F-024

Deep Q-Networkを用いた模倣による動作自動獲得
Imitation of Motion Pattern using Deep Q-Network

中村 格† 飯塚 博幸† 山本 雅人†
Itaru Nakamura Hiroyuki Iizuka Masahito Yamamoto

1. はじめに

ロボットの動作を設計する手法は数多く存在する．たとえば，人間が関節角を指定し，動作を設計する方法があるが，関節の数が多くなるほど指定しなければならない値が多くなり，制御が困難になるという問題がある．

そこで，近年はロボットに動作を学習させることで動作を生成する手法が注目されており，学習手法として強化学習やニューラルネットワークを用いた研究が盛んに行われている[1]．また，動作生成の手法の中に，既にある動きから模倣して動作を生成する手法がある．舞踊ロボットのように，伝統的な技術をロボットで再現するなど，人間にできることをロボットに再現させる研究や，すでにある動作を学習元とし，新たな動作を獲得させる研究など模倣学習による動作獲得に関しても多くの研究が行われている[2].

本研究ではロボットに模倣から動作を獲得させることを目的として，Deep Q-Networkを用いて学習を行う．模倣による動作生成には，模倣対象の関節角の情報や，マーカー等を用いて関節位置を取得して模倣を行う手法も存在する．しかし，これらの手法は模倣対象と同じ形状の模倣ロボットにのみ利用可能で，異なる形状のモデルに対して利用不可能である．また，YouTubeなどから入手した動画からは関節角などの情報を得ることはできないため，このような場合は利用できない．本研究では，容易に入手できる動画に対して模倣することが可能な学習手法を構築することを目的とし，関節角等の情報を用いず，複数のモデルに対しての動作の模倣を生成する手法を提案する．

この目的のため，模倣対象と模倣ロボットの両方をシルエット化することで，その色や質感といった特徴を用いず，形状を合わせることで模倣を行う手法を提案する．

2. 提案手法

動作を模倣するために，ロボットアームのモデル用いて，Deep Q-Network(DQN)による模倣動作を行う．

模倣対象の運動を画像の系列で与え，自分自身の運動も画像として与える．各時刻において模倣対象と自身の2枚の画像をDQNに入力し，行動に対応するQ値を出力する．行動の候補それぞれに対してQ値を求め，得られたQ値が最大の行動をとり，ロボットアームのモデルの状態を更新し，次の時刻での入力画像とする．これを繰り返すことで動作の模倣を行う

2.1 ロボットアームのモデル

本研究に用いるロボットアームのモデルはOpenGLを用いて作成した．図1に示すように肩に2自由度，肘に1自由度をもつ腕が2本あるロボットアームのモデルを用いる．

肩部の関節は，−90度から+90度の範囲で可動する．肘部の関節は上腕部から直線上にある状態を0度とし，90度折れ曲がった状態まで可動する．関節はすべて10度ごとに可動する．ロボットアームのモデルは関節角を指定することで，形状を変化させ，動作を生成する．

本研究では，動画から得られた画像系列をシルエット化して模倣対象としているため，生成したロボットアームも同様にシルエット化することで形式を統一した．

2.2 Deep Q-Network

DQNは強化学習の代表的な手法であるQ学習と深層学習のConvolutional Neural Network(CNN)を組み合わせた手法で，2013年に提案された[3].

DQNは毎ステップ，現在の状態，１ステップ後の状態，行動，報酬をReplay Memoryに保存し，そこから学習の際にランダムサンプリングし，ミニバッチ学習を行うExperience Replayと呼ばれる手法と，学習時に期待値の計算に用いるTarget networkの更新を一定の間隔で行う手法によって学習を安定させている．

2.2.1 Q学習

Q学習はエージェントが環境に対し，ある方策に基づいて行動をおこない，その行動に対し得られる報酬に基づいて行動価値関数(Q値)を更新し，最適な行動を選択できるようにする手法である．Q値の更新式は以下で表される．

$$Q(s_t,a) = Q(s_t,a) + \alpha\left[r_{t+1} + \gamma \max_p Q(s_{t+1},p) - Q(s_t,a)\right] \quad (1)$$

ここでs_t，a_t，r_tは時刻tにおけるエージェントの状態，行動，および報酬を表し，αは学習率，γは割引率を表す．

2.2.2 CNN

CNNは，フィルタを用いた畳み込みを行う畳み込み層とプーリング層を接続した構造を持ち，主に画像認識の分野で用いられている．DQNではCNNを画像からQ値を推定するために用いている．

本研究では7層で構成されるネットワークを用いた．畳み込み層をC1，C2，プーリング層をP1，P2，全結合層をN1，N2とし，C1，P1，C2，P2，N1，N2の順に配置する．畳み込み層のフィルタサイズを3×3，プーリング層のフィ

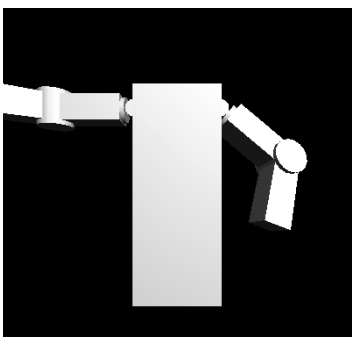

図1 模倣ロボットモデル

† 北海道大学 Hokkaido University

ルタサイズを2×2とし，特徴マップの縦×横×チャネル数を順に40×40×64, 20×20×64, 20×20×64, 10×10×64, 1×1×13として，40×40×2の画像を入力する．CNNの入力は現在のロボットアームの画像と目標画像のグレイスケール画像２枚を入力としているため，２チャネルとなっている．

また，CNNの出力は，関節を＋方向に曲げる，−方向に曲げるという操作を６個の関節に対し行う12種類の動作に停止を加えた13個の動作のQ値となっている．

3. 静止画模倣実験

ロボットアームのモデルの画像をDQNへの入力とし，毎ステップ行動を決定し，新たな画像を取得する．これを繰り返すことで目標画像との差分が小さい形状になる行動を学習する．取得した画像は400×400ピクセルの画像を40×40ピクセルに縮小して入力画像とする．

本研究では模倣ロボットと同自由度，同一形状の目標画像を複数用意し，エピソードごとに切り替えることで１つのネットワークに対して複数の目標画像に対する学習を同時に行う．

目標画像を与えて，それに近づける行動を出力することを目的としているため，目標画像と現在のロボットアームのモデルの画像の差分が大きくなるほど小さい報酬(負の報酬)を与えるように設定する．そのために，目標画像と現在のモデルの画像の各ピクセルの比較を行い，その差を最大値が１になるように正規化し，マイナスをかけて負の報酬として与える．

3.1 学習の流れ

学習はエピソードごとに目標画像から1つを選択し，毎ステップロボットアームの状態をDQNに入力し，行動を出力する．これを100ステップ繰り返し，100ステップでの報酬の平均値を評価値とした．訓練に ε -greedy方策を用いているため，毎ステップ ε の更新を行い，一定エピソードごとに ε の値を0としたテストを行い，性能を確認する．

学習開始時は，数枚の目標で学習を行い，テストの評価値が規定値を超えるごとに目標画像を追加する追加学習を行った．目標画像を追加する際に ε の値を高くすることで，ランダムに探索を行う頻度を上げている．ここでは目標画像の初期値は10枚とし，目標画像追加後の ε の値は0.7とした．

3.2 追加学習用のExperience Replay

本研究ではReplay Memoryを2つ用意して学習を行った．Replay Memory1(RM1)にはその時点までに学習したすべての目標画像に対する状態，行動，報酬の組を保存し，Replay Memory2(RM2)には最後に追加された目標画像に対する状態，行動，報酬の組のみを保存する．

学習時には，バッチサイズの半分をRM1から取り出し，残りをRM2から取り出す．これにより，追加された目標画像がバッチ学習の際に選ばれずに学習が行われない問題を解消し，学習に時間がかかるDQNにおいて高速かつ安定な学習ができることを期待した．

4. 実験結果

2000エピソードの学習を行ったところ，訓練を繰り返すにつれて評価値が上がることが確認できた．テストにおける評価値が規定値を超え，目標画像を追加した際は一時的に評価値が下がったが，学習を進めると評価値が再度上がり，新たな目標画像にも対応できるようになったことが確認できた．目標画像とDQNによる模倣動作の結果できた画像系列を図２に示す．ここで目標画像は，実際の人間の動きを元にした画像を用いた．目標画像はカメラで撮影した1920×1080の写真をグレイスケールに変換し，グレイスケール値100以上を0，それ以外を255とすることで2値化し，その後肩の位置がロボットアームの肩の位置にとなるようにトリミングと画像の拡大縮小を行い，400×400の画像とした．関節を曲げる動作と伸ばす動作を繰り返すなどの挙動が見られ，評価値が上下するなどの特徴が見られたが，目標画像とロボットアームのモデルの画像の差分を小さくするような動きは学習できており，多くの目標画像に対し，概形の模倣はできていた．

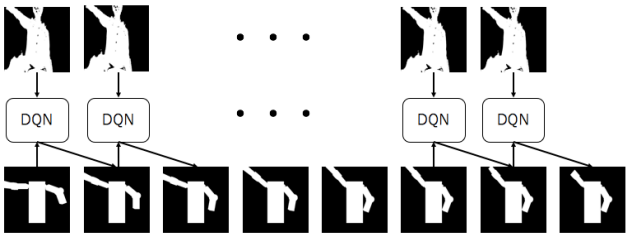

図2　(上段)人物の動作を元にした模倣対象
(下段)ロボットアームモデルによる模倣動作

この学習済みネットワークを用いた行動の動作モデルと異なる形状のロボットアームのモデル及び実際に人間がとった動きを元にした目標画像に対する評価値及びランダムに行動したときの評価の比較を行った結果，DQNを用いた手法では，別のロボットアームのモデル，人間の動きを元にした目標画像の両方に対する模倣において，ランダムに生成した行動より評価値が高くなっていたため，有効であるといえる．

5. おわりに

本研究では，Deep Q-Networkを用いてシルエットに対する模倣動作の学習が可能であることを確認した．また，学習したネットワークは形状の異なるモデルや実際の人間の動きに対しても模倣可能であることも確認した．

しかし，実際の人間の動きを模倣する際，画像の下準備といった問題が存在するため，ロボットアームのモデルと人間の各部位の対応付けを行うなど，大きく異なるモデルに対しても適用可能にすることが今後の課題である．

参考文献

[1] S.Levine,C.Finn,T.Darrell,P.Abbeel,"End-to-End Training of Deep Visuomotor Policies",JMLR17,2016

[2] 中岡慎一郎，中澤篤志，横井一仁，池内克史，"シンボリックな動作記述を用いた舞踊動作模倣ロボットの実現"，電子情報通信学会技術研究報告 : 信学技報，103(390)，55−60，2003.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves,Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller."Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602, 2013.