ISSN 2186-5647

2-53

稼働域が干渉しあうロボットアームの強化学習による動作最適化

日大生産工(院)　○日暮　一登　　日大生産工　風間 恵介
日大生産工　丸茂 喜高

1. 緒言

現在の産業用ロボットは, ティーチングコスト削減のため, 強化学習の導入が模索されている. しかし, 既存のロボットアームと強化学習に関する研究は, 既存の研究はロボットアームと運搬対象が1対1のものが大半である. 加えて, 複数のロボットアームの協調動作に関しては, 机の上に置かれた6つのボタンを押すという簡易的なもののみである[1]. そこで, 2台のロボットアームの稼働エリアが干渉する状況での動作に着目した. ロボットアームの可動エリアが干渉するという, 複雑な動作の最適化は, 強化学習による経路生成と相性が有効だと考えた.

強化学習を用いた制御を行う場合, 学習を早く完了させるには, シミュレーション上で学習を行い, 完成した学習ファイルがそのまま実機で扱えるものであることが望ましい. しかし, シミュレーションで動作を生成する際と, その動作を実機に適応する際とでは, 課題が異なるケースが大半であり, それぞれの事情が相反するパターンも少なくない. 例えば, 物理演算機の接触判定, シミュレーション環境での画像学習などは, 動作生成自体はうまくいったとしても, 実機において, センシングによる観測や報酬関数を与える条件を再現するのは難しい.

本研究では, 強化学習にて, 2台のロボットアームの稼働エリアが干渉する状況で動作する協調制御を実現し, その再現性を検証すること目的する. 今回はシミュレーション上の学習および実機の動作の双方の観点で, 上記の条件における最適なセンサ類の配置や学習手法などの比較検討を行う.

2. 強化学習の概要

学習環境を設定するに当たって議論すべき項目を説明するために, 強化学習の基本的な概要について説明する[2]. まず強化学習を行うにあたって設定する必要があるのが, 状態sと行動aである. 状態sは数字であれば何でも扱えるため, 画像のデータや座標など多岐にわたって代入することが可能である. 続いて, 行動aは, 離散値と連続値を扱える. ここで, 強化学習に使用する関数の1つである状態価値関数を(1)に示す.

$$V\pi(s) = \pi(a|s) \cdot p(r, s'|s, a)\left\{r + \gamma \cdot V\pi(s')\right\} \quad (1)$$

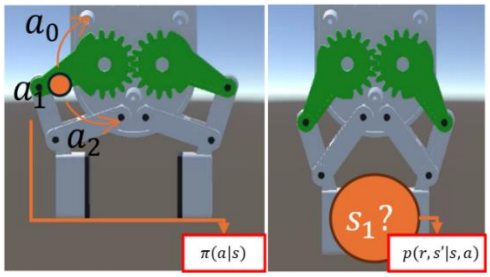

Fig. 1 Correspondence between robot arm and state value function

Fig.1に, この状態価値関数とロボットアームの先端との対応を示す. a_0の場合開く方向, a_1の場合停止, a_2の場合閉じる方向の3段階の離散値で定義し, 先端が何も掴んでいない状態ではs, 何かを掴んだ状態になればs'に遷移するものとする. このとき, どの行動aを選択するのが最適であるかは, 先端の先の搬送物の有無で変わる. 搬送物ない場合, 先端は開いたままの方がよいためa_0あるいはa_1が, 有る場合は, 搬送物を掴むためにa_2が出力されるのが理想である. しかしこれは手動制御ではないため, 搬送物の有無を判断してa_2を選択かどうかは, 機械が確率的に決定する. その確率を最適化するために, 状態価値関数$V\pi(s)$には2つの確率が定義されている.

1つ目は, $\pi(a|s)$で, これは状態sに対して機械が行動aを起こす確率であり, 今回は3種類のaが存在する. しかし, それらは同様に確かではなく, 学習を進めることでその比重を変化させ最適化されていく. この比重をポリシーという. 2つ目の確率は, $p(r, s'|s, a)$である. これは状態sが特定の行動aに対して, 状態s'に遷移する確率を表す. 学習初期はこの確率は不確かなものであるが, 学習を進めることでサンプリングを行い, その精度は上昇する.

状態価値関数$V\pi(s)$は, 始めは0から始まるが報酬rを発見すると増加し, 現在の状態sがどれ

Motion Optimization by Reinforcement Learning for Robot arms with Interfering Operating Ranges

Kazuto HIGURASHI, Keisuke KAZAMA and Yoshitaka MARUMO

くらい報酬に近いかを評価する指標となる．また，実際に行動aを行い，状態価値関数を更新することはstepという単位で管理される．stepをただ重ね続ける場合，状態sも変化し続ける．最適経路問題を取り扱う上では，この状態sをリセットしたいケースが存在する．このリセットする回数をepisodeという．強化学習においては状態s，行動a，報酬r，そしてstep，episodeなどをどのように設定するかが重要となる．

3. 実験概要

行動a，状態s，episodeの有無について比較検討する．比較検討を行わない項目がある理由は，状態s，報酬r，その他ハイパーパラメータは比較可能な幅が極めて広く，組み合わせが膨大となる点，step数を増やしすぎるとシミュレーションに時間がかかりすぎてしまう点を考慮し，厳選した．行動aは，順運動学・逆運動学・動的計画法の3項目，状態sは，タスク管理変数+搬送物座標+アーム先端座標，アーム先端に設置したカメラの画像データ，定点に設置したカメラの画像データの3項目，episodeは，End episodeの有無の2項目である．これら3×3×2=18パターンの組み合わせとなるシミュレーションを作成し，それぞれ完成した学習モデルを用いてシミュレーション上での搬送物取得率，および実機環境での搬送物取得率を試験する．また，それぞれの学習モデル作成時の報酬関数なども比較し，学習の成績を評価する．

4. 実験結果および考察

Fig.2に，コンベアが停止した環境で学習を行っている様子を示す．中央部の橙色の太線①がコンベアであるが，今回は動作させない．この上に搬送物②を4つ設置する．②はロボットアーム③が回収し，上部のエリア④へ②を移動させる．これを4回繰り返すとepisodeが終了する．また，動作中に床や他のロボットアームに接触した場合もepisodeは終了する．Fig.3に，コンベア①を停止した環境での報酬関数の変化を示す．関数が5に収束しているIは，ロボットアームのそれぞれの関節角度，ロボットアームの先端座標，搬送物座標を状態sに代入し，Table 1のAのように報酬関数を設定した．10 に収束しているIIは，Iに対して報酬関数の設定Bに変更した状態であり，35 に収束しているIIIは，Table 1の1列目に記載したタスクを管理する変数をbool型で状態sに代入して，それぞれ学習を行った．IIIはすべてのタスクを完了し，IおよびIIは搬送物を1つ運ぶ程度で学習が終了した．このことから，報酬関数の重み付けがほとんど学習結果に寄与しなかったことがわかる．さらに，状態sは状況の変化を時々刻々と詳細に伝えるようなパラメータするのではなく，離散的に状態の変化を認識させるようにすることで，学習が改善されることが分かった．

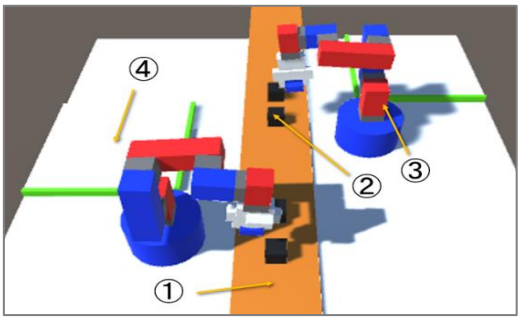

Fig. 2 Simulation of conveyor stopped

Table 1 Setting of the reward function

Target conveyed object	①・②		③・④	
	A	B	A	B
Grab the conveyed objects(+1)	1	1	13	19
	2	2	14	20
Lifting conveyed objects(+1)	3	3	15	21
	4	4	16	22
1st axis rotation(+1・+3)	5	7	17	25
	6	10	18	28
Place conveyed items(+1)	7	11	19	29
	8	12	20	30
Robot arm rise(+1)	9	13	Transport completed (+3・+1) 23・31	
	10	14		
1st axis rotation(+1・+2)	11	16		
	12	18		

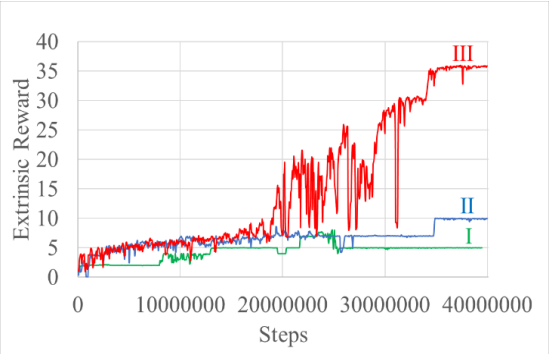

Fig. 3 Reward in simulation of conveyor stopped

参考文献

1) 北野郁弥，内方英利，松下光次郎，志賀元紀，佐々木実，”複数ロボットアームの協調動作における機構と制御の最適化”，日本AEM 学会誌Vol.29，No.1 (2021)，pp. 161-162

2) 中井 悦司，”ITエンジニアのための強化学習理論入門”，(2020) pp40-162