6L-6

強化学習によるロボットの歩行動作獲得

橋本佳典　大藪又茂

金沢工業大学大学院 工学研究科 システム設計工学専攻

1.はじめに

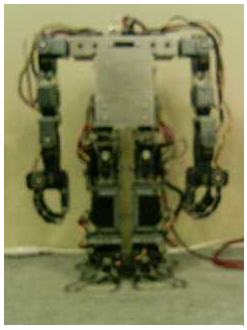

Fig.1　ロボット

現在、社会の中には多くのロボットが存在する。しかし実用化され活動しているロボットの多くは車輪や多足での移動が主流となっていて、人間のように二足歩行のロボットは少ない。二足歩行をする利点は、人間が受け入れやいロボットをデザインできることである。二足歩行ロボットの問題点は、移動中の姿勢が不安定であることがあげられる。車輪などのロボットは簡単な計算で姿勢を安定させることができるが、二足歩行ロボットは「ZMP 制御」や「順・逆運動学」といった複雑な計算を行なわなければ、姿勢を安定させることができない。また、「ZMP 制御」や「順・逆運動学」などを盛り込んだプログラムは複雑となり、ロボットのハードウェアの変更があった場合に、その都度パラメータなどの変更を行なわなければならず、ソフトウェア開発への負担が大きかった。強化学習では従来の力学的なパラメータの設定法に代わり、自発的な繰り返し学習により最適値に近いパラメータを獲得することができる。そのためソフトウェア開発の負担の軽減が期待される[1]。

本研究では、強化学習を採用してロボットの二足歩行獲得を試みた。

2.問題設定

2.1　二足歩行ロボット

ロボットは片足 6 関節、片手 5 関節の合計 22 関節を持っている。距離を測ることのできる赤外線センサと、重力加速度を計測できる加速度センサの 2 種類のセンサを搭載している。アクチュエータには市販されているロボット用サーボモータ（以下モータ）を使用した。ロボットの制御にはマイコンを使用した。マイコンにはロボットの制御以外にも強化学習のプログラムも実装した。

Acquisition of Robot Waking by Reinforcement Learning
Yoshinori Hashimoto, Matashige Oyabu
Department of Systems Design Engineering Kanazawa Instituteof Technology

2.2　学習環境

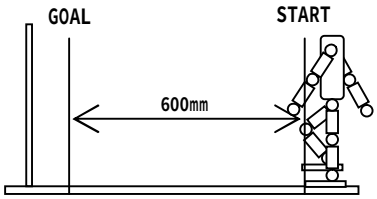

Fig.2　学習環境

Fig.2 に示すような一定の距離（600 mm）を一定の時間で歩くことを学習させた。時間の計測は、マイコンに内蔵されているタイマを使用しタイマ割り込みにより行なった。割り込み間隔は 10msec とした。また、本実験ではあらかじめ設定した歩行パターンを使用して学習を行なった。この歩行パターンは 3 つの基本動作を組み合わせたものである。Fig.3 に基本動作 1 を示す。基本動作 1 は重心を左右に移動させる動きである。この動きで学習したのは、Fig.3 に示すθ_1の角度である。

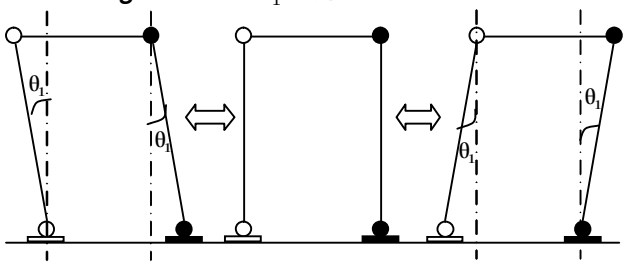

Fig.3　基本動作 1

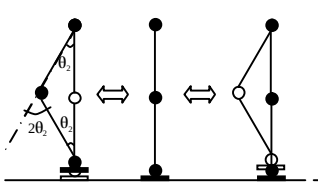

Fig.4　基本動作 2

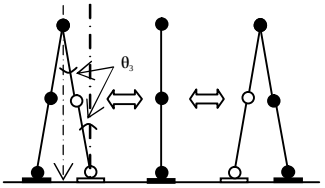

Fig.5　基本動作 3

基本動作 2 を Fig.4 に、基本動作 3 を Fig.5 に示す。Fig.4 と Fig.5 はロボットを左から見た図で黒丸が左足の関節を示している。基本動作 2 は足を上げる動作である。θ_2は 15 度の一定にして実験を行なった。Fig.5 の基本動作 3 は足を広げる動きであり、θ_3の角度を学習した。

3.学習について[2]

3.1 学習目的

学習の目的は 600mm の距離を 5.7sec で歩行するものとした。価値の更新は式(1)に示す。

$$Q_{k+1} = Q_k + \alpha[r - Q_k] \quad \cdots (1)$$

Q_{k+1}：更新後の価値　Q_k：更新前の価値

α：学習率　　r：報酬

学習率αは 0.8 と一定で学習を行い。報酬 r は式(2)により計算した。

$$r = |5.7 - Ta| \times 100 \quad \cdots (2)$$

Ta：ゴールするまでに掛かった時間

つまり、価値と報酬は小さいほど良いように設定した。

3.2 学習範囲

学習範囲はθ_1とθ_3の積数だけ存在する。使用したモータの動作範囲は 180 度であるが、フレームなどが干渉する角度と、明らかに倒れてしまう角度については学習範囲から除外した。これによりθ_1の範囲は１～25 度の 25 通り、θ_3の範囲は１～20 度の 20 通りとなった。よって学習範囲は 25×20＝500 の 2 次元配列で表記することができる。本研究ではマイコンを使用して学習を行なったので実数の扱いが難しく、整数のみを扱うようにプログラミングした。そのため学習法として Actor-Critic ではなく Q-learning を使用した。

3.3 アルゴリズム

(1)価値の初期化

本研究では学習範囲が 25×20 で Q-learning を使用したので 500 個の価値Qが存在することになる。まず、この 500 個の価値Qを一定の値に初期化した。

(2)行動の選択

まず、θ_1の角度の選択を行なった。行動の選択には ε グリーディ政策を使用した。1 -εの確率で貪欲に行動する。この場合は式（3）によりθ_1の価値Q_{θ_1}を求め、25 個のθ_1全てのQ_{θ_1}を計算し、最小のQ_{θ_1}を持つθ_1を選択した。

$$Q_{\theta_1}(\theta_1) = \sum_{\theta_3=1}^{20} Q(\theta_1, \theta_3) \quad \cdots (3)$$

また、εの確率で探索的な行動を選択する。この場合はランダムに角度を選択した。

次にθ_3の角度を選択した。この角度もθ_1と同様に選択するが、貪欲に行動する場合に違いがある。式(4)にθ_3の価値Q_{θ_3}を示す。20 個のθ_3から最小の$Q_{\theta 3}$を探し、そのθ_3を行動として選択した。

$$Q_{\theta_3}(\theta_3) = Q(\theta_1, \theta_3) \quad \cdots (4)$$

またεの初期値は 0.8 とした。

(3)行動

（2）で選択したθ_1とθ_3の値で歩行し、ゴールするまでの時間を計測し、式（2）により報酬を計算した。もし転倒や足踏みを行い 8sec 以内にゴールできない場合の報酬はペナルティとして 600 を与えた。

(4)価値の更新

価値の更新を式(1)により計算した。

(5) εの値を減少させる。

(1)～(4)までの行程を 1 回の学習とし、これを 100 回繰り返した。εの値は学習回数が偶数の時に 0.01 減少させた。

4.実験結果

Fig.6 に 100 回の学習を 5 回繰り返し、平均価値を計算した結果を示す。全体的に右下がりのグラフになっている。これは学習が進むにつれて転倒や足踏みがなくなっていき、価値の大きい悪い行動の選択が少なくなっていることを示している。

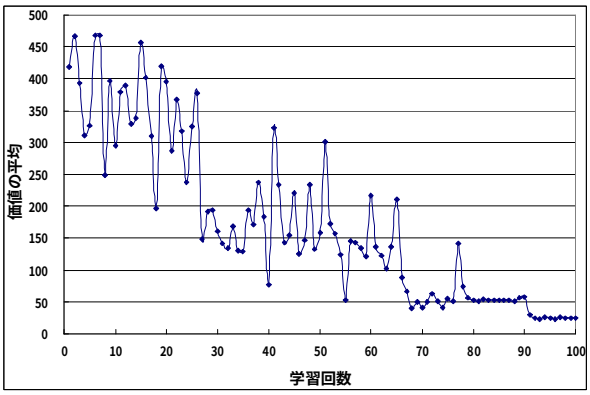

Fig.6　5 回の実験結果の平均

5.考察

学習回数 60 回を過ぎた辺りからゴールできない行動はほとんど見られなかった。すなわち、ここで採用した学習システムでは、ほぼ 60 回の学習でゴールに到達できる歩行を獲得したといえる。また 60 回を過時点でεの値は 0.5 以下となっている。このようにして、学習を進めることによりロボットは二足歩行を獲得することができる。

・参考文献

[1]木村元・山下透・小林重信：強化学習による 4 足ロボットの歩行動作獲得，電気学会 電子情報システム部門誌，Vol.122-C，No.3 ，pp.330-337 (2002)

[2]Richard S.Sutton・Andrew G.Bart,三上貞芳・皆川雅章共訳:強化学習 Reinforcement Leaning，森北出版，2003