1R-6

二輪倒立ロボットのための強化学習による動作制御と行動選択

藤原 滉司† 平石 広典‡
秋田工業高等専門学校生産システム工学専攻†
秋田工業高等専門学校電気情報工学科‡

1 はじめに

本研究では，二輪倒立ロボットにおけるライントレース制御に強化学習の枠組みを適用する．強化学習とはロボットがある状態においてある動作を選択し，その結果が良ければ報酬を得ることが出来る．そして，最も高い報酬を得るような動作を選択するようになるといった学習方式である．しかし，ある特定の環境への適応は可能であるが，緩いカーブやきついカーブなど，環境の変化への対応が難しいのが現状である．

そこで，本研究では，様々なコースでの学習を行い，学習によって得られた行動を，新たな環境においてロボット自身が選択することのできる学習方式を提案する．

2 強化学習

強化学習（Reinforcement Learning）とは，ある環境内におけるエージェントが，現在の状態を観測し，取るべき行動を決定する問題を扱う機械学習の一種である．エージェントは行動を選択することで環境から報酬を得る．強化学習は一連の行動を通じて報酬が最も多く得られるような方策を学習する．また，未知の学習領域を開拓していく行動と，既知の学習領域を利用していく行動とをバランス良く選択することができるという特徴も持っている．その性質から比較的環境変化のない未知の環境下でのロボットの行動獲得に良く用いられる[1].

図 1 は本研究で採用した学習方式である．これは，ラインに対してロボットが現在位置から線に最も近くなる行動に報酬を与え，線から離れる行動は報酬を減らし，ロボットがラインを突き抜けたり，ラインに近づきはするが，最も近づく方法ではない場合は報酬を与えないという方式である．ロボットは，より多くの報酬をもらうように行動するため，より正確なライントレースの学習をするようになる．

Action control and behavior selection based on reinforcement learning for a two-wheeled robot inverted
† Koji Fujiwara, Akita National College of Technology.
‡ Hironori Hiraishi, Akita National College of Technology.

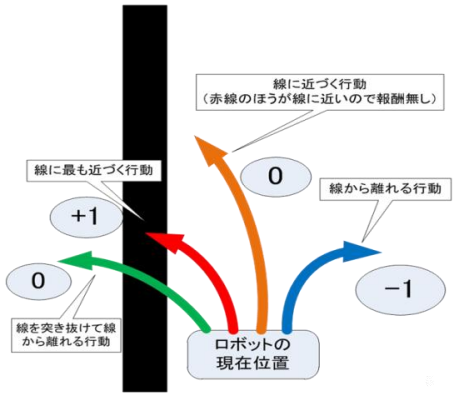

図 1 行動に対する報酬の与え方

3 全体コースと学習コース

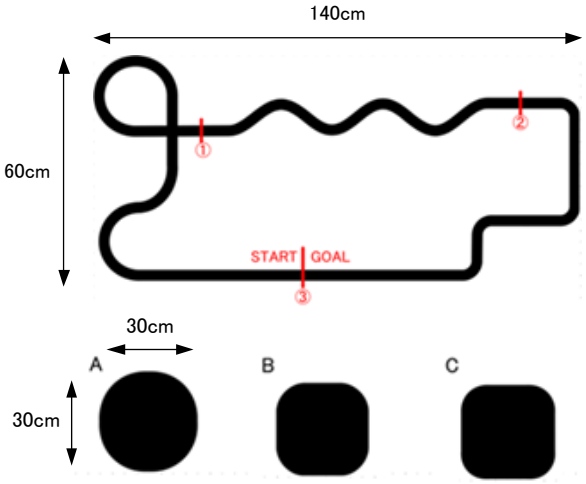

図 2 全体コースと学習コース

本研究では，図 2 に示した全体コースを設定した．また，ロボットが直進速度と回転速度を学習するためのコースとして全体コースのカーブを基に 3 種類の学習コースを作成した．全体コースの途中にチェックポイント（CP）を設置した．A コースは CP③と CP①の間のカーブ，B コースは CP①と CP②の間のカーブ，C コース CP②と CP③の間のカーブに対応している．A コースのカーブが一番緩く，B コース，C コースとなるにつれカーブがきつくなっている．

4　実験結果

表 1　強化学習結果

	テストコース			
	A	B	C	全体
直進速度 , 回転速度 [%]	30 , 30	30 , 45	20 , 60	30 , 60
第1CP通過タイム(成功回数)	21.5秒 (3)	21.9秒 (3)	25.4秒 (3)	21.6秒 (3)
第2CP通過タイム(成功回数)	× (0)	10.8 秒 (3)	9.8秒 (3)	10.0秒 (3)
第3CP通過タイム(成功回数)	× (0)	× (0)	16.4秒 (3)	14.0秒 (2)
合計タイム			51.6秒	45.6秒

表 1 の直進速度と回転速度はロボットの最高速度を 100 としたときの割合である．A，B，C とカーブがきつくコースを学習するにつれ，通過できる CP が増え，通過タイムが遅くなっている．これは，直進速度を落とし，回転速度を上げて走行するためである．表 1 より A コースの学習では B コース，C コースに対応したカーブを曲がりきることができず失敗している．B コースの学習では，A コースと B コースに対応したカーブを曲がることはできるが，C コースに対応したカーブは曲がりきれずに失敗している．そして，C コースの学習では，A コース，B コース，C コースのいずれのカーブにも対応できている．また，表 1 の全体という項目は，初めから全体コースで学習させたときの結果であり，テストコースでの部分的な学習よりも通過タイムも短く，速度も速い．しかし，速度が速いことで第3CP において失敗するケースが見られた．

5　ハイブリット方式

前章より，全体コースの学習データ（全体学習）は，速度は速いがコースアウトすることがあるため確実ではない．そこで，それぞれの環境で得られたデータの特徴を反映させるため，各環境に応じた学習データを使用し，行動選択の強化学習を階層的[2]に行うといった，速さと確実性を両立できるハイブリッド方式を提案する．

図 3 にハイブリッド方式の概要を示す．まず，A 学習から選択し．走行時に得られる報酬の合計は 0 から始める．上位の学習データへ移行するときの値と下位の学習データへ移行するときの値は別の値である．今回の実験では，上位の値を 97，下位の値を-43 とした．この上位の値と下位の値の設定によってデータの切り替わるタイミングが変わるため走行に大きな影響が出てしまう．そのため，最も成功率の高い値を模索した結果このような値となった．

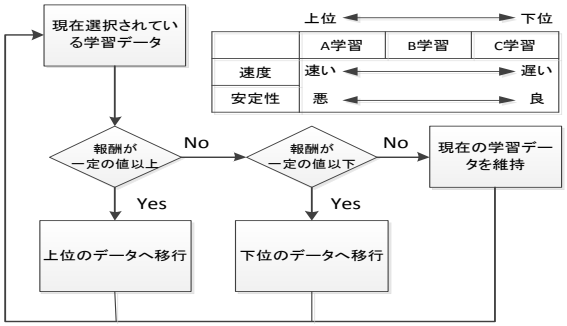

図 3　ハイブリッド方式

表 2　ハイブリッド方式結果

	ハイブリッド方式			
	1回目	2回目	3回目	平均
第1CP通過タイム	21.3秒	21.7秒	21.7秒	21.6秒
第2CP通過タイム	11.4秒	10.8秒	11.4秒	11.2秒
第3CP通過タイム	15.1秒	16.4秒	15.9秒	15.8秒
合計タイム	47.8秒	48.9秒	49.0秒	48.6秒

表 2 はハイブリッド方式の結果を示している．表 1 と表 2 の通過タイムと比べると，表 1 の A，B，C それぞれの CP の通過タイムは表 2 の平均とほぼ一致している．これは，各学習データの特徴がそれぞれの CP に現れており，学習データの切り替えがうまくできていることを示している．

ハイブリッド方式は C 学習のみで走行した時と比べれば速度は速く，全体学習と比べると速度は劣るがミスがなく，速さと安定性を両立していることがわかる．

6　おわりに

本研究では，二輪倒立ロボットにおけるライントレース制御と状況に応じた行動選択に強化学習を用いることで異なる環境において，ロボット自身により適した行動を選択させることを可能にした．

学習コース，全体コースで得られたデータのみではより安全な行動を選択するようになるため，全体的な移動スピードが遅くなる傾向がある．そこで，より正確で速く，様々なコースに対応できるようなハイブリッド方式を実現した．

今後の課題としては，学習したデータをしっかり活用して複雑な環境にも対応できるように各データの精度を上げ，より良いタイミングで切り替えられるようにすることが挙げられる．

7　参考文献

[1] 三上 貞芳，皆川 雅章：“強化学習”森北出版株式会社，第 1 版(2000).

[2] 日野 慎一，Ruck Thawonmas “階層型強化学習の RoboCup エージェントへの適応”高知工科大学 情報システム工学科.