知識選択型転移強化学習を用いた移動ロボットによる動的障害物回避

高矢　空*　　河野　仁　　須賀　哉斗　　鳥谷部　悠希（東京電機大学）
池　勇勳（北陸先端科学技術大学院大学）　　藤井　浩光（千葉工業大学）
鈴木　剛（東京電機大学）

Dynamic Obstacle Avoidance by Mobile Robot Using Transfer Reinforcement Learning with Knowledge Selection

Takaya Sora*, Kono Hitoshi, Suga Kanato, Toriyabe Yuki, (Tokyo Denki University)
Ji Yonghoon, (Japan Advanced Institute of Science and Technology)
Fujii Hiromitsu, (Chiba Institute of Technology),
Suzuki Tsuyoshi, (Tokyo Denki University)

In recent years, machine learning technologies have been actively implemented in society. Especially technologies such as reinforcement learning and transfer learning are being implemented in intelligent robot systems. The authors have proposed knowledge-selective transfer reinforcement learning based on the spreading activation model, which is a knowledge in the field of cognitive science. In past research, although static obstacle avoidance has been achieved in mobile robots with knowledge-selective transfer reinforcement learning, dynamic obstacle avoidance has not been investigated. In this paper, it is realized that dynamic obstacle avoidance by tuning the hyper-parameters of transfer reinforcement learning with knowledge selection.

キーワード：転移学習，強化学習，知識選択，活性化拡散モデル
(Transfer learning, reinforcement learning, knowledge selection, spreading activation model)

1.　緒言

近年，自動運転技術の進展により，利用者が運転の負担から解放され，交通の安全性と効率性が向上する可能性が高まりつつある．自動運転車は機械学習技術を駆使して，リアルタイムの状況判断や障害物検知，交通ルールの遵守などを行い，自律的な運転を実現する．しかし，自動運転車が直面する課題の一つとして，動的な障害物の回避がある．現行の自動運転技術では静的な障害物への対応は従来の経路計画手法などで適応可能であるが，動的な障害物の予測と回避は依然として課題とされている．このような課題に対して著者らは，知識選択型転移強化学習を用いた移動ロボットにおける動的障害物回避の実現を提案している(1) (2).

著者らの従来の研究においても同様に自動運転シニアカーにおける静止障害物の回避は実現しているが，動的障害部の回避は実現出来ていない．そこで，本研究では，知識選択型転移強化学習を用いた移動ロボットによる動的障害物回避について基礎的な検討と実験を行ったので報告する．知識選択型転移強化学習のハイパーパラメータを調整することで，動的障害物の回避が実現できることを確認し，知識選択型転移強化学習は，過去の学習経験を転移させることで学習速度の向上や新しい環境への適応度の向上を図る手法であることを示す．

2.　提案手法

〈2･1〉　強化学習と転移学習

本研究では強化学習に多くの研究で用いられている Q 学習を用いる(3)．強化学習は，エージェントが試行錯誤的に環境を探査し，得られた報酬情報を基に報酬を最大化するような行動を学習するアルゴリズムである．そこで獲得される情報はエージェントが観測可能な状態に対する行動の行動価値であり，それらの情報が格納されているものを方策や行動価値関数などと呼ぶ．近年では，方策や行動価値関数を他のエージェントや他のタスクで再利用する手法が議論されており，転移学習と呼ばれる手法である．本研究では強化学習における転移学習を転移強化学習と呼び，Taylor らの転移強化学習手法（価値関数転移）を用いる(4)．価値関数転移は，転移元である Source task にて強化学習した行動価値関数を転移先である Target task のエージェントが再利用して再学習を行い，Target task という新たな環境やタスクに行動価値関数を適合させる学習を行う．これにより，Target task における学習速度の改善や，学習初期からの高パフォーマ

ンスを実現する．

〈2･2〉　知識選択手法 SAP-net

SAP-net (Spreading Activation Policy Network)とは，Kono らが開発した知識選択型の転移学習手法である(1)．転移学習前の手続きとして，予め学習した複数の強化学習の方策や行動価値関数をそれらの接続関係や距離を定義したグラフで構成し保存する（図 1）．また，それらの方策や行動価値関数には活性値というパラメータを付与されている．強化学習エージェントやロボットにおけるセンサ入力などの刺激が入力されると，その刺激に対応した方策や行動価値関数の活性値が活性化する．活性化した値はグラフ上で接続された他の方策や行動価値関数にも拡散される．これにより，各方策や行動価値関数の活性値を閾値で選択し，これを想起と呼ぶ．想起された方策や行動価値関数を基に強化学習エージェントやロボットは転移学習を行う．さらに，活性値は時間に比例して減衰するプロセスも SAP-net に実装されており，時間経過に対する活性値の累積を抑制する効果がある．

〈2･3〉　ハイパーパラメータチューニング

先述の SAP-net(Spreading Activation Policy Network)では，移動ロボットに実装されたケースにおいて，移動ロボットが静的障害物を回避するような検証はなされている(2)．しかし，本研究では SAP-net にグラフ構造を調整し，刺激となる入力値を調整することで，動的障害物の回避も可能になることを確認する．

本研究では，知識選択型の転移学習モデルにより，強化学習した知識を選択し，ハイパーパラメータを調整することで動的障害物をよける手法の提案を行う．

3.　知識選択実験

〈3･1〉実験目的・条件

本提案手法を用いて移動ロボットによる動的障害物回避の実現を評価するために物理演算シミュレータ内に構築した．物理演算シミュレータには Cyberbotics 社製の Webots 2023a を使用し，図 2 のような環境を構築する．ロボットは初期座標として図 2 のように配置され，障害物配置は計 5 種類の環境でそれぞれ強化学習を行い，これが Source task となる．障害物の先にはゴールエリアが設定され，報酬がロボットに与えられる．また，障害物や壁に接近すると負の報酬が与えられるように設定されている．ロボットの行動は直進と後退，右旋回，左旋回の 4 種類の行動を選択可能となっている．

本実験における SAP-net の構成としては，Source task で獲得した 5 つの行動価値関数に，ロボットの初期座標から見た障害物の方向と距離$(r,\ \theta)$をラベルとして付しておき，各知識の持つ角度と距離をベクトルとする．例えば，行動価値関数 1 の持つ角度をθ_1，距離をr_1と置き，行動価値関数 2 の持つ角度をθ_2，距離をr_2と置きいたとき二つの距離ベクト

本実験における SAP-net の構成としては，Source task

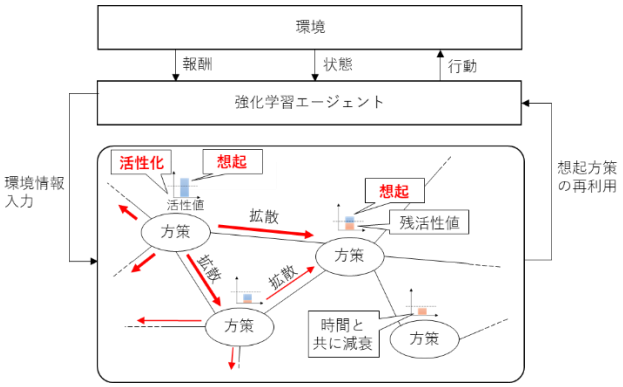

図 1　SAP-net の概念図

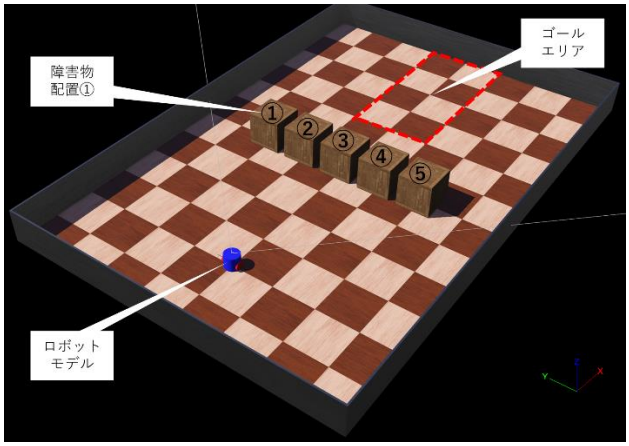

図 2　構築した Webots の環境

で獲得した 5 つの行動価値関数に，ロボットの初期座標から見た障害物の方向と距離$(r,\ \theta)$をラベルとして付しておき，各知識の持つ角度と距離をベクトルとする．例えば，行動価値関数 1 の持つ角度をθ_1，距離をr_1と置き，行動価値関数 2 の持つ角度をθ_2，距離をr_2と置きいたとき二つの距離ベクトルのユークリッド距離を求める．これにより，各行動価値関数間のグラフ上における距離を算出する．さらに，各行動価値関数の持つ角度と距離と，エージェントが行動中に取得する障害物までの角度と距離を比較することで活性値の活性化を行う値として用いる．さらに求めたユークリッド距離の最大値を使用して正規化を行う．ユークリッド距離の最大値をD_{max}とする．ある行動価値関数間の距離をDと置く．次式により正規化距離D_{norm}を求めることで行動価値関数も持つラベル間距離を 0 から 1 の範囲に収めることが出来る．

$$D_{norm} = \frac{D}{D_{max}} \quad \cdots\cdots(1)$$

このラベル間距離の正規化は，異なるスケールの距離値を比較しやすくしている．また，これらの計算値がどのような値であるかを比較しやすくするために，図 3 のようなヒートマップを作製した．なお，図 3 では図 2 における①から⑤までの障害物配置で学習した行動価値関数のラベルとしてユーザインタフェース側で認識しやすいよう次のような表示ラベルを用いている．

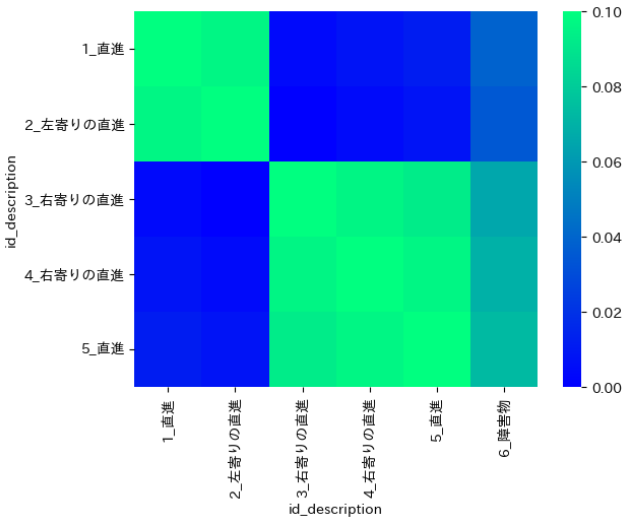

図 3 学習済行動価値関数のラベル間類似度ヒートマップ

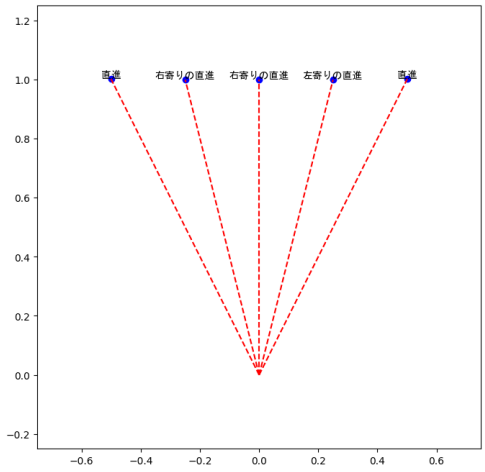

図 4 知識ベクトルの類似度ネットワーク

- 障害物配置①：直進
- 障害物配置②：右寄りの直進
- 障害物配置③：右寄りの直進
- 障害物配置④；左寄りの直進
- 障害物配置⑤：直進

図 3 に示したヒートマップを構成するラベルなどの具体的な構成やシステムの流れとしては以下の通りである．各行動価値関数でクロス表を作りそれらの持つベクトルのユークリッド距離を正規化した値でヒートマップを作成している．そのため，0.1 に近ければ近いほど類似度が高く，0.0 に近ければ近いほど類似度が低いといえる．

次に，各知識の類似度を求めるため，距離と角度の点 P 群を二次元マップにプロットした．次に，プロットした知識のベクトルとエージェントを線で結ぶことで，障害物への距離を図 4 のように可視化した．

図 4 に示した類似度ネットワークでは，強化学習で得た行動価値関数の位置と行動のラベルが 1 つの図にまとまっ

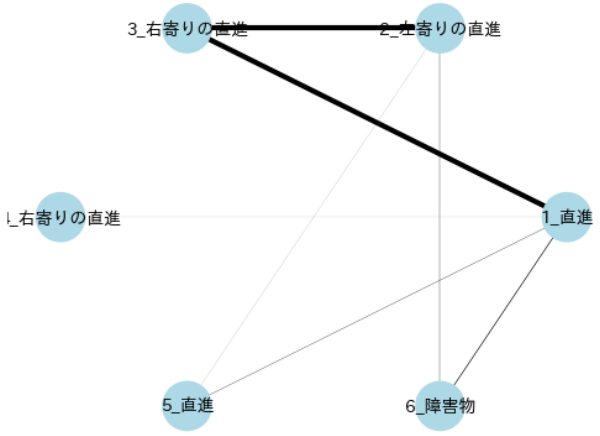

図 5 実験に使用した SAP-net

ている．次に，図 5 のようなネットワーク図も作成した．このネットワーク図では，知識（行動価値関数）間のつながりや活性化の際の値の伝搬がどのように行われていくのかが一目で分かるようになっている．よって，つながりの太さは類似度の高さに比例するため，活性値も高くなる．

これらのシステムを用いて Target task では，図 2 に示した障害物は障害物配置の①から⑤に移動しながらロボットがゴールに向かって移動する実験を行い，動的障害物配置の回避が発現するような知識選択がなされるか評価する．Target task ではロボットが 10 行動毎にセンサから障害物までの方向と距離$(r,\ \theta)$を取得し，SAP-net による知識選択を行い選択された行動価値関数による転移学習で行動を行う．本実験では，提案システムの基礎的検討であるため，Target task を 1 エピソード分実行する．

Source task と Target task 共に，強化学習のパラメータとして学習率$\alpha = 0.5$，割引率$\gamma = 0.9$，行動選択関数にはボルツマン選択を用いており温度定数 $T = 0.1$で実験を行う．学習エピソード数は各障害物配置の条件で 4000 エピソード実行する．

〈3･2〉結果・考察

まず Source task の実験結果として，図 2 の各障害物配置における強化学習した移動経路を図 6 に示す．移動経路は，ロボットの 1 行動毎に座標を記録している．各障害物は静止しているため，その障害物や壁に衝突しないでゴールにたどり着いていることが確認できた．次に，知識選択を行いながら移動障害物を回避した経路を図 7 と図 8 に示す．Webots での物理演算シミュレーションはセンサのノイズによる移動誤差が存在するため，複数の経路が Target task にて発現した．図 7 では，移動障害物に対して衝突するような状況になったら知識選択を行い，一度旋回して障害物から遠ざかる方向に移動し，再度ゴール方向へ旋回を繰り返すことで障害物を回避しながらゴールエリアへの到達を行っている．図 8 の結果では，移動障害物に衝突しない範囲で直進し，知識選択をしながら障害物から遠ざかる方向に旋回してゴールエリアへの到達を達成する移動経路となった．

図 7 と図 8 の移動軌跡は共に，前進してから後退する行

動がしばしば見られた．これは Source task における学習が十分でないことも考えられ，とりわけロボットの移動スタート地点ではゴール報酬の行動価値が伝搬しているが行動価値による行動選択において行動が固定化されない探査の余地が残っていることも考えられる．

4. 結言

本論文では，自律移動ロボットの動的障害物回避を目的として，知識選択型転移強化学習である SAP-net のハイパーパラメータを調整することで動的障害物回避が実現可能であることを示唆した．実験では物理演算シミュレータである Webots を採用し，2 輪型移動ロボットが前進や後退，右旋回，左旋回の行動が実行できる設定において，5 種類の障害物配置で強化学習を行い，それらの行動価値関数を選択して，Target task である移動障害物を回避する実験を行った．結果として複数の移動軌跡となる振る舞いが発現したが，シミュレータのセンサノイズや移動誤差などによるものだと考えられる．しかし環境に応じて適応的に移動障害物を回避しながらゴールエリアへの達成を実現した．

今後の課題は次のとおりである．本論文の実験では，少ない知識（行動価値関数）数における選択であったことや動的障害物が単調な移動しか行わない条件でのシミュレーション検証であった．今後はより複雑な条件でのシミュレーションや実際の移動ロボットへの実装も実施する．また，知識選択における選択順序などの妥当性や解析を行うことも今後の課題である．

謝　辞

本研究の一部は JSPS 科研費 JP23K11276 の助成を受けたものである．

文　献

（1） H. Kono, R. Katayama, Y. Takakuwa, W. Wen, and T. Suzuki: "Activation and Spreading Sequence for Spreading Activation Policy Selection Method in Transfer Reinforcement Learning", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, pp. 7-16 (2019)

（2） 河野仁，坂本裕都，温文，藤井浩光，池勇勳，鈴木剛，“知識選択型転移強化学習を用いたシニアカーの自律運転”，2022 年電気学会電子・情報・システム部門大会，pp. 714-718，広島，2022.

（3） R. S. Sutton and A. G. Barto (1998). “Reinforcement learning: An introduction. ” MIT press.

（4） M. E. Taylor: "Transfer in Reinforcement Learning Domains", Springer, Vol. 216 (2009)

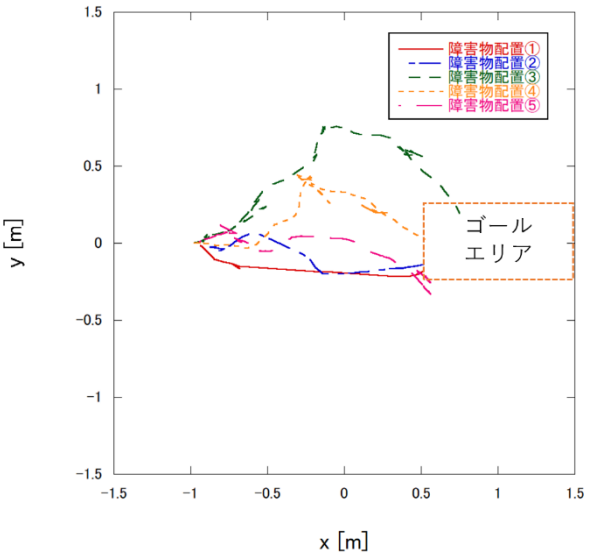

図 6 各障害物配置における学習した移動軌跡

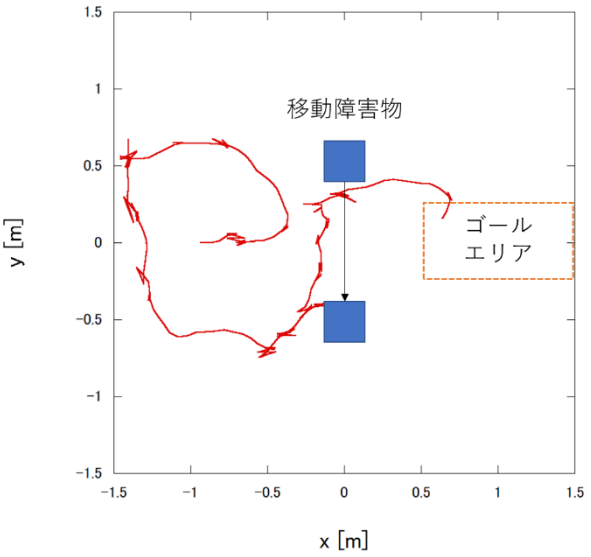

図 7 障害物回避の移動軌跡例

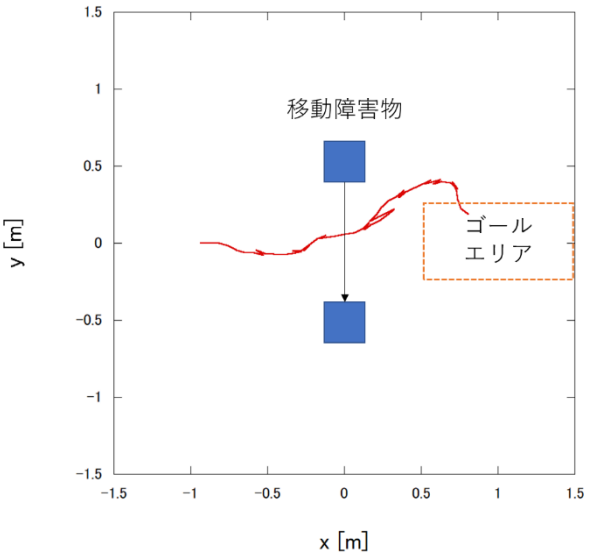

図 8 障害物回避の移動軌跡例