画像認識に基づくロボットの行動を制御する強化学習の取組み

恒川英里　(指導教員：小林 一郎)

1　はじめに

近い将来、家庭にロボットが導入され高齢者の支援や居住者の生活を支援することが予想される．その際，ロボットが現実世界に存在する様々な課題をこなす必要がある．このことから本研究では，ロボットが現実世界において視覚から得た情報を用いて自らの適切な行動を獲得する強化学習の枠組みについて考察する．

具体例として，テーブル上に置かれた物体を色によって決められた順番に従うように取得するという行動知識をヒューマノイドロボットを使って実現することに取り組む．

2　ロボットの行動知識獲得

2.1　作業課題

使用するロボットは (株) 川田工業社製ヒューマノイドロボット HIRO を用いる．取り付けられたハンドカメラを用いて，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．画像中の物体に対して，画像処理ライブラリ OpenCV を用いた色認識および領域抽出による物体の認識を行う．HIRO はテーブル上の物体を取って来た際に，正解となる順番と比較し相当する報酬を獲得し，最終的に正解の順番にとってくるという行動知識を獲得する．図 1 に作業課題の概観を示す．

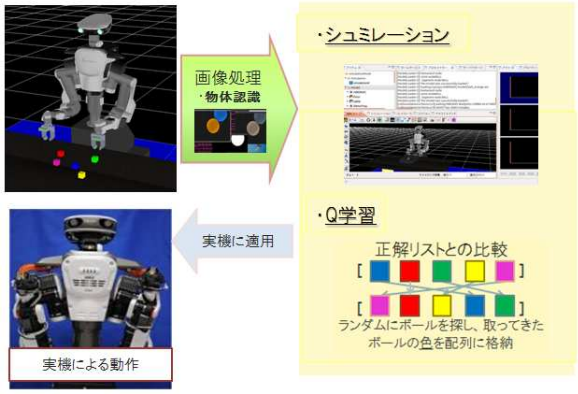

図 1: 作業課題の概観

2.2　画像処理による物体の位置推定

HIRO は備え付けられたハンドカメラにより画像の取得を行い，色認識，二値化処理および輪郭抽出処理を行う．それにより物体の座標推定を行い (図 2)，その座標に手を移動し物体の把持を行う．

2.3　強化学習への定式化

2.3.1　Q 学習

本研究では，強化学習の枠組みにおいて最適な行動を学習する Q 学習 [1] を用いることにより HIRO の行動知識を獲得する．Q 学習による行動価値の更新は，

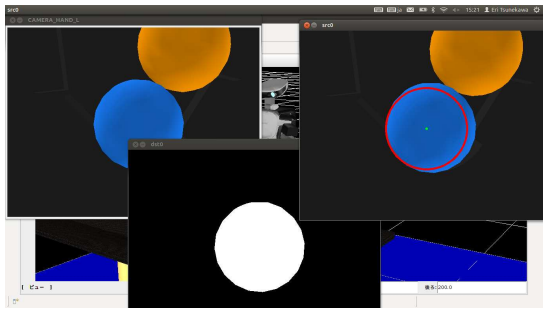

図 2: HIRO による画像処理

式 (1) によって示される．

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

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 は将来の報酬に対する割引率を表す．

2.3.2　状態

HIRO が認識する状態は，カメラから得られる視覚情報と物体の取得状態の二つから構成される．ただし，色によって状態が異なることとするため，以下に示す 3 つとする．

- 画像中に指定された色の物体が映っている
- 画像中に物体が映っていない
- これまでに取得した物体の順番

2.3.3　行動

HIRO による物体取得の行動においては，まず，カメラに物体が映っていない場合は，物体を探すために適当な範囲で手を動かす動作が必要になる．また，カメラに物体の一部が映しだされた場合 (本研究では物体を色で認識しているため，正確にはその色が認識された場合)，物体を把持できる位置に手を動かす動作を行う，最後にその状態において対象となる物体を取得する /しないの行動が選ばれる．このことから本課題において対象となる動作は以下の 3 つとなる．

- ランダムに手を動かす
- 物体が映ったらその方向に手を動かし，物体の重心を捉える．
- 物体を掴む

2.3.4　報酬

互いに異なる色の物体がテーブル上に n 個存在し，それを取ってきた際の報酬は予め決められた順番との差異がペナルティとして与えられる．上記のペナルティによって物体を正しい順番に整列するための工夫として，配色に対して異なる価値を付与することを考える．いま，取得したい順番が [青, 緑, 橙, 白, 黒] とし，リストの右にいく (順番が遅くなる) につれて，その価

値が小さくなっているとする．ここでは，例えば，価値を青: 5，緑: 4，橙: 3，白: 2，黒: 1 と設定する．評価は物体が取得される毎に行われ，それぞれの試行において取得した物体の色と正解順番との差分をペナルティして掛け，報酬とする．例えば，[黒, 緑, 青, 橙, 白] という順番を取得した時その時刻の，報酬は 5 × (-2)，4 × 0，3 × (-1)，2 × (-1)，1 × (-1) として与えられる．

3 実験

3.1 画像処理に基づく物体取得

作業課題に従って，机の上に無造作に置いてある物体を選択された色に合わせてランダムもしくは greedy 選択によって探し出し，物体を把持した後，指定した場所に物体を移動させ，併せて，獲得した物体の色を記録する．図 3 に画像処理に基づく物体取得の様子を示す．

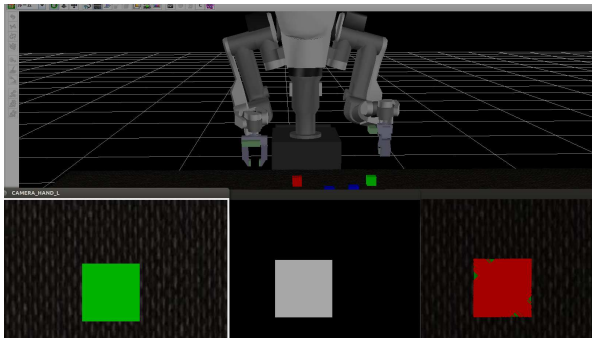

図 3: 画像処理に基づく物体取得の様子

3.2 Q 学習による行動知識獲得

作業課題を達成する Q 学習を実装し，課題に対する検討を行った．作業課題は，予め決まった順番に 5 色を並べるという知識を獲得するというものであり，その状態からの逸脱度をペナルティとした報酬を与える．状態および行動の表現形式は共に同形であり，(取得順番，色) として与えられる．また，現在の状態は，過去に取得した物体の情報を保有しているとする．エージェントの行動選択方法として，ϵ-greedy 選択を用いた．

3.3 シュミレータを用いた実験

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，学習の収束状態を確認した (図 4)．図 4 より，約 20 回ほど学習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 学習を通じて獲得できることをシュミレータ上で確認した．

3.4 実機を用いた実験

シュミレータ上で動かしたものを実際に実機に適用した (図 5)．指定した色を探し出し，把持し，物体を探索する範囲の外に移動させることができた．

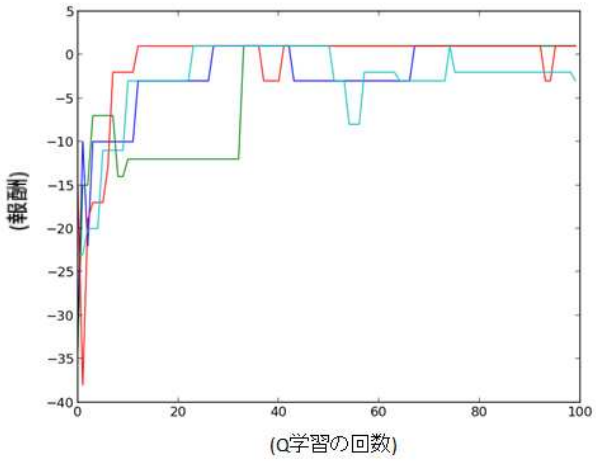

図 4: 最終結果の評価値グラフ

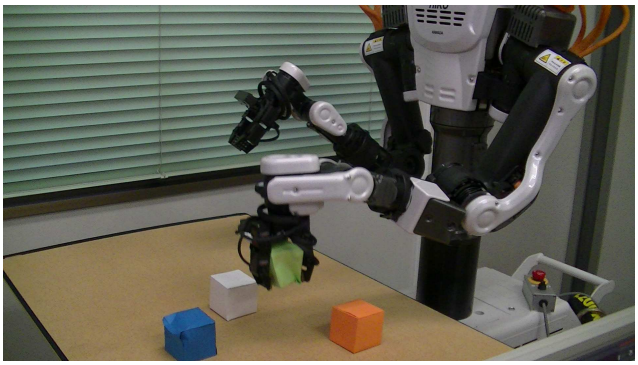

図 5: 実機での実験の様子

3.5 実験結果と考察

画像処理に基づく物体取得プログラムについて，シュミ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正解を導き出すのが速かったと考えられる．

4 おわりに

本研究においては，画像認識に基づくヒューマノイドロボット HIRO の行動知識を強化学習を用いて行うための基礎的な実験を行った．具体的には，HIRO に備え付けられたカメラから得られた画像中に映る物体の領域および色の認識を行い，物体を把持した．今後の課題として，ロボットの制御部分を拡張すること，画像から行うようにすること，及び，より現実世界を考慮した少ない情報の中での学習，例えば，物体を上に積み，高さを報酬にする等，さらに，学習回数が少なくても正解を導くことのできる，強化学習の新しい枠組みについて検討し，改良を加えていきたい．

参考文献

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

[2] 浅田稔，野田彰一，俵積田健，細田耕，視覚に基づく強化学習によるロボットの行動獲得，日本ロボット学会誌，Vol.13, No.1, pp.68〜74, 1995.