}}$ のほうがよいといえる．しかしながら，その獲得動作は地面と点で接するような形状をとることによって転がり落ちていくような受動的な動作である (Fig.7(a))．そのような動作では突発的に移動を停止したい場合などにおいて対処することができないと考えられるため適しているとは言えない．

一方，報酬設計 $\mathrm{R_{dev}}$ では低い姿勢を保ち高さの変位を小さくした安定した動作であり (Fig.7(b))，移動速度も平らな環境と同程度である．よって，下り坂環境において安定した動作を獲得することに報酬設計 $\mathrm{R_{dev}}$ は有効であるといえる．

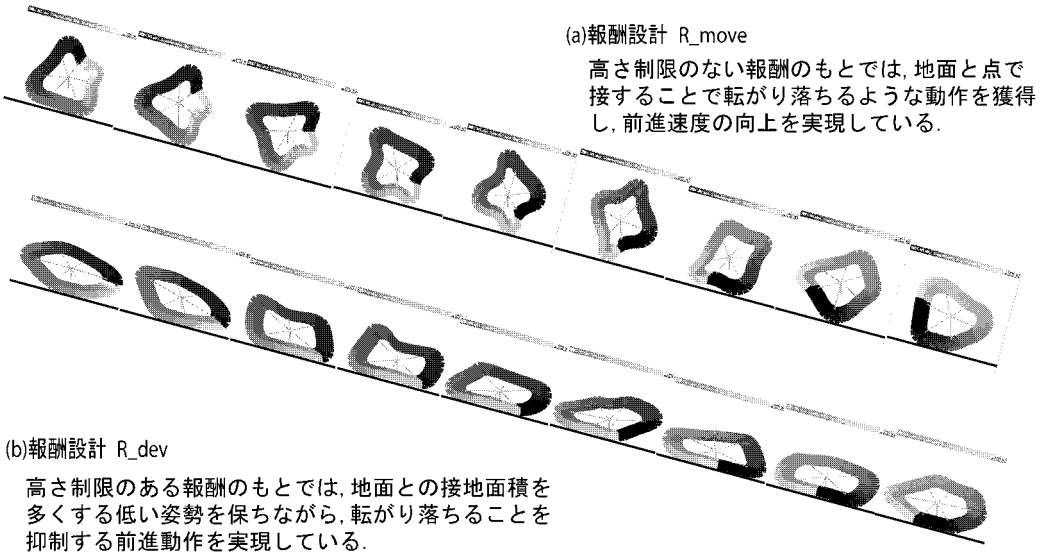

Fig. 7: 下り坂環境での獲得動作の様子

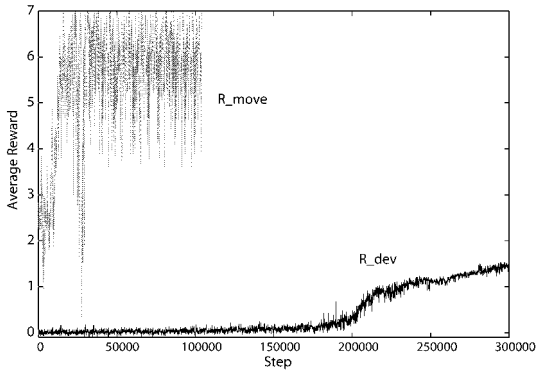

Fig. 5: 下り坂環境における学習曲線

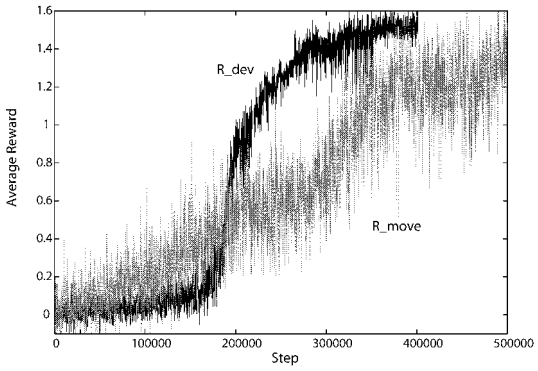

Fig. 6: 月重力環境における学習曲線

月重力環境での挙動:

月重力環境での結果からは，報酬設計の違いによる獲得動作の違いはあまり見られなかった．この要因としては重力が低いために回転モーメントが発生しにくいということが考えられる．つまり，前方へ回転するためにはある程度の高さが必要となり結果として得られた動作に大きな違いがなくなってしまったのだと考える．よって月重力環境においては高さ制限を加えた報酬設計の影響が少ないといえる．

5. おわりに

本研究では，二つの報酬設計を用いて各環境での動作獲得実験を行うことにより，提案した発達段階を考慮した報酬設計 R_{dev} が 4 つの環境において有効であることを確認した．

今後は，未知環境に対して適応していけるような制御システムを構築していく．そのために未知環境をロボット自身が判断できるようにする必要も生じてくると考え，そのような判断に用いることができる知見を様々な環境に対して行う実験から探していきたい．

参考文献

[1] A. Kamimura. et al. : Automatic Locomotion Pattern Generation for Modular Robots, Proc. of IEEE Int. Conf. on Robotics and Automation, pp.714-720 (2003)

[2] Y. Sugiyama, S. Hirai: Crawling and Jumping of Deformable Soft Robot, Proc. of IEEE Int. Conf. on Intelligent Robots and Systems, pp.3276-3281 (2004)

[3] 木村 元，山村 雅幸，小林 重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近，人工知能学会誌, Vol.11, No.5, pp.761–768 (1996)

[4] 木村元，小林重信: Actor に適正度の履歴を用いた Actor-Critic アルゴリズム，人工知能学会誌，Vol.15，No.2，pp.267-275 (2000)

[5] 角田英太郎，青木圭，佐久間淳，小林重信: 強化学習によるカササギの歩容獲得，計測自動制御学会 第 17 回自律分散システムシンポジウム，pp.159-164 (2005)