2ZD-06

強化学習を用いたガイドキャラクタの経路誘導行動の生成

石川友哉† 森博志† 外山史† 東海林健二†
宇都宮大学大学院工学研究科†

1. はじめに

CG キャラクタの人を模した外見特徴や特性を活かして VR におけるサイバー空間や AR 技術を利用した実空間でのバーチャルガイドとしての利用例が見られるようになってきた[1,2]．CG キャラクタによるバーチャルガイド（以降，ガイドキャラクタ）は，対象空間において目的地までの誘導や説明をする用途に用いられる．そのため，ガイドキャラクタに要求される機能として，ユーザの状態に応じた適切な立ち位置の調整や，目的地まで案内するための移動動作制御が挙げられる．

ガイキャラクタの立ち位置，移動動作の制御問題に対して，ユーザや目的地の位置関係を考慮に入れて，移動可能な座標を事前に設定し，接続関係を考慮して設定した動作データを組み合わせることで立ち位置の制御が可能になる[3]．しかし，事前にコンテンツに応じた設定が必要であり，また，移動可能な位置が制限されてしまうため，ユーザの移動に対して適応的に移動することは難しい．

そこで本稿では，ユーザの移動に対するガイドキャラクタの適応的な移動動作の生成手法を提案する（図 1）．本手法では強化学習を用いてガイドキャラクタが適切な立ち位置を取りながらユーザを目的地まで先導する経路誘導行動を獲得する．これにより，ユーザと環境に適応したキャラクタの経路誘導行動を得られることが期待できる．

2. 強化学習を用いた経路誘導行動の生成

2.1 強化学習

強化学習は，得られる報酬を基に環境に適した行動を強化する学習手法である．強化学習では，環境の状態 s_t を観測し，状態 s_t において行動 a_t を出力する．行動 a_t の出力の結果，状態 s_{t+1} へと遷移し，状態 s_{t+1} において観測される報酬 r_{t+1} を学習器に与え，最終的な累積報酬を最大化するように学習する．

状態・行動系列を $h=(s_1, a_1, s_2, a_2, \ldots, s_T)$，累積報酬を $R(h)=r_1+\gamma r_2+\gamma_2 r_3+\ldots$，状態 s における行動則（方策）を $\pi_\theta(a|s)$とすると，期待報酬 $E^{\pi\theta}[R(h)]$を最大化するように方策パラメータ θ を最適化する問題に定式化される．

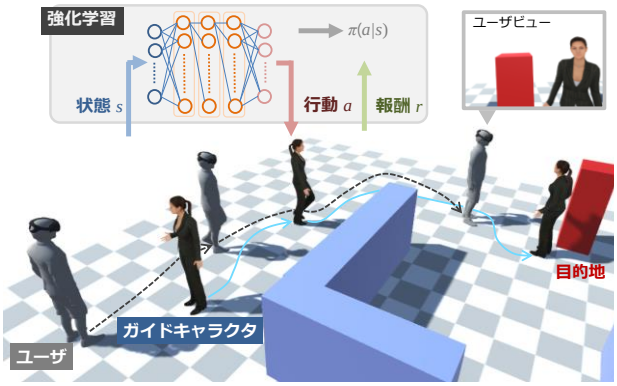

図 1 提案手法の概要

2.2 ガイドキャラクタの経路誘導行動制御への強化学習の適用

実際のガイドが被案内者（ユーザ）を案内する場合には，ガイドは目的地に被案内者が到達できるように先導する．そのため本稿ではガイドキャラクタの経路誘導行動を次のように定義する．

- ガイドはユーザの視界内にいる
- ガイドはユーザから一定の距離内にいる
- ガイドはユーザが目的地に到着するように先導するように位置をとる

以上を満たすような行動を強化学習により獲得する．そこで，観測される状態とガイドキャラクタの行動，報酬を次のように定義する．

観測される状態は，ガイドキャラクタとユーザとの相対位置(x_u, z_u)と角度 α_u，ガイドキャラクタから目的地までの相対位置(x_d, z_d)と角度 α_d，ガイドキャラクタの移動速度(Δx_g, Δz_g)，ユーザの移動速度(Δx_u, Δz_u)，ガイドキャラクタと周囲の障害物との相対的な位置情報とする．本稿では周囲の障害物との相対的な位置情報は周囲 16 方向にレイを飛ばし障害物の識別子と衝突の有無，相対距離を取得する．以上の 56 次元ベクトルをエージェントであるガイドキャラクタが観測する状態とする．

行動はキャラクタのモーションクリップとし，報酬はユーザの視野領域内に位置し目的地に到着したときに与える．

3. 実験

3.1 実験条件

本実験では，経路誘導行動を構成するガイドキャラクタの行動であるモーションクリップとして，前進歩行，左右 45，90，180 度の回転，左右の斜め前方への旋回歩行，待機の計 10 種を用いた．

Generating Route Guidance Behavior of Character using Reinforcement Learning
Tomoya Ishikawa†, Hiroshi Mori†, Fubito Toyama†, Kenji Shoji†
†Utsunomiya University, Graduate School of Engineering

報酬は，ユーザキャラクタの視野領域を前方 120 度，半径 3m の扇状の領域と設定し，視野領域内に位置しつつ，目的地の周囲 1.5m に到達した際に，r=1.0 を与えた．

学習に用いるニューラルネットワークは入力層，出力層，中間層 3 層の計 5 層で，入力ユニット数は観測情報の 56，出力ユニット数はガイドキャラクタの行動の 10 である．本稿では，方策ベースのアルゴリズムとして，PPO(Proximal Policy Optimization)[4]を用いた．

3.2 学習

図 2 に示す環境において初期位置と目的地をランダムに設定し，経路探索結果に基づいてユーザキャラクタが目的地に移動する．ガイドキャラクタは移動するユーザキャラクタを目的地に誘導するような適切な立ち位置を学習する．

1 ステップを 1 つのモーションクリップの実行，1 エピソードを複数のモーションクリップを連続して実行しユーザキャラクタが目的地に到着するまでとする．ただし，ガイドキャラクタが障害物またはユーザに衝突した場合と，ユーザキャラクタの視野領域から外れた場合にはその時点でエピソードを終了とする．

学習を 50 万ステップ行ったときの累積報酬の推移を図 3 に示す．図 3 から累積報酬が増加し収束していることが確認できる．

3.3 実行結果

3.2 の学習結果を用いて学習時とは異なる環境においてシミュレーションを行った結果を図 4 に示す．ユーザキャラクタは学習時と同様に目的地まで移動する．実行結果より，ガイドキャラクタがユーザキャラクタを目的地まで誘導するように移動していることが確認できる．

4. おわりに

本稿では，強化学習を用いて適切な立ち位置を取りながらユーザを目的地まで誘導するガイドキャラクタの経路誘導行動の生成手法を提案した．実験結果より，ユーザキャラクタの位置移動に対してガイドキャラクタが経路誘導を行うような立ち位置を取りながら移動動作を実行できていることが確認できた．今後の課題として，ガイドキャラクタの行動数を増やし，より柔軟な立ち位置制御への対応が挙げられる

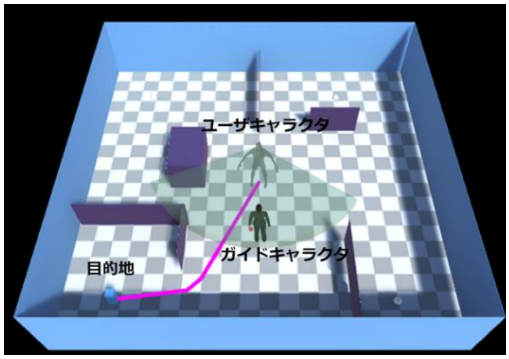

図 2　学習環境

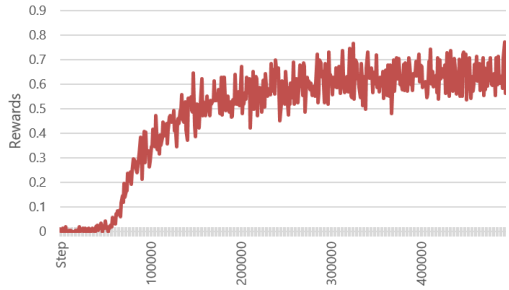

図 3　累積報酬の推移

参考文献

[1] KDDI，シリコンスタジオ（2017）プレスリリース「バーチャルアテンダントが案内する VR 不動産コンテンツ」，（https://www.au.com/information/topic/mobile/2017-065/），参照 2018-08-02

[2] パソナテックシステムズ，KDDI，（2018）プレスリリース「パソナテックシステムズと KDDI、AI を活用したバーチャルキャラクターによるガイド業務の実証実験を大手町牧場で開始」(https://prtimes.jp/main/html/rd/p/000000381.000016751.html)，参照 2018-08-02

[3] Rei Chan, Junichi Hoshino, ”Building immersive conversation environment using locomotive interactive character”, Journal of Universal Computer Science, Vol.12, No.2, pp.149-160, 2007.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, “Proximal Policy Optimization Algorithms”, arXiv preprint arXiv:1707.06347v2, 2017.

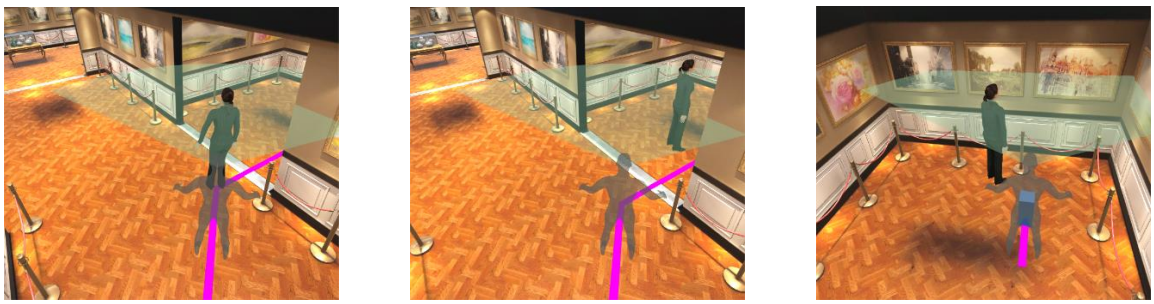
(a)移動開始時　(b)移動時　(c)到着時

図 4　学習結果に基づく経路誘導行動の生成結果