階層型強化学習を用いた実ロボットによる起立運動の獲得

奈良先端科学技術大学院大学・川人学習動態脳プロジェクト (科技団)　森本 淳 (PY)　銅谷 賢治

Acquisition of Stand-up Behavior by a Real Robot using Hierarchical Reinforcement Learning

Nara Inst. of Sci. and Tech., Kawato Dynamic Brain Project, JST : Jun MORIMOTO (PY)　Kenji DOYA

Abstract :　In this paper, we propose a hierarchical reinforcement learning method which enables a real robot to learn nonlinear control tasks in a realistic number of trials. In the upper level, the learner coarsely explores the low-dimensional state space. In the lower level, the learner finely explores the high-dimensional state space. A real robot successfully learned to stand up based on the subgoal sequence acquired by the learning with an internal model.

Keywords : reinforcement learning, hierarchical, motor control, stand up, real robot

1　はじめに

近年，人間が制御則を作り込んでロボットを動かすことに関しては，高度な動作を行うロボットが出現している[1)]しかし，将来人間と共存し，人間の代わりとなって働くようなロボットを実現するためには，固定された制御則を用いるだけでなく，動的に変化する環境の中で，ロボット自身が学習によって制御則を獲得することが要求される．そのような要求に応えるため，ロボットに未知環境において行動を獲得させる手法として，強化学習が注目を集めている．強化学習とは，模範となる出力系列が与えられなくても，最終的に課題がどれだけ達せられたかという評価信号から望ましい制御則を発見する学習の枠組である．

一般に強化学習を用いると，あるタスクを獲得するのに多くの試行錯誤を必要とする．しかし，実ロボットで学習することを考えた場合，たとえ低自由度のロボットであっても，その耐久性を考慮すると，少ない試行回数でタスクを達成することが望まれる．そこで我々はこれまで強化学習に階層構造を導入する手法を検討してきた[2, 3)]．本研究ではそれらの知見を用いて，まず内部モデルを用いた仮想的な試行錯誤を通じた学習を行い，それらの経験をもとに実ロボットでのタスク獲得を可能にする手法を提案する．

本研究で扱うタスクは，図 1のようなロボットの動的起立運動である．起立運動のような過渡的な運動は，歩行運動のような定常的な運動に比べて，理論的に制御則を導いたり，人間が発見的に制御則を求めることは困難であり，強化学習を用いて制御則を獲得する必然性が大きい．

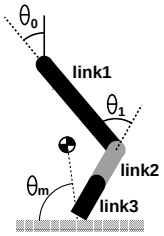

図 1: ロボットの形状：実ロボットの全長は 0.7[m] (link1:0.4[m]，link2:0.15[m]，link3:0.15[m])，重さは約 5.0[kg]．用いたセンサは，関節角におけるエンコーダと，link3 下部に取りつけたポテンションメータおよび link1 に取りつけたジャイロセンサである．

2　階層型強化学習

上位階層では，低次元化・離散化された状態空間上で行動系列を学習し，その行動系列を下位階層にサブゴールとして与える．上位階層の学習方法として，本研究では $Q(\lambda)$ 学習[4)]を用いる．一方，下位階層では，上位階層によって与えられるサブゴールを達成するための高次元連続状態空間中の軌道を学習する．学習方法としては連続系 $TD(\lambda)$学習[5)] を用い，その実現方法として actor-critic 法を用いる．各サブゴールは上位階層の低次元状態空間で決定されるため，下位階層の高次元空間においては，サブゴールは点ではなく超平面となる．報酬は，上位階層においてはタスクが成功したかどうかに応じて与えられ，下位においては上位階層の決めたサブゴールに到達したかどうかによって与えられる．ロボットの状態がサブゴールを中心としたある範囲内に入ったときにサブゴールに到達したと判断し，上位階層での状態が更新され，次のサブゴールが下位階層に提示される（図 2参照）．

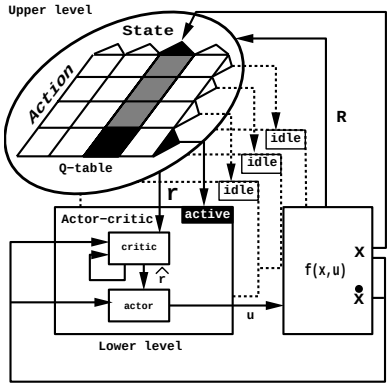

図 2: 階層構造: 灰色の部分が現在の状態を示しており，黒の部分が上位階層によって取られた行動を示す．また，r は上位階層によって下位階層に与えられる報酬を示す．

3　起立運動学習

上述の階層型強化学習法を用いて，図 1のような 3 リンク 2 関節のロボットによる起立運動学習を行った．ただし，今回は簡単のため，足下に近い方の関節をサーボによって固定し，2 リンク 1 関節のロボットとして用いた．

3.1　上位階層での学習

上位階層への入力としては，ロボットの重心位置の（足下から見た）角度と関節角 ($\boldsymbol{X}=(\theta_g, \theta_1)$) を 30[deg] 単位で離散化した表現を用いた．上位階層での状態遷移は下位階層においてサブゴールが到達されたときのみ行われるため，上位階層では時間的にも粗い探索を行うことになる．出力も同様に，ロボットの重心位置の角度 θ_g と関節角 θ_1

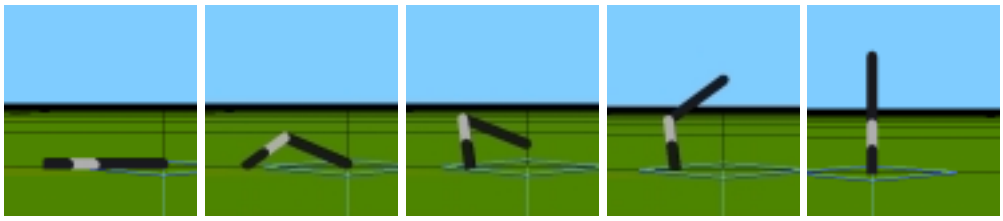

図 3: 上位階層の行動系列 (サブゴール系列)

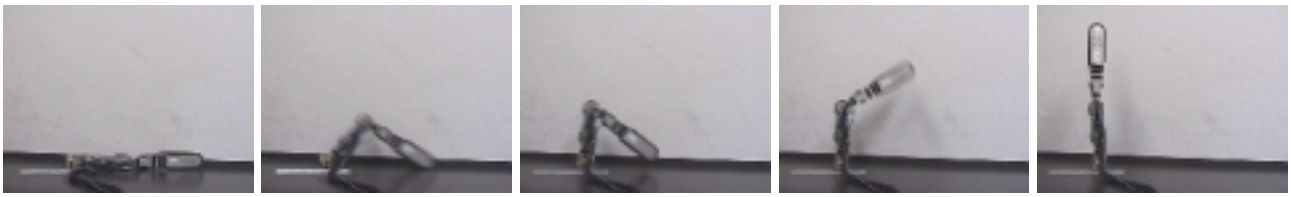

図 4: 実ロボットによる下位階層の行動系列

を用いた．$\Delta ACT = 30[deg]$ として，$\boldsymbol{U}$=$(\Delta\theta_g, \Delta\theta_1)$ = $(\pm 2\Delta ACT,\ \pm\Delta ACT,\ 0)$ という離散的な行動を取る．よって，下位階層に与えるサブゴール姿勢は$\boldsymbol{X} + \boldsymbol{U}$となる．しかし，下位階層での状態表現に合わせるため，実際には重心位置の角度 (θ_g) を頭のピッチ角 (θ_0) に変換したものをサブゴールとして用いる．また，上位階層に与える報酬は $R = 0.5(\frac{y}{L} + 1) + R_s$ とした．ただし，y はサブゴールが達成された時点でのロボットの頭の高さ，L はロボットの全長を表す．R_s は起立運動全体が最終的に達成されたときに $R_s = 1$，転倒等によって達成されなかったとき $R_s = 0$ とした．1 回の試行は試行が 10 秒間続くか，一度起き上がろうとして腰または頭を地面につけば（転倒すれば）終了とした．

3.2 下位階層での学習

下位階層への入力としては，ロボットの頭の傾き角（ピッチ角）θ_0，関節角 θ_1 とそれらの角速度を用いた $(\boldsymbol{x}$=$(\theta_0, \theta_1, \dot{\theta}_0, \dot{\theta}_1))$．出力は関節で発生するトルクとした $(u = \tau_1)$．また，下位階層の評価関数を近似する critic と，関節での非線形フィードバック制御を行う actor には，効率的な関数近似方法である INGnet を用いた[6]．下位階層での行動出力は線形サーボと actor の出力 $f_a(\boldsymbol{x})$ の和により次式のように与えた．

$$\tau_b = k(\theta_{s1} - \theta_1) - b\dot{\theta}_1 + f_a(\boldsymbol{x}) \qquad (1)$$

下位階層に与える報酬は，サブゴール姿勢に近ければ高い報酬が得られるよう次式のようにした，

$$r(\boldsymbol{\theta}, \boldsymbol{\theta_s}) = \exp(-\sum_{i=0}^{1} \frac{|\theta_i - \theta_{si}|^2}{\sigma_{ri}^2}) - 1 \qquad (2)$$

$$r = -1.5\ (on\ failure) \qquad (3)$$

ただし，$\boldsymbol{\theta}$=(θ_0, θ_1) は現在の姿勢，$\boldsymbol{\theta_s}$=$(\theta_{s0}, \theta_{s1})$ はサブゴールの示す角度，σ_r は報酬関数の幅を表す (本実験では $\sigma_r = 40[deg]$ とした)．また，$\sum_{i=0}^{1} |\theta_i - \theta_{si}| < 10\ [deg]$ のとき，サブゴールに到達したと判断する．

4 実験結果

内部モデルを用いた仮想的な試行錯誤を通じて階層型強化学習を行った結果 220 回（5 回の実験の平均）の試行で起立運動を獲得することができた．その結果，上位階層の学習によって得られたサブゴール系列は図 3のようになった．さらに，上述の仮想的な経験をもとに，実世界と内部モデルの誤差を吸収するための学習を実ロボットを用いて行った．実ロボットによる学習の結果，約 180 試行で図 4のような起立運動を獲得した．図 4の右から 2 番目の図のような姿勢 (静的つり合いの取れない姿勢) を通過していることより，ロボットは動的な起き上がりを獲得していることがわかる．

5 まとめ

本研究では，実ロボットでの学習を可能にするための枠組みとして，強化学習に階層構造を導入した．内部モデルを用いた仮想的な学習によって得られた経験をもとに，実ロボットで学習を行うことで，起立運動の獲得を行うことができた．

今後はより自由度の大きい制御対象に本手法を適用することを試みる．また，実ロボットを用いる場合に生じるセンサの遅れの補償や，内部モデルの学習による獲得についても検討する予定である．

参考文献

1) 広瀬真人, 竹中透, 五味洋, 小澤信明. 人間型ロボット. 日本ロボット学会誌, Vol. 15, No. 7, pp. 23–25, 1997.

2) J. Morimoto and K. Doya. Hierarchical reinforcement learning of low-dimensional subgoals and high-dimensional trajectories. In *The 5th International Conference on Neural Information Processing*, Vol. 2, pp. 850–853, 1998.

3) J. Morimoto and K. Doya. Reinforcement learning of dynamic motor sequence: Learning to stand up. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 3, pp. 1721–1726, 1998.

4) J. Peng and R. Williams. Incremental multi-step Q-learning. *Machine Learning*, Vol. 22, pp. 283–290, 1996.

5) K. Doya. Reinforcement learning in continuous time and space. *Neural Computation(In Press)*.

6) 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)(印刷中).