第 30 回日本バーチャルリアリティ学会大会論文集（2025 年 9 月）

# Unity ML-Agents を用いた狭隘丘陵カーブ環境における車両間協調戦略の提案

Proposal of a Cooperative Strategy Between Vehicles on Narrow Hilly Curved Roads Using Unity ML-Agents

増田琉利 [1)]，川西保裕 [2)]，水谷賢史 [3)]
Ryuto MASUDA, Yasuhiro KAWANISHI, and Kenji MIZUTANI

1) 東海大学大学院工学研究科医用生体工学専攻 （〒259-1193 神奈川県伊勢原市下糟屋 143, 4CEYM008@tokai.ac.jp）
2) 東海大学工学部医工学科 （〒259-1193 神奈川県伊勢原市下糟屋 143, 3CEW1157@tokai.ac.jp）
3) 東海大学大学院工学研究科医用生体工学専攻 （〒259-1193 神奈川県伊勢原市下糟屋 143, mk3149@tokai.ac.jp）

**概要**：道路の老朽化や自然災害による通行制限が増える中，安全運転と効率的な協調行動が求められている．本研究では，Unity ML-Agents を用いた強化学習により，坂道・カーブ・狭路を含む複雑な環境における自動運転車の協調行動を実装し，報酬設計とセンサー角度の最適化を行った．成功率や衝突率に基づく定量評価の結果，障害物や対向車の検出精度が向上し，安定した協調行動が促進されることが示された．

**キーワード**：自動運転，強化学習，協調行動

## 1. はじめに

自動運転技術の進展に伴い，強化学習を活用した研究が注目を集めている.Qizwini ら [1] は,航空画像と実際のセンサーデータを組み合わせてシミュレーション環境を構築することで,自動運転車が車線の中央を維持して走行するよう効果的に誘導できることを示した．

近年では，都市部・農村部を問わず,老朽化したインフラや工事による通行止めが増加している．さらに，気象現象により，山間部の道路における土砂崩れなどの自然災害も増えている．このような状況においては,危険な環境下での運転をシミュレートし，自動運転の車両間の協調行動を促すための指針を提供することが重要である．不規則な道路状況にも適応し，安全に走行できる自動運転車の開発が求められている．

これまで，平坦な環境における強化学習の研究 [2] や，丘陵や連続したカーブといった地形特徴を持つ環境における単一エージェント学習の研究 [3] は数多く行われてきたが，丘陵道路における対向車との干渉を考慮した研究は，著者が渉猟した限りでは,狭隘な丘陵道路において対向車との干渉を考慮した研究は依然として少ない．

本研究では,坂道やカーブ，狭路といった特徴を持つ道路環境においてマルチエージェント学習を行い，対向車が存在する状況でのエージェント間の協調行動を評価した．その際，最適なセンサー角度と報酬設定の探索に加え，学習済みモデルを用いた成功率や衝突率に基づく定量評価を行い，これらの要素が学習効率と協調動作に与える影響について検証する．

## 2. 実験原理

### 2.1 強化学習

本研究では，強化学習に基づく手法により，自動運転エージェントが連続した行動を選択しながら安全かつ効率的に走行する戦略を学習する．エージェントの目的は，状態 $s_t$ において行動 $a_t$ を選択し，累積報酬の期待値を最大化することである．この学習過程は，以下の目的関数により定式化される．

$$\pi^* = \arg\max \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t)\right] \quad (1)$$

ここで，$\pi$ は方策,$R(s_t, a_t)$は時刻 t における報酬，$\gamma$ は割引率を示す．

また，各状態と行動の組に対して将来得られる報酬の期待値を表す行動価値関数は以下のように示され，学習においては，エージェントがこの $Q^{\pi}(s, a)$を最大化するように方策 $\pi$ を更新していく．

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t)\right] \quad (2)$$

Abouelazm ら[4] によれば,自動運転における強化学習では，安全性・進捗・快適性・交通規則の遵守といった複数の目的を加味した報酬関数が重要であり，これらを重

み付きの線形和で統合する手法が広く用いられている．本研究においても報酬関数 R(s , a)は，複数の評価指標を線形和として定義する．具体的には,安全走行の度合い $F_{safe}$, ゴール到達状況 $F_{goal}$, および衝突条件 $F_{collision}$ などに重み w1, w2, w3 を付加し，以下のように表される．

$$R(s,a) = w_1 F_{safe} + w_2 F_{goal} + w_3 F_{collision} + \cdots \quad (3)$$

これにより，エージェントは報酬を最大化するように行動を選択し，行動価値 $Q^{\pi}$(s , a)を更新しながら，方策の最適化を図る．

### 2.2 使用アルゴリズム

本研究では，複数の自動運転エージェントが狭隘丘陵カーブ環境において相互に干渉し合う状況を再現するため, Multi-Agent Proximal Policy Optimization（MA-PPO）を適用した. MA-PPO は，各エージェントが個別に観察・報酬を受けながら，共有された物理環境を通じて方策を更新する強化学習手法である [5]. この手法では，環境を通じた間接的な相互作用によって協調行動の学習が可能となる．各エージェントは他車両の存在を把握しながら，報酬最大化を目指す学習を行うため，譲り合いや衝突回避といった分散型の協調行動が報酬値などにより獲得される仕組みとなっている．

### 2.3 エージェントの観察と行動

エージェントの観察について説明する．図 1 に示した Ray Perception sensor を利用した Ray cast Observation という観察方法を使用する．実装によりエージェントは壁，障害物などの要素を報酬値獲得によって学ぶことができる．次に行動について説明する.エージェントは 3D の car モデルを使用し, 1)前進, 2)後退, 3)左旋回, 4)右旋回, 5)ブレーキの 5 種類の行動を実行することができ，現実の車両の物理挙動を模したシミュレーション環境で訓練した．

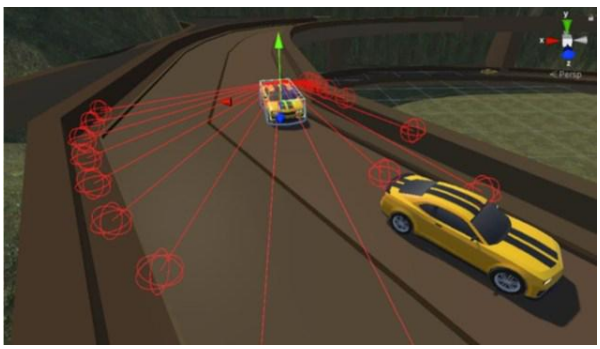

図 1: Ray Perception sensor

## 3. 実験方法

### 3.1 環境構築

作成した環境を図 2 に示す.本環境には，坂道，カーブ，狭路が含まれており，下り車両（①, ②）および上り車両（③, ④）が配置されている．その際，壁や障害物への衝突及びそれぞれの最終地点にゴールを用意し，ゴールに到達するとエピソードが終了し，初期位置に戻る仕様とした．表 1 にセンサー設定および報酬設定の概要を記載した．


図 2: 実験環境

表 1: センサーと報酬設定

| Item | | | Reward Value |
|---|---|---|---|
| Reward | Default | over the legal speed limit | -0.05 |
| | | Forward at appropriate speed | 0.015 |
| | | Forward at low speed | 0.002 |
| | | Forward at very low speed | -0.01 |
| | | Crashing into a wall or obstacle | -4.0 |
| | | Crashing into a car | -5.0 |
| | | Crashing into goal | 9.0 |
| | Added | Retreat | -0.05 |
| | | Prolonged stop | -0.1 |
| Sensor | Angle | 0° | |
| | | 45° | |
| | | 90° | |
| | | 135° | |
| | | 180° | |
| | Vertical offset | Start: 0<br>End: 0 | |
| | | Start: 1.8<br>End: -5.5 | |

### 3.2 実験内容

#### 3.2.1 予備検証

表 1 に示す報酬設定の下で，センサー角度の違いがエージェント行動に与える影響を確認した. 90°では前進・回避といった行動がある程度可能であったが, 180°では壁への衝突が多発し，安定した学習が困難であった．

#### 3.2.2 報酬設計の工夫

予備検証の結果より，センサー角度を 90°に固定した上で，停止や後退を抑制する報酬設定を追加し，学習効果を

比較した. その結果, 衝突回数と停止頻度が減少し, より安定した行動方策が獲得された.

#### 3.2.3 最適センサー角度の探索

改良後の報酬設定を用い, エージェントのセンサー角度が学習結果に与える影響を評価するため, センサー角度を 0°から 180°まで 45°刻み（0°, 45°, 90°, 135°, 180°）で変更し学習を行った. 各モデルは同一の環境条件および報酬設定のもとで訓練され, 角度ごとの学習傾向や行動の違いを比較可能とした. またステップごとの平均累積エピソード報酬の比較を TensorBoard によって評価した.

#### 3.2.4 学習モデルの評価

学習モデルの性能は, 統一された評価環境下で 100 エピソードを実行することで評価した. 評価環境は学習時とは異なる構成で設計されており, 各車両がゴール地点に向けて走行する. エピソードは, ゴール到達・壁衝突・車両衝突のいずれかが発生した時点で終了し, 環境が即時にリセットされる設定とした. 評価指標としては, 以下の式に基づいて成功率と衝突率を算出した.

$$\text{成功率} = \frac{Nsuccess}{Ntotal} \quad \text{衝突率} = \frac{Ncollision}{Ntotal} \tag{4}$$

ここで, $N_{success}$ はゴール到達回数, $N_{collision}$ は壁または他車両との衝突回数, $N_{total}$ は総試行回数（100 回）を表す. 衝突については壁衝突および車両衝突に分類し, それぞれの割合を積み上げ棒グラフにより視覚化した.

## 4. 結果

### 4.1 センサー角度による学習傾向の違い

センサーの観測範囲の違いに基づく, 累積エピソード報酬の移動平均の推移の事例を図 3 に示す. 本結果から, センサー角度によって学習の成否および収束の速度に明確な差が見られた.

成功例としては, センサー角度が 45°, 90°, 135°の条件下において累積報酬が正の値で安定し, 学習が有効に進んだことが確認された. 特に 45°の条件では, 約 300k ステップ時点で報酬が正に転じ, その後は一貫して増加し, 500k ステップ以降は+30 付近でほとんど変動せずに安定した. これは, 極めて高い安定性と学習効率を示す結果である. 90°の条件では, 700k ステップで報酬が正に転じ, 800k ステップ以降に+30 付近で収束した 135°の条件でも 700k ステップで報酬が正に転じ, 800k ステップ以降は+20 付近で安定した. 上昇傾向は緩やかだったが, 安定的な学習が確認された.

一方, 失敗例に該当する 0°および 180°の条件では, 累積報酬は終始負の値にとどまり, 学習の改善は見られなかった. 0°では 200k ステップ以降も-20〜-10 の範囲で停滞し, 180°では 500k ステップ以降でも-40〜-10 の範囲で停滞した.

### 4.2 学習モデルの評価

学習モデルの性能は, 学習時とは異なる評価環境において, 各条件で 100 エピソードを実行し評価した. 図 4 に, センサー角度および報酬設計の違いによる成功率および衝突率を示す. その結果, センサー角度 45°かつ報酬設計を改良したモデルが最も高い性能を示し, 成功率 100%, 衝突率 0%を達成した. 90°条件においても比較的良好な結果が得られたが, 報酬設計の有無によって成績にばらつきが見られ, 環境変化に対する適応力は 45°条件より劣る傾向があった. また 135°では, 図 3 の事例で報酬が正に転じていたが, 学習モデルで評価した際は, 衝突率が 84%であった.

一方で, 0°および 180°では成功率が 0%にとどまり, 衝突率は 100%となった. 特に 180°では, 前方の対向車両を認識できず, 車両衝突が集中して発生した. また, 報酬を改良していない条件では, 適切なセンサー角度を設定していても, 壁への衝突や停滞行動が多発し, 十分な走行行動を学習できていないことが確認された.

## 5. 考察

本研究では, 強化学習による自動運転エージェントの協調行動獲得において, センサー設定および報酬関数の設計が学習成果に及ぼす影響を評価した.

まずセンサー角度の変更に関する実験結果から, 45°の角度設定が最も安定した学習成果を示した. この設定では, 車両前方に加え左右方向への適度な視野を確保でき, 対向車や障害物の早期検知につながったと考えられる. 対照的に, 0°や 180°では視野不足または過剰による観測ノイズが増大し, 行動選択の不安定化が見られた. また 135°条件においては,TensorBoard 上では累積報酬が正に転じていたものの, 学習モデルを用いた評価では衝突率が 84%と高く, 実運用上の安定性が欠如していた. この結果は, 「累積報酬が高い＝必ずしも安定した運転行動を意味しない」ことを示唆しており, 今後の評価指標設計において注意を要する必要があることが示唆される.

次に報酬設計の工夫により, エージェントの行動探索が大きく変化した. 表 1 に示したデフォルト報酬のみでは, 前進・停止・後退といった基本動作に偏りが生じ, 衝突や無駄な動作が多発した. 一方で,「動き続けること」や「後退・停止への罰」を設けたことで, 学習の停滞が解消され, より積極的な回避行動が誘導された. これは,Berta ら [6] の報告と一致しており, エージェントが自由に探索しつつ報酬を受け取れる設計が, 強化学習における収束の鍵となることが示された.

また学習された回避行動には, 内側・外側双方への回避が確認されたが, これは報酬関数が「車線維持」よりも「衝突回避」「ゴール到達」に重点を置いていたためである. この点は, 今後の報酬バランス調整において現実的な交通ルールや通行優先権を反映させる必要性を示唆している.

## 6. むすび

本研究では, Unity ML-Agents を用いて, 狭隘丘陵カーブ

環境における複数車両の協調行動を強化学習により獲得することを試みた．その結果，適切なセンサー設定と報酬関数の設計が，安定的かつ現実的なエージェント行動の学習に不可欠であることが明らかとなった．

今後は，車線維持や通行優先といった交通規範に準じた報酬項目の導入，および動的交通環境下での適応行動の獲得を目指す．これにより汎用性の高い協調運転アルゴリズムの確立を進めていく予定である．

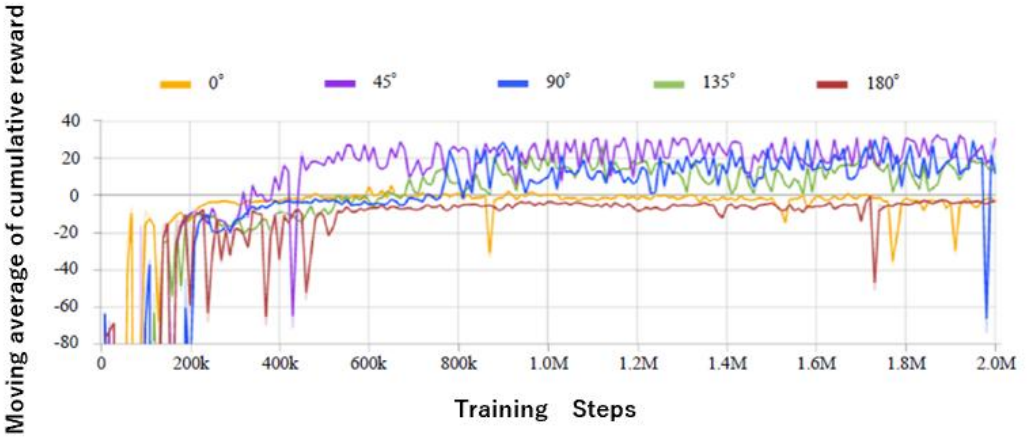


図 3：センサー角度の違いによる累積エピソード報酬の移動平均のトレーニングによる推移の事例

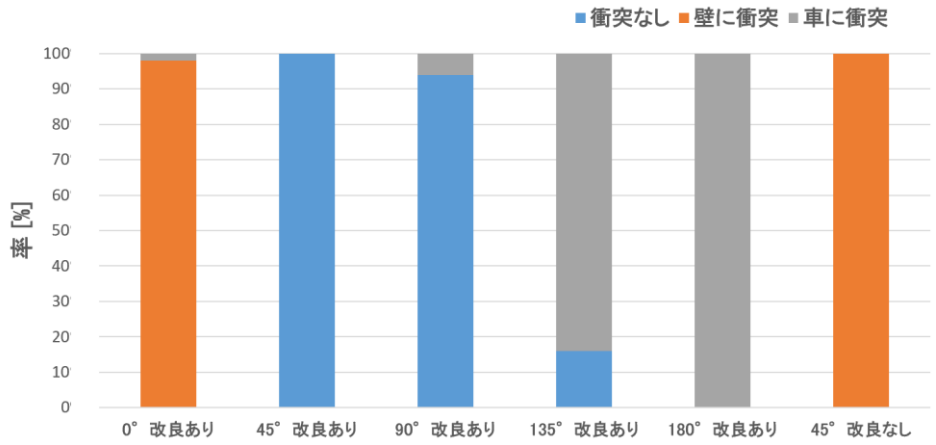


図 4：学習モデル評価

## 参考文献

[1] M. Al-Qizwini, O. Bulan, X. Qi, Y. Mengistu, S.Mahesh, J.Hwang, and D.Clifford, "A Lightweight Simulation Framework for Learning Control Policies for Autonomous Vehicles in Real-World Traffic Condition," IEEE Sensors Journal, vol. 21, no. 14, pp. 15762-15774, 2021.

[2] F. Shafique, T. Naeem, A. Hussain, “Path Tracking and Obstacle Avoidance Control Using Deep Reinforcement Learning for Autonomous Vehicles,” IEEE proceedings, 2023.

[3] X. Li, Z. Cao, Q. Bai, "A Novel Mountain Driving Unity Simulated Environment for Autonomous Vehicles," Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), pp. 16075-16077, 2021.

[4] A. Abouelazm, J. Michel, and J. M. Zöllner, "A Review of Reward Functions for Reinforcement Learning in the Context of Autonomous Driving," IEEE Intelligent Vehicles Symposium (IV) proceedings, pp. 156–162, 2024.

[5] M. Q. Phan, V. H. La, and H. H. Nguyen, "Simulated Autonomous Driving Using Reinforcement Learning," *Information*, vol. 14, article 290, pp. 1-22, 2023.

[6] R. Berta, L. Lazzaroni, A. Capello, M. Cossu, L. Forneris, A. Pighetti, F. Bellotti, "Development of Deep-Learning-Based Autonomous Agents for Low-Speed Maneuvering in Unity," Journal of Intelligent and Connected Vehicles, vol. 7, no. 3, pp. 229-244, 2024.