The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

人から評価を得てリアルタイムに強化学習する移動ロボット

A mobile robot based on real-time reinforcement learning using human evaluation

山根 健 *1　植月 宏昌　横松 秀康 *2
Ken Yamane　Hiromasa Uetsuki　Hideyasu Yokomatsu

*1帝京大学　*2株式会社タイトー
Teikyo University　Taito Corporation

In develping intelligent robot systems, it is necessary for designers to make robots by fine-tuning parameters and modifying programs in detail. However, this conventional method is a big burden not only for designers but also general users. In this paper, we present a new design method based on reinforcement learning in real time through human-robot interaction. We also develop an autonomous mobile robot system and show its possibilities.

1. はじめに

知能ロボット開発において，一般的には，ロボットのために環境を整えて，パラメータやプログラムを細かく調整するなど設計者がシステムを作り込む．しかし，作り込みには大きな労力が必要であるし，実環境では想定外のことが頻繁に起こる．また，非技術者である一般ユーザがロボットの行動を望む方向へ変更したい場合，いわゆるプログラミングによって機能を修正したり追加したりすることはできない．

これらに対して，我々は人との相互作用を通じてロボット自らが行動を学習する方法を検討している．例えば，ユーザがロボットへ行動を教示したり，ロボットの行動に対してその良し悪しを評価したりする中で，ロボットがリアルタイムに学習して必要な機能を獲得する方法である．これは設計者やユーザにとっても負担の少ない設計方法として期待される．

その実現のために，強化学習 [Sutton 98] の枠組みを利用できると考えている．しかし，実環境において強化学習を行う場合，膨大な学習時間が必要であるなど幾つかの大きな問題がある．これに関して，強化学習の価値関数近似器として選択的不感化ニューラルネット（SDNN）を用いる方法が提案され，単純なアルゴリズムである Q 学習でも従来手法に比べて効率よく学習できることが示された [新保 10]．しかし，実環境で強化学習するための具体的な方法や人という不確定な要素を系に含んでも安定して学習できるかについては明らかではない．

そこで本研究では，SDNN を価値関数近似器として用いて強化学習するロボットを実際に構築し，人との相互作用の中で学習可能か検証する．具体的には，事前知識をほとんど与えずに，ロボットの行動に対する良し悪しの評価を人から得て，ロボットが適切な行動をリアルタイムに学習可能か，ウェイポイント経路を追従させる課題において調べる．

2. 方法

2.1 ハードウェアとセンサ情報の前処理

使用するロボット，測域センサおよび評価用ボタンを図 1 に示す．ロボットは独立 2 輪駆動方式であり，市販のノートパソコンにより走行制御される．なお，安全のために，緊急停止スイッチを取り付け，最大速度を 0.15m/s に制限する．

連絡先: 山根健，帝京大学理工学部情報電子工学科，〒 320-8551 栃木県宇都宮市豊郷台 1-1，028-627-7224，yamane@ics.teikyo-u.ac.jp

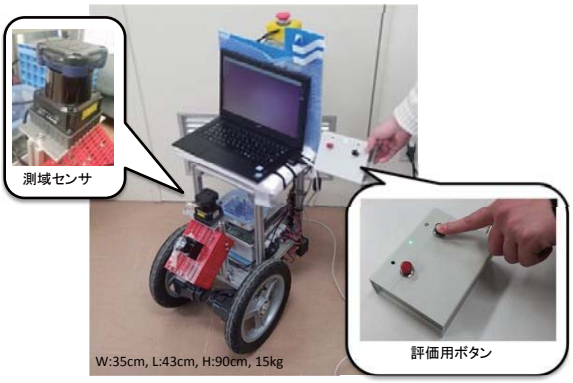

図 1: ハードウェアの構成

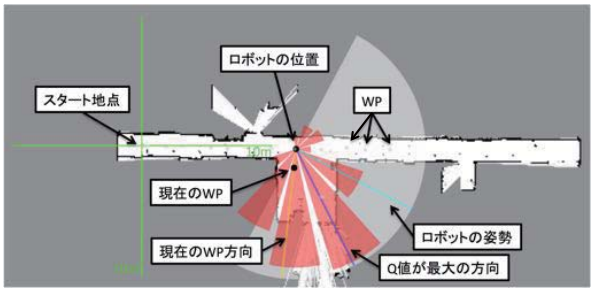

図 2: 実験コースと可視化したロボットの状態

環境情報を取得するため，測域センサ（UTM-30LX-EW，北陽電機）を用いて床面から 0.49m の高さにおける周囲 ±135° の範囲で 1080 点の距離情報を取得し，前処理して状態変数とする．具体的には，5m を超えるものに関して 5m とし，30° ずつの範囲でグループに分け（120 点ずつ 9 つに分け），それぞれの範囲における平均値を求める．次に，最大値 5m で割ることで 0〜1 の値に正規化し，これらを x_1〜x_9 と表す．また，推定したウェイポイント方向とロボットの姿勢の差（±90°）を 0〜1 の範囲で正規化したものを x_{10} として状態変数に含める．これら状態変数の構成イメージを図 2 に示す．

2.2 学習方法

ロボットの学習方法を図 3 に示す．ここでは強化学習アルゴリズムとして Q 学習を，価値関数近似器として SDNN[新保 10, 小林 15] を用いる．SDNN は前述した入力変数 x_1〜x_{10} を受

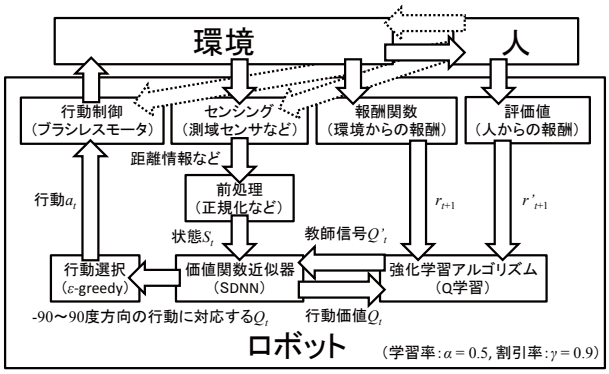

図 3: 学習方法

けて，ロボットの現在の姿勢から $-90°$〜$90°$ 方向へ直進または旋回する行動の価値（Q 値）を計算する．この連続的な Q 値を表現するため，小林らが提案する連続状態行動空間における学習に対応した SDNN[小林 15] を用いて，181 個の行動（1° 刻みの行動）に対する Q 値を出力する．

行動選択では，最大の Q 値に対応する行動を選択する．なお，学習初期のみある一定の確率でランダムに行動を選択すると，連続状態行動空間を効率よく探索できる．

行動実行後，予め設計した報酬関数に基づいて環境から報酬が与えられる．加えて，図 1 に示す評価用ボタンを用いて，人から行動の良し悪しの評価を受ける．2 つのボタンはそれぞれ正の評価（良い）と負の評価（悪い）に対応し，ロボットの行動に対して人が自由なタイミングでボタンを押すことができる（押さなくてもよい）．このボタン押しの頻度から一次元的な値である評価値が計算されて報酬として与えられる．

最終的に，Q 学習の更新式で得られた新しい Q 値（図 3 の Q'_t）を教師信号として SDNN を 1 回だけ学習する．なお，学習には誤り訂正学習を用いる．以上の処理を行動毎に繰り返す．

3. 実験

実際にロボットシステムを構築して，複数のウェイポイント（WP）を順番に繰り返し通過させる課題を行った．具体的には，図 2 に示す 1 周およそ 46.5m の屋内コースにおいて，事前知識なしの状態でスタート地点にロボットを置き，人の評価がある条件とない条件でそれぞれ 4000 step まで行動させた．

このとき，Q 学習のパラメータについて，学習率 $\alpha = 0.5$，割引率 $\gamma = 0.9$ とし，ランダム行動の確率については 1000 step まで 20%，さらに 1000 step まで 10%，その後は 0%とした．報酬関数について，障害物との距離が 0.55m 未満になった場合に-7，WP から離れる場合に-3，ロボットの姿勢と WP 方向の差が 5.6° 未満の場合に 2 を与える．さらに，SDNN を構成する素子数について，入力層 1,000 個，中間層 9,000 個，出力層 3,000 個とし，出力層において -30〜30 までの Q 値を表現した．また，学習係数 $c = 0.01$ とした．

それぞれの条件で 3 回ずつ実験を行った，コース 1 周にかかるステップ数の変化を図 4 に示す．評価の有無に関係なく学習が進むとステップ数が減少した．また，2 周目以降は「評価あり」の方が「評価なし」よりも少ない結果が得られた．

「評価あり」における報酬について見ると，環境から得られた報酬（予め設計した報酬関数から得られた報酬）については，スタート地点，道幅が狭いところや障害物付近などで負の報酬が多かった．一方，人の評価から得られた報酬については，それらの付近でも正の報酬が与えられており，全体として

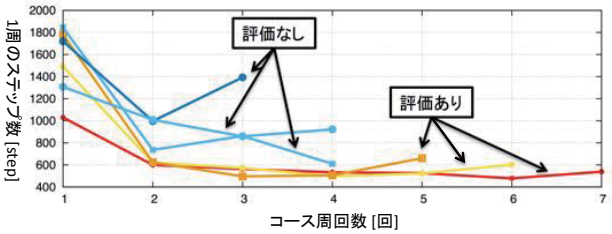

図 4: 1 周にかかるステップ数（4000 step まで）

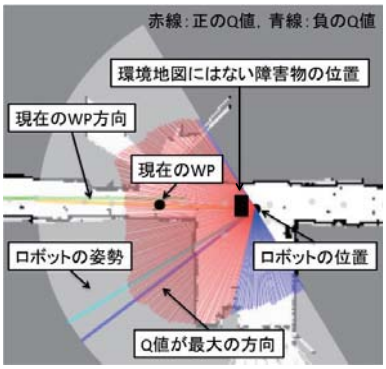

図 5: 障害物を回避しているときの Q 値の構成

も正の報酬が多かった．そこで両者を比較したところ，報酬の正負について 79%が一致していなかった．また，壁に向かう行動に対して人が良い評価をするなど，人が非協力的な条件ではまったく課題を遂行できなかった．これらから報酬関数では対応できない部分を人がカバーしていたと考えられる．

詳しく見るために，障害物付近での Q 値を図 5 に示す．WP 経路追従が本来の目的であるから，現在の WP 方向へ向かう Q 値が学習により高くなると予想される．しかし，実際には WP 方向の Q 値は抑えられており，障害物を回避する方向の Q 値が最も高くなっている．このような Q 値の構成は曲がり角や道幅が狭いところでも見られ，人の評価に基づいて学習する中で簡単な障害物回避機能が獲得できた．

4. おわりに

本研究では，人の評価を利用して強化学習するロボットを示した．詳しい分析は課題として残されているが，人を含む系でもリアルタイムに学習でき，人が協力的な条件では学習効率が高いことや簡単な障害物回避も可能であることがわかった．

今後の課題として，人の評価の効果について調べることや複数の行動次元（例えば，進む方向と速度の組み合わせなど）を扱えるようにシステムを拡張することなどが挙げられる．

参考文献

[小林 15] 小林高彰, 渋谷長史, 森田昌彦:選択的不感化ニューラルネットを用いた連続状態行動空間における Q 学習, 信学論 D, Vol.J98-D, No.2, pp.287-299 (2015).

[新保 10] 新保智之, 山根健, 田中文英, 森田昌彦:選択的不感化ニューラルネットを用いた強化学習の価値関数近似, 信学論 D, Vo.J93-D, No.6, pp.837-847 (2010).

[Sutton 98] Sutton, R. S. and Barto, A. G.:Reinforcement Learning, MIT Press (1998), (邦訳：強化学習, 三上貞芳, 皆川雅章 訳, 森北出版 (2000)).