強化学習による歩行者シミュレーションにおける行動規則の自動生成

○小林姫華 [1] 森山甲一 [1] 松井藤五郎 [2] 武藤敦子 [1] 犬塚信博 [1]
([1] 名古屋工業大学 [2] 中部大学)

Automatic Generation of Behavioral Rules for Pedestrian Simulations Using Reinforcement Learning

∗H. Kobayashi [1] , K. Moriyama [1] ,T. Matsui [2] , A. Mutoh [1] and N. Inuzuka [1]
([1] Nagoya Institute of Technology , [2] Chubu University)

Abstract– In designing pedestrian simulations, it is common for model designers to prepare in advance behavioral rules of the pedestrians. To reduce the burden on the designers, we aim to automatically generate the rules using reinforcement learning. In this work, we consider counter-flow simulations where agents move toward each other. They decide their actions based on visual information. This work designs their states to avoid collisions in the simulations. From the simulation experiments, we found that the proposed agents learn appropriate behavioral rules to arrive at the destination while avoiding collisions.

Key Words: Reinforcement learning, Pedestrian simulation, Behavioral rules

1 はじめに

現在、施設を設計する際や、イベント等の混雑を予測する際には、そこで歩行者がどのように動くかを検討する必要がある。しかし、当然のことながら、設計時や予測時に実際の歩行者の動きを調査することはできない。そこで、歩行者の動きをシミュレーションにより再現して、対象とする環境で歩行者がどのように振舞うかを観察し、生じる問題点を検討することが広く行われている。

一般的に用いられている歩行者シミュレーションでは、歩行者がどの状態でどの行動をとるのかという行動規則を、モデルの設計者が予め決めておくことが多い。例えば、自身を基準としたグリッド空間を作り、そのマスの中に存在する障害物の有無に応じてあらかじめ作成した行動規則を適用して衝突回避する方法 [1)] がある。しかし、この方法は人数や状態数が増えるほど設計が困難で非現実的になる。他にも障害物との距離から引力・斥力を求めて衝突回避を行う方法 [2)] などもあるが、現実社会の人間のように行動選択を柔軟に行えないという欠点が挙げられる。

本研究では、歩行者シミュレーションに必要な行動規則を自動生成することを目的とする。シミュレーション時に想定する人数や状態数がどれだけ増えても設計者に負担をかけず対応できることに加え、経験した状態の違いから人間のような個性も習得できると考えられるため、機械学習手法の一種である強化学習を使用する。本研究では、シミュレーションの中でも基本的な動作である衝突の回避が自動で学習できるかどうかを、平均衝突回数と平均獲得報酬により検証する。

また、本研究では、対向流と呼ばれる二つの歩行者の群れが互いに向き合って移動する歩行者流におけるエージェントを考える。対向流は施設内の移動等で広く生じることから実用性が高く、歩行者の流れが層状になることが報告されるなど、先行研究による知見が多い。このような知見をもとに、獲得された行動規則をエージェントに適用することで、現実に発生する層状の流れが確認できるかも調査する。

2 準備

2.1 強化学習

強化学習とは、機械学習の分野の一つである。強化学習エージェントは、自身が遭遇したそれぞれの状態でどのような行動を選択すると報酬がどれだけ得られるのかを繰り返し経験する。そして最終的に、一連の行動選択の後に得られる報酬の総和の期待値が最大となる行動選択をするように学習する。強化学習は、教師データの代わりに、結果の望ましさを示す報酬を与えることでエージェントが試行錯誤することによって学習を行う。また、ある状態で、ある行動を選択したときにどれだけ報酬を得られそうかを表す値を行動価値と呼ぶ。エージェントは繰り返し様々な状態と行動選択、報酬を経験する中でこの行動価値を更新して学習し、真の価値に近い値を導き出し、行動選択の際にこの行動価値が高い行動を選択することによって、より多くの報酬が得られると考える。

強化学習には、以下の式を用いて行動価値関数を更新して学習を行う Q 学習と呼ばれる学習法がある [3)]。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \tag{1}$$

式中の s_t、a_t、r_t は時刻 t における状態、行動、報酬を、α は学習率、γ は割引率を表している。エージェントは行動価値関数の値である Q 値を学習し、それぞれの状態で Q 値に基づいて、より報酬が得られると期待できる行動を選択する。しかし、単純に Q 値が最大の行動のみを選択すると、他の全く更新されていないような状態・行動ペアや、更新の際にたまたま運悪く Q 値が低くなってしまったが、本来はもっと価値の大きな行動が選択されなくなってしまう。

そこで、自分の今持っている知識を利用して多くの報酬が得られそうな行動を選択しつつ、まだ十分に探索し切れていないものを経験するために、ϵ-greedy 法

と呼ばれる方法がある。この方法は、低い確率 ϵ で行動をランダムに選択してまだ十分に更新されていない状態や行動を探索し、残りの確率 $(1-\epsilon)$ で行動価値が高い物を選ぶことによって探索と利用のジレンマに対応する。この ϵ の値は学習を開始したばかりの時には大きく設定して様々な状態と行動とそれに対する報酬を経験させ、学習が進むにつれて徐々に小さくすることで効率的に多くの状態を探索することができる。本研究では Q 学習と ϵ-greedy 法を組みあわせて用いる。

2.2 歩行者流

歩行者流とは人々が移動する際に向かう方向別に群衆が生じて水や粒子のように流れる状態を指し [4]、一方向流、対向流、交差流の 3 つの種類が存在する [5]。

一方向流とは、歩行者がすべて同じ方向を向いて移動するものである。一方向流では、可能な限り迂回はせず、一人一人が固有の速度でかつ急いでいない平常時は障害物とは一定の間隔を保ちながら流れることがわかっている [6]。ただし、急いでいる場合や人口密度が高い場合はこの間隔が狭まる。

対向流は、歩行者の群れが互いに向かい合って流れる歩行者流を指す。対向流では、主に以下の 6 つの相が確認できる [1]。

静的層流
: 人々の群れが進行方向別に交互に 2 列以上できる現象を指す層流が形成され、向かい合う歩行者同士の回避動作がない

安定層流
: 層流が形成されるが、向かい合う歩行者同士の回避行動が見られる

層乱混合流
: 向かい合う歩行者同士の回避動作により層流が形成と崩壊を繰り返していて一定の層は見られない

乱流
: 層流も閉塞も形成されない

不完全閉塞
: 閉塞が部分的に見られ、発生と崩壊を繰り返す

閉塞
: 閉塞が発生し、歩行者は移動できなくなる

交差流は、渋谷のスクランブル交差点のように 2 つの歩行者の群れが斜めに移動し，経路の中心付近で互いに交わるような流れを指す。よって、中心付近では多くの歩行者が衝突する可能性がある。交差流では、人数が少ない状態では回避行動が容易だが、人数が増え出すと回避が複雑かつ困難になり、レーン形成と呼ばれる前の人に続いて歩行する状態が見られる [4]。

2.3 強化学習による行動規則の獲得

木村ら [7] は交差流を想定した歩行者シミュレーションで、強化学習を用いて行動規則を自動生成する手法を提案した。ここで提案されたエージェントは、Fig. 1 のように正面を中心として左右 45 度ずつ視野が広がっており、その中に 7 本の視線を等間隔で設けている。行動はいずれかの視線の方向への移動または停止で、視線に当たった他者、壁、ゴールの有無を状態とした。また、他者や壁との衝突で負、ゴール到達で正の報酬を得る。エージェントは各状態、行動、報酬をもとに Q 学習を行い、学習された Q 値から最適な行動を選択する。実験では、エピソードが進むにつれて衝突回数が減少し、平均獲得報酬は増加していることが確認され、さらに学習された Q 値からエージェントは衝突回避を行う適切な行動規則を獲得したことが示された。しかし、この研究では交差流を想定した場合の行動規則が獲得されたのであって、他の歩行者流を想定した際の行動規則の獲得については言及していない。

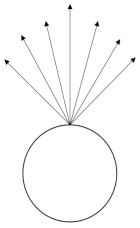


Fig. 1: Lines of sight of Kimura et al.'s method.

3 提案手法

木村らの手法は交差流を想定したものであった。本研究では対向流でのエージェントを想定するため、対向流の特徴を踏まえて回避動作に必要な視野を検討する。

1 つめの特徴は自分と同じ方向へ進む者が急に進行方向を大幅に変更し、衝突してしまうような事象は少ないことが挙げられる。つまり、対向流の行動規則を生成するためには、視野内の他者の向きを取得することが重要である。そこで、木村らの手法では Fig. 2 のように他者の向きは利用していなかったが、本研究では、提案手法 1 として Fig. 3 のように進行方向に合わせてエージェントのタグをわけて状態を区別する。

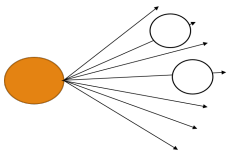


Fig. 2: Recognition of others by Kimura et al.'s method. The large circle indicates the agent while the small circles are others.

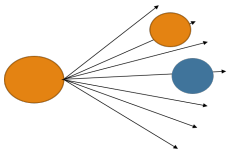


Fig. 3: Proposal 1: Recognizing others' moving directions. The large circle indicates the agent while the small circles are others. The colors (orange and blue) indicate the moving directions.

2 つめの特徴として、互いに向かい合って進むことが挙げられる。よって、歩行速度の 2 倍以上の距離がある場合は回避が可能であるため、視線の長さは最高速度の 2 倍以上あれば問題ない。木村らの手法では他者との距離を判断せず衝突回避を行っていたが、本研究では提案手法 2 として、視線の距離を Fig. 4 のように設定する。

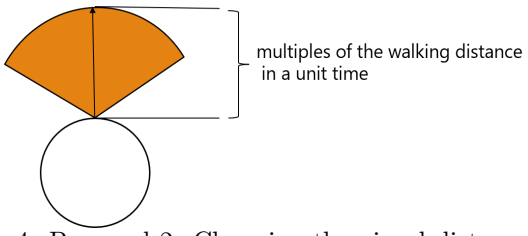


Fig. 4: Proposal 2: Changing the visual distance following the walking speed.

3 つめの特徴として、互いに向かい合った状態ですれ違うため衝突しやすい他者は正面付近に固まることが挙げられる。よって、木村らの手法では視野内の視線は等間隔に広がっていたが、本研究では提案手法 3 として、Fig. 5 のように視線は正面を 0 度として左右 10 度、20 度、45 度に視線を伸ばすように設定した。

人間の視野には中心視野と周辺視野という 2 種類の視野が存在している [8]。中心視野に入った物は正確に認識できるが、周辺視野に入った物は大まかな情報しか認知できない。周辺視野は端に行くにつれておおまかな形や色程度しか把握できない範囲から、物があったとしても見えず明るさのみしか認識できない範囲へと移ろう。つまり、正面方向の視線の密度を高くして視界の端に向かうほど疎になるように変更することは人間の視野の特徴を踏まえたうえでも矛盾しないことがわかる。

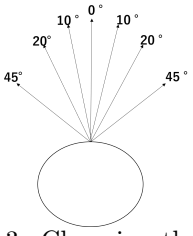


Fig. 5: Proposal 3: Changing the density of lines of sight.

次に、エージェントの学習方法について説明する。各エージェントは初期位置から状態認識、行動選択、報酬獲得をエピソードが終了するまで繰り返す。エピソード終了後に、エピソード開始から終了までに経験した状態・行動・報酬・次状態を順番に用いて、Q 学習で行動価値を更新する．以上を指定されたエピソード数だけ繰り返す。

4 実験

提案手法を用いたシミュレーション実験を統合開発環境の Unity[9] 内で行った。環境には Fig. 6 のように、長辺が 20 で短辺が 1.5 の範囲を作る 4 つの壁と長辺の両端に設置された一辺が 1 の赤い立方体であるゴールが存在し、各ゴール前にはエージェントを 6 体ずつ配置する。エージェントは配置された場所とは逆の端に設置されたゴールを目指す。ここで、本研究の目的は行動規則の獲得であって、ゴール位置の探索ではないため、目指すゴールの座標は予めエージェントに与えられている。エージェントの直径は 0.2、歩行速度は 0.3 で、視線の距離は 1.0 である。ただし、提案手法 2 では視線の距離を歩行速度の 1〜3 倍にできるよう設定した。

各ステップでは、12 体のエージェントが状態を認識し、それぞれの方策に基づいて行動を選択する。全員の行動後に報酬を各エージェントに与え、次のステップへ遷移する。行動選択の際には、各エージェントは向かうべきゴールを視野の中心にとらえて、視線のうちから一方向を選択して移動するか停止する。報酬は他者との衝突で −100、壁との衝突で −5、ゴール到達で (500 − 経過ステップ数)、1 ステップ毎に −0.1 である。また、3000 エピソードかけて学習した後、100 エピソードで学習結果を観測する。エピソードは 500 ステップが経過するか全員がゴールに到着すると終了する。学習率 α は 0.1、割引率 γ は 0.99、ϵ はシミュレーション開始時には大きくし、エピソード数が増えるにつれて徐々に小さくするため、以下の式で設定した。式中の n はエピソード数である。

$$\epsilon = 0.5 \times 0.99^n \tag{2}$$


Fig. 6: Experimental environment.

また、木村らの手法では全てのエージェントが決められた順番に状態を把握し、その後再び順番に行動選択を行う更新方法を用いていた。しかし、現実社会では全ての人間が決められた順番で動くわけではない点、また全員が状態を把握した直後に移動することで相手の行動が予測できず衝突してしまう点を踏まえて、エージェントの行動選択の順序について以下の 3 種類の実験を行った。

実験 1 全員が状態を確認した後再び順番に行動選択

実験 2 全員が状態を確認した後ランダム順で行動選択

実験 3 ランダムな順番で状態を確認して行動選択

4.1 実験結果

Fig. 7 から Fig. 12 にそれぞれの実験での学習中の 1 エピソード毎の 12 体のエージェントの平均衝突回数と平均獲得報酬の推移を、Table 1 から Table 3 に各実験での 12 体のエージェントの学習後 100 エピソードの計 1200 サンプルにおける平均衝突回数と平均獲得報酬を示す。ただし、見やすさの関係で、学習中のグラフは木村らの手法と提案手法を全て使用した場合の結果のみを示す。また、提案手法を全て使用した場合及び実験 2 と実験 3 での提案手法 2 の視線の距離は速度の 2 倍を用いた。

Fig. 7 から Fig. 12 より、学習中はどの手法もエピソード数が増加すると平均衝突回数は徐々に減少し、平均獲得報酬は増加していることが確認できる。また、Table 1 から Table 3 より、提案手法を用いた場合の学習後の平均衝突回数は木村らの手法よりも減少したことが確認できたが、0 に収束することはなかった。実験 1 では、提案手法 1 と提案手法 3 と提案手法全ての場合に Fig. 13 のような層流を確認でき、提案手法 1 と

提案手法全ての場合には同じ方向へ進む者の後に続くエージェントが見られた。実験 2 は実験 1 と比べて見られる結果と層流に差はあまりみられなかった。実験 3 では実験 1，実験 2 と比べて大幅に良い結果が確認できたが、どの場合も層流は確認できなかった。

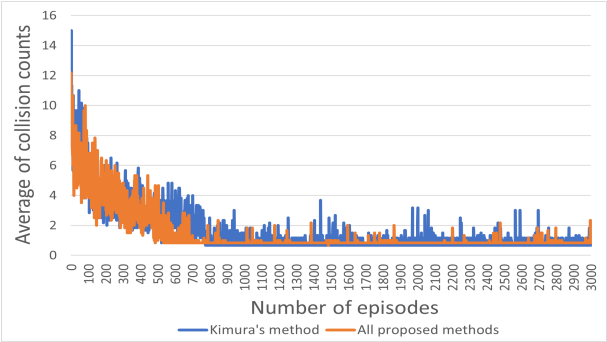


Fig. 7: Average of collision counts in Experiment 1.

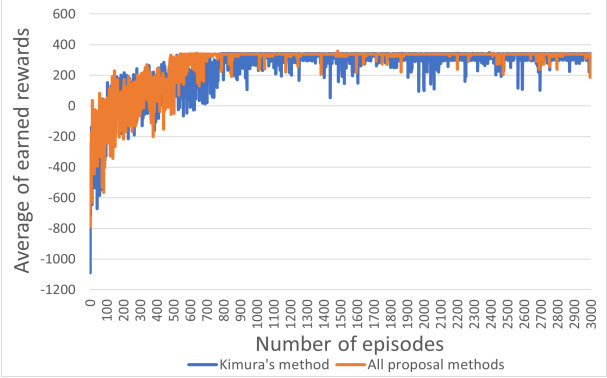


Fig. 8: Average of earned rewards in Experiment 1.

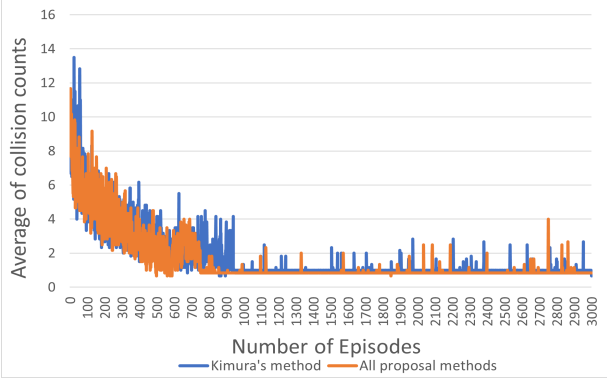


Fig. 9: Average of collision counts in Experiment 2.

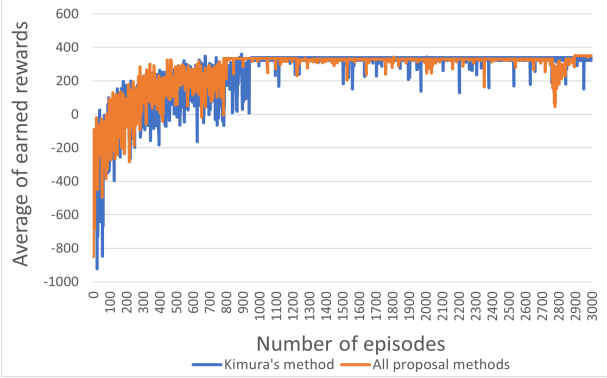


Fig. 10: Average of earned rewards in Experiment 2.

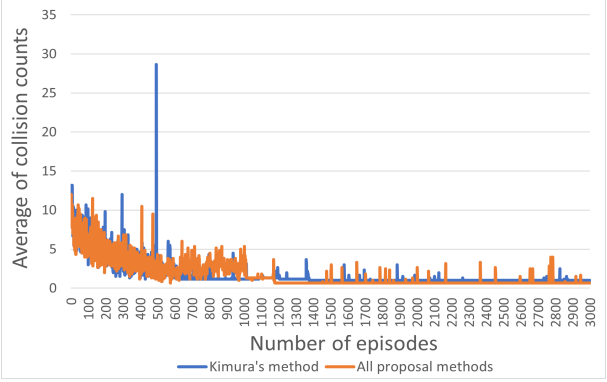


Fig. 11: Average of collision counts in Experiment 3.

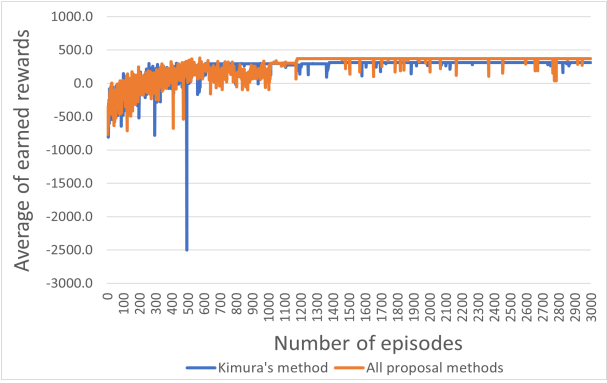


Fig. 12: Average of earned rewards in Experiment 3.

Table 1: Result of Experiment 1 (rounding off to one decimal place)

Method	Average of collision counts	Average of earned rewards
Kimura et al.	0.9 ± 0.7	320.1 ± 102.2
Proposal 1	0.8 ± 0.5	327.9 ± 80.4
Proposal 2 (1x distance)	1.8 ± 0.7	246.3 ± 61.2
Proposal 2 (2x distance)	0.8 ± 0.5	354.2 ± 67.9
Proposal 2 (3x distance)	1.2 ± 0.8	294.8 ± 95.0
Proposal 3	0.8 ± 0.6	328.6 ± 79.2
All proposals	0.8 ± 0.6	338.1 ± 68.3

Table 2: Result of Experiment 2 (rounding off to one decimal place)

Method	Average of collision counts	Average of earned rewards
Kimura et al.	0.9 ± 0.5	332.5 ± 70.1
Proposal 1	0.8 ± 0.5	337.8 ± 84.3
Proposal 2 (2x distance)	1.0 ± 0.4	322.0 ± 59.0
Proposal 3	1.2 ± 0.8	317.4 ± 88.8
All proposals	0.8 ± 0.6	349.6 ± 48.2

Table 3: Result of Experiment 3 (rounding off to one decimal place)

Method	Average of collision counts	Average of earned rewards
Kimura et al.	1.0 ± 0.9	311.0 ± 120.5
Proposal 1	0.5 ± 0.8	406.1 ± 75.5
Proposal 2 (2x distance)	0.5 ± 0.5	393.0 ± 50.1
Proposal 3	0.5 ± 0.5	397.1 ± 46.8
All proposals	0.6 ± 0.7	369.7 ± 65.1

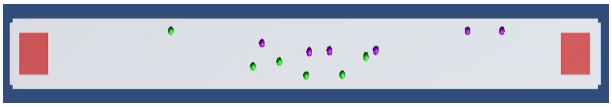
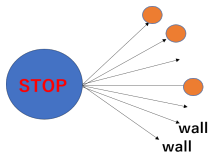
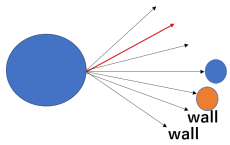


Fig. 13: Example of a confirmed laminar flow. Green agents are moving to the right, while red agents are moving to the left.

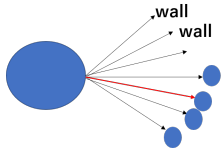
Fig. 14 に学習された Q 値の例を示す。これらが示す通り、学習した Q 値からは障害物が多くある場合には停止する、障害物を避ける、同じ方向の他者の後を進む等の行動選択を確認できた。また、積極的に衝突回避を行うものや逆に道を譲ってもらうよう仕向けるもの、わざわざ大回りをして衝突を回避するものや最低限の動きで回避するもの、同じ向きに進む二者が前後に並んだ場合、前者が停止すると後者も同じように停止するものやわざわざ追い越して先を急ぐものなど、同じ状態でもエージェントが異なると衝突を回避する方法に個性が見られた。



(a) The agent chose to stop to avoid collisions.



(b) The agent chose the red direction to avoid the other agents and walls.



(c) The agent chose the red direction to follow the other agent.

Fig. 14: Examples of learned Q-values. The large circle indicates the agent while the small circles are others. The colors (orange and blue) indicate the moving directions.

5 考察

実験 1 と 2 の提案手法 1 と全てを用いた場合で層流が確認できたことから、他者の方向の判断は層流の形成に重要であることがわかる。

実験 1 の提案手法 2 で視線の長さを 1、2、3 倍に変更した場合、2 倍の時は平均衝突回数が最も小さいと同時に標準偏差が最も小さく良い結果が出ている。速度の 1 倍では視野内に他者が入ってきたと判断するタイミングが遅く、回避行動が間に合わないので、特にエージェントが集中する中央付近ではどうしても衝突が回避できない場合が増えて結果が悪くなっていると考えられる。また、速度の 3 倍では 2 倍の時よりも衝突が多く累計報酬が少ないことがわかる。例えば、あるエージェントが視野内に他者を確認し、視線の距離が長いためにそのまま直進してもぶつからなかった状態を学習したとする。しかし他のステップで同じような状態に遭い、そのまま進んだところ学習した時とは相手の位置が異なり、かなり近くにいたせいでぶつかってしまった場合などが考えられる。つまり、視線の距離が長いと今自分がすぐに衝突回避をしなければならない場合なのか、それともそのまま直進しても構わない場合なのかがうまく判断できず、学習が上手くいっていないと考えられる。

実験 1、2 のどの提案手法でも衝突回数を 0 にできなかった原因としては、全員が状態を確認した後で一人一人行動を選択しているため、相手の向かう先は状態を確認している時には判断できないことが考えられる。例えば、自分が移動する前に視野内の他者が移動を完了していたために衝突しなかった場合を学習したとする。しかし、同じ状態に遭遇したとしても、この他者が実験 1 では自分より後に動く者だった場合や、実験 2 ではランダム順で行動するため、相手よりも先に自分が動いてしまった場合に、衝突が起こってしまうと考えられる。

また、実験 3 の結果が最も良い理由は、エージェントが状態を確認してそのまま行動を選択できることによる。状態確認の際に、ある方向に他エージェントが存在しなかったならば、その方向に移動して衝突することはあり得ない。一方で実験 1、2 では、状態確認と行動選択に時間的なずれが存在するため、複数のエージェントが同じ場所に移動して衝突することがあり得る。これが、実験 3 において層流が発生しない原因になっているとも考えられる。行動選択の際に衝突のリスクが存在しないため、比較的安全である同じ方向へ進む者の後を追う必要がなくなるのである。つまり、対向流では、人々がリスクを予測して回避するために、同方向の人の後に続いて列をなすことで層流が生じていると考えられる。

それでも実験 3 の衝突回数が 0 回にならなかった原因は、他者との衝突が起こった際の報酬の大きさによると考えられる。本研究では、木村らの手法と比べてエージェントが歩く距離が長くなり、エピソード開始からゴールに到達するまでのステップ数が増加したため、ステップ数を木村らの手法の 100 から 500 に増やしている。木村らの手法では、他者との衝突が一回でも生じると -100 の報酬が生じ、1 エピソードの報酬の総和は必ず負になる。一方、本研究では他者との衝突

による報酬は木村らの手法と同じ値を用いたので、ステップ数が増えていることを踏まえると、1 回程度なら衝突しても報酬の総和は負にはならない。そこで、他者が多い中央付近で身動きがとれなくなった場合、衝突してでもできるだけ早くゴールに到達する方策を学習したと考えられる。実際、視野内のほとんど全てに他者がいる状態の Q 値からは、停止せずに、他者がいたとしてもゴールに最も近づく正面方向への行動を選択していることが確認された。

6 まとめと今後の課題

本研究では、対向流の特徴を踏まえて衝突回避用の視野を提案し、対向流を想定した歩行者シミュレーションでの行動規則を強化学習を用いて自動生成した。エージェントの更新順を 3 つに分けて実験を行い、それぞれで既存手法と比べると提案手法を用いた場合には衝突回数が減少していること、Q 値から周囲の他者や障害物を適切に避けていること、層流の形成がリスク回避の結果かもしれないことを確認することができた。

本提案手法では他者の向きを目指すゴール別に 2 つに分けて大まかに判断できるようにしている。しかし、現実社会の人々は視野内の他者の顔や足の向きから進行方向を知り、自分と衝突する可能性があるか判断している。また、本提案手法では視界に入る障害物との距離は、あらかじめ速度を元に決めて衝突回避に用いている。しかし、人々は視野内の他者との大体の距離をその場その場で測って衝突回避する際に役立てている。さらに、人々の歩行速度はそれぞれ異なる。以上より今後の課題は、正確な他者の向きの把握、他者との距離の計測、歩行速度を変化させる方法を提案・実装して、より人間らしく、より正確に衝突を回避する行動規則を学習させることである。

さらに、本研究ではエージェントの初期位置を固定としたため、前列にいるエージェントと後列にいるエージェントでは学習に差が生じてしまう。例えば、最初のステップで後列のいずれかのエージェントが前にいるエージェントより先に動くと、前列のエージェントが動かずまだ存在しているために衝突回避を行おうと停止する行動を学習してしまう。すると、エピソード途中で前列にいたエージェントが後列のエージェントの前に出てくると、後者は直進しても構わないのに停止する適切ではない行動を選択してしまうことがある。したがって、エージェントの初期位置をエピソードごとに変化させることが必要である。

謝辞

本研究は、JSPS 科研費 JP18H03825 および JP19K12118 の助成を受けたものである。

参考文献

1) 正光将大, 兼田敏之: 行動ルールを用いた歩行者エージェントモデルによる対向流の相転移の分析, 日本建築学会技術報告集, **23**-54, 721/724 (2017)

2) 上水流友望, 富井規雄: マルチエージェントモデルによる繁忙期における新幹線駅ホーム上の旅客流動シミュレーション, 電気学会論文誌 D, **134**-8, 750/759 (2014)

3) Christopher J.C.H. Watkins, Peter Dayan: Technical Note: Q-learning, Machine Learning, **8**, 279/292 (1992)

4) 石橋竹志, 鈴木章彦, 渋谷秀雄: マルチエージェントシミュレーションを用いた歩行者流の解析, 日本機械学会論文集（C 編), **74**-744, 1985/1992 (2008)

5) 中裕一郎: 交差流動の構造 鉄道駅における旅客の交錯流動に関する研究 (1) , 日本建築学会論文報告集, 258, 93/102 (1977)

6) 兼田敏之: 歩行者流のエージェントシミュレーション, 計測と制御, **43**-12, 944/949 (2004)

7) 木村哲, 森山甲一, 武藤敦子, 松井藤五郎, 犬塚信博: 強化学習による衝突回避エージェントモデルの自動生成, 第 18 回情報学ワークショップ論文集, L-4A-1 (2020)

8) 福田忠彦: 図形知覚における中心視と周辺視の機能差, テレビジョン学会誌, **32**-6, 492/498 (1978)

9) Unity: `https://unity.com/ja` , 2021 年 2 月 3 日参照