5ZD-06

Advantage 関数を用いたコントローラキャリブレーションによるユーザ好みのマッピングの獲得

長谷川麟太郎† 福地庸介‡ 奥岡耕平‡ 今井倫太†
慶應義塾大学理工学部† 慶應義塾大学理工学研究科‡

1 はじめに

多くのインタフェースの操作方法は，設計者があらかじめ決めた方法に依存しているため，ユーザは操作方法を学習し適応する必要がある．しかし，ユーザの操作の癖や障害といった身体的特徴により，設計者が設定した操作方法に適応しづらい場合もある．もしユーザの操作と操作の結果の適切なマッピングが可能ならば不適合の問題を解決することができる．

Li らの研究では，ロボットの動きを観察したユーザが，対応するジョイスティックの操作を入力し，半教師あり学習によってマッピングを学習する[1]．さらに人間がジョイスティックを操作する際のいくつかの前提を仮定することでマッピングの学習が効率化できることを実験で示した．また里形らの研究では，強化学習で得られるQ関数により計算される価値が高い結果をユーザが期待していると仮定し，価値の大きさに応じて「操作」と「操作の結果」のマッピングを学習する[2]．しかし，Li らの手法ではユーザがシステムを操作する前に手動でラベリングするため実際のプレイによるマッピングの学習ができない．また里形らの手法におけるQ関数は，直近の状態遷移によって得られる行動価値と遷移以降に期待される状態価値の和であり，両者を区別しない．状態価値のみが高い場合は，エージェントがどのような行動を取ってもQ関数の値は高く，ユーザにとって重要な場面でなくとも操作によるマッピングが大きく更新されてしまうため，マッピングの学習をうまく行うことができない．

本研究ではユーザがシステムの操作を学習するのではなく，システムがユーザの操作を学習することで，ユーザが自分好みの方法でシステムを扱えることを目指す．そこで強化学習で得られるAdvantage関数を用いたマッピングであるAdvantage Mapping及びマッピングのキャリブレーションのための場面抽出手法を提案する．Q関数と異なりAdvantage関数では行動価値のみを求めることができる．またマッピングの獲得のためユーザに特定の場面での操作を行わせることでキャリブレーションを行う．場面は Advantage 関数から自動的に選択されたものを利用する．

Controller calibration with advantage function for acquisition of user-preferred mappings.
Rintaro Hasegawa†, Yosuke Fukuchi‡, Kohei Okuoka‡, Michita Imai†
† Keio University Faculty of Science and Technology
‡ Keio University Graduate School of Science and Technology

2 Advantage Mapping

2.1 Advantage 関数によるマッピング

マッピングの学習はAdvantage関数$A(s,a)$によって行う．ここでsはエージェントが観測できる状態である．$A(s,a)$は行動価値関数$Q(s,a)$と状態価値関数$V(s)$の差分であり，状態の価値を含めない行動のみの価値を求めることができる．

$$A(s,a) = Q(s,a) - V(s) \quad (1)$$

本研究で提案するAdvantage Mappingはユーザの操作oが与えられた時，行動aを選択するマッピングの関数A_Mである．このマッピングの更新は以下のように行う．

$$\forall a, A_M(o_t,a) \leftarrow (1-\beta)A_M(o_t,a) + \beta A^{\pi}(s_t,a) \quad (2)$$

βはマッピングのための重みづけ係数であり，Q学習における学習率と同様の役割を果たす．提案手法ではAdvantage関数とマッピングの関数のバランスを取る役割を持つ．提案システムでは時間減衰するβを利用した．また[1]の研究を踏まえて，対称の位置のマッピングを行う．つまりコントローラ上でユーザの操作o_tと反対方向への操作をo'_t，エージェントの行動aと反対方向の行動をa'としたときに以下のような更新も行う．

$$\forall a, A_M(o'_t,a) \leftarrow (1-\beta)A_M(o'_t,a) + \beta A^{\pi}(s_t,a') \quad (3)$$

2.2 キャリブレーション

キャリブレーションには特定の方向の行動だけ価値が高く，人間にとって動くべき方向が明らかな場面が相応しい．つまり，エージェントが取れる行動の中で，1 つの行動の価値のみが際立って高い場面を抽出する．そのために，以下の式(4)から各行動の価値のエントロピーLを計算した．各行動の価値が均一なときはエントロピーが高く，ばらつきがある時はエントロピーが低くなる．提案手法ではLが 0.1 以下の時の場面を抽出してキャリブレーションに用いた．

$$L = \sum_{a} A(s,a) \log A(s,a) \qquad (4)$$

3 実験

3.1 実験で使用した環境・コントローラ

実験参加者は OpenAI が提供する roboschool pong という2次元のテーブルテニスの環境をプレイした．また参加者は図1に示すコントローラを用いて，テーブルテニスの板の操作を行った．コントローラは縦方向と横方向の4点の圧力の値を取得することができる．あまり馴染みのないコントローラでもマッピングを獲得できるか検証するため，このような形状のものを使用した．

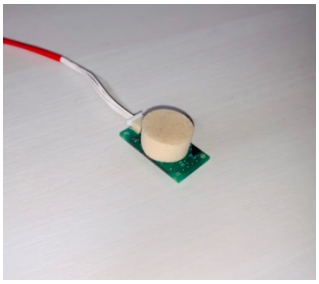

図1 参加者が使用したコントローラ

3.2 実験方法

実験ではマッピングを，プレイによる学習で獲得する条件1とキャリブレーションにより獲得する条件2で比較した．各条件でマッピングを学習後，獲得したマッピングにより本番のプレイを行なった．また評価として以下の5項目について5段階の評価をアンケートにより調査した．

Q1: スムーズに操作することができた
Q2: 思い通りの操作が実現できた
Q3: システムがやりたいことを推測してくれた
Q4: 操作方法を把握することができた
Q5: 操作方法がわかりやすかった

さらに実際のゲームの場面について，Advantage 関数で一番価値が高い行動と，実際のユーザの入力の方向を記録した．

3.3 実験結果

図2に実験のアンケート結果を示す．各群の平均値の比較には，対応のある t 検定を用いた．結果として Q1, Q2, Q3 で有意差が見られた．条件 1 では Q4, Q5 の項目が高いにもかかわらず，Q1, Q2 の項目が低いという結果になっている．これはマッピングの操作方法を理解することはできたが，把握した操作方法は自分のやりたい操作とは異なっていたということを意味する．対して条件2では Q1 から Q5 の項目が総じて高いことから，マッピングの操作方法が思い通りの操作であったことを意味している．図3, 4 はゲームの場面について押すべき方向と実際に押された方向についてのヒートマップである．条件2では対角成分のみが高いのに対し，条件1では押された方向にばらつきがあり，キャリブレーションがマッピングの獲得に有用であったことがわかる．

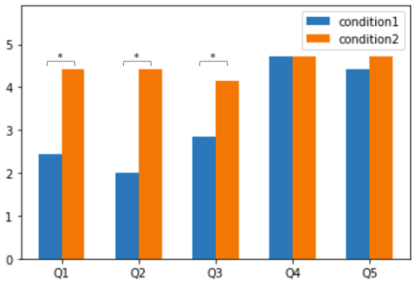

図2 提案システムの評価結果

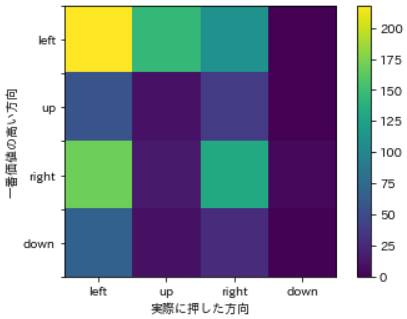

図3 条件1で実際に押された方向と一番価値の高い方向の比較

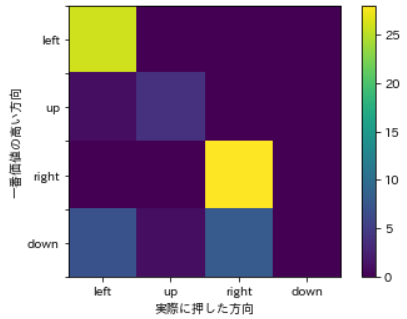

図4 条件2で実際に押された方向と一番価値の高い方向の比較

5 おわりに

本研究では，キャリブレーションによりユーザが思い通りのマッピングを獲得できることを示した．提案手法では自動的にβを決定することができないため．今後の課題として，動的にパラメータを決定する手法を検討していく．

参考文献

[1] Mengxi Li et al. (2020)“Learning User-Preferred Mappings for Intuitive Robot Control” IROS
[2] 里形 et al. (2020) “Q-Mapping: 行動価値関数を利用したユーザ操作に対する解釈の適応的獲得”. JSAI