マルチモーダル情報を用いた強化学習による行動知識獲得への取り組み

理学専攻・情報科学コース　恒川英里

1　はじめに

将来，人とロボットが共に暮らし，高齢者や居住者の支援をすることが考えられる．しかし，人の生活は常に変化しており，ロボットはそのような変化にも対応出来る必要がある．このような背景から，本研究ではロボットに経験から正しい行動の学習を可能にさせる強化学習を用いることによって，行動に関する概念を獲得させることを目的とする．まず，色付きの物体を積むという行動を，次に色付きの物体を適切な場所に置くという片付け行動を学習させる．しかし，本研究で用いている Q 学習は状態を離散的に扱うため，連続した空間の中で，状態を適切に区切ることが難しいという問題を持っている．そこで，片付け行動を通して，報酬を得た行動データから多層マルチモーダル LDA(mMLDA)[3] を用いて，適切な状態設定を獲得することを行う．

2　物体積載行動知識の獲得

2.1　作業課題

ロボットの持つ，ハンドカメラからテーブル上に置かれている色付きの物体の画像を取得する．物体の認識は，画像処理ライブラリ OpenCV を用いて色認識と領域抽出から行う．物体の把持は，得られた画像から重心を計算し，手を移動させることによって行う．用いているロボットは (株) 川田工業社製ヒューマノイドロボット HIRO である．作業課題として，HIRO はテーブル上の物体を把持した後，決められた場所に正解順番に積むという行動を学習する．物体を取得した際に，最終的な正解の順番と比較し，相当する報酬を得る．図 1 に作業課題の概観を示す．

図 1: 作業課題の概観

2.2　強化学習への定式化

2.2.1　Q 学習

強化学習手法は逐次報酬を得られる Q 学習 [1] を用いている．報酬が得られた際，その行動の価値を式 (1) を用いて計算し，更新を行う．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha[r_{t+1}+\gamma \max_a Q(s_{t+1},a)-Q(s_t,a_t)] \quad (1)$$

上式において，s は状況，a は行動，r_t は時刻 t における報酬，$Q(s,a)$ は累積報酬 $E\{R_t|s_t=s,a_t=a\}$ で表現される行動価値を表し，α は学習率，γ は割引率を表す．

2.2.2　定式化

課題に対し，定式化を行った．その概要を図 2 に示す．

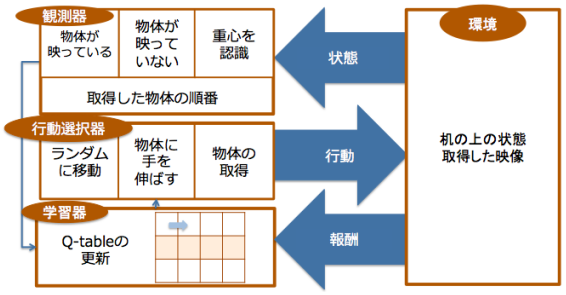

図 2: 定式化概要

2.3　実験

作業課題に対し，定式化した枠組みを用いて実験を行った．シミュレータを用いた実験では Q 学習において，エピソード回数に対する報酬の変化を見ることにより，収束状態を確認した (図 3)．約 20 回ほど学習

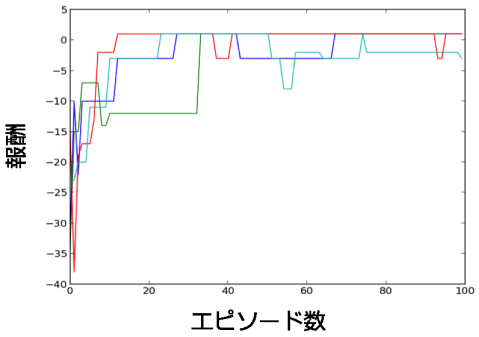

図 3: エピソード回数における報酬の遷移状況

させた辺りから異なる結果が出ることもあるが，正解が現れ始め，およそ収束し始めることがわかる．画像処理に基づき色付きの物体を希望する順番で取得する行動知識を，Q 学習を通じて獲得出来ることをシミュレータ上で確認したのち，実際に実機に適用した．

指定した色を探し出し，把持，物体を探索する範囲の外に移動させることができた．

2.4　結果と考察

画像処理に基づく物体取得プログラムについて，シミュレータでの試行と同様に，指定した色の物体を見つけ，把持し，その記録を配列に収めることができた．強化学習アルゴリズムについては，報酬の与え方が逐次的であったため，より正解を導き出すのが速かったと考えられる．

3　片付け行動知識の獲得

未知な課題に対する行動を考慮した時，多くの観測情報を持つことにより，その状況に対して，適切な行動が出来ると考えられる．また，経験が蓄積されると，それまで観測した情報全てが必要でない場合があり，観測した情報を全て用いると，学習に時間がかかるというデメリットが発生する．そこで，Q 学習を効率良く学習するため，mMLDA を用いて課題達成のための状態空間を適切な次元数に削減しながら学習させることに取り組む．

3.1　作業課題

机の上を 4 × 4 のグリッドとし，色は赤と青，形は丸か四角の物体に対して，赤い色の物体は「右下」，青い色の物体は「左上」に置くことを学習させる．課題の概要を図 4 に示す．

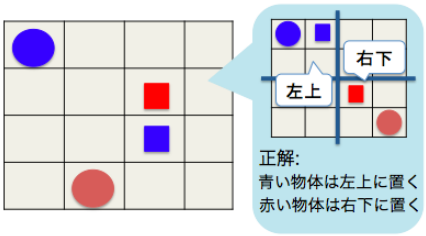

図 4: 片付け課題概要

3.2　適切な状態設定獲得への取り組み

3.2.1　mMLDA を用いた状態数削減

mMLDA とはマルチモーダル LDA（MLDA）[2] を多層化したモデルであり，このモデルの統合概念は教師なし学習によって学習することが出来る [3]．今回，下位概念として使用するカテゴリは，物体，場所，行動の 3 つである．まず，Q 学習を用いて，報酬の得られた状態，場所，行動情報を収集し，そのデータを用いて mMLDA を学習させる．分類結果を元のデータに適用し，報酬が得られた行動を予測できているかを確認する．そして，一番正解率の高かった分類結果を Q 学習の枠組みに適用し，学習を再度行う．図 5 に作業の概観を示す．

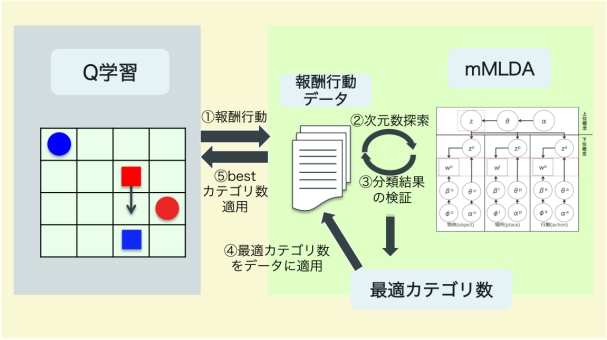

図 5: mMLDA を用いた適切な状態設定獲得

3.2.2　Q 学習の設定

状態，行動，報酬を以下のように設定する．後ろに次元数を記す．

- 状態 : 円形度 (10)，rgb 値 (10^3)，物体の面積 (10)，場所（机の上の座標）(16)，動き (3)
- 行動 : (座標に) 掴んで置く，押して置く，そのまま
- 報酬 : 正しい場所に置かれたら正の報酬（10）

3.3　実験

Q 学習を行いながら，報酬の得られた行動のデータを収集し，適切な次元数を探索する．Q 学習を 100 エピソード，を 10 回行い，報酬の得られた行動 715 回分を用いて実験を行った．今回は人手で正しいと感じる 2 種類の設定を行い，学習させた．実験 1 では物体に合わせて，物体カテゴリ 4 つ，行動カテゴリ 3 つ，場所カテゴリ 8 つと設定し，実験 2 では報酬に合わせて，物体カテゴリ 2 つ，行動カテゴリ 3 つ，場所カテゴリ 2 つに設定した．結果を表 1 に示す．値は予測精度の確率である．

表 1: 実験結果

カテゴリ	実験 1	実験 2
物体	0.51	0.95
行動	0.44	0.40
場所	0.28	0.52
適切な報酬行動	0.62	0.51

3.4　結果と考察

実験 1 と 2 を比較すると，物体カテゴリと，行動カテゴリでは，実験 2 の方が良い数値が出ている．これは mMLDA が他のカテゴリ情報も考慮して学習をしているためと考えられる．しかし，適切な報酬行動を予測できているのは実験 1 である．報酬行動を一番高く獲得するカテゴリ数の探索が必要である．

4　おわりに

本研究では，ロボットが強化学習を用いて未知な課題を解決することを目的に取り組んできた．実機を使った実験では，単純な課題ではあるが，強化学習を用いて物体を積むという行動を獲得することが出来た．片付け行動の学習に関しては，多次元の状態から mMLDA を用いて次元を削減することが出来た．今後最適なカテゴリ数を探索し，どれだけ学習が効率的になるかさらなる考察が必要である．

参考文献

[1] Watkins, C.J.C.H.,Learning from Delayed Rewards. PhD thesis, Cambridge University, Cambridge, England. 1989.

[2] T. Nakamura et al., Grounding of Word Meanings in Multimodal Concepts Using LDA, in Proc. of IROS 2009, pp.3943-3948, 2009.

[3] アッタミミ，ムハンマド，阿部，中村，船越，長井，多層マルチモーダル LDA を用いた人の動きと物体の統合概念の形成，日本ロボット学会誌，Vol.32，no.8，pp89-100，2014.